UNIVERSITÄT
DES
SAARLANDES

Aus dem Bereich Klinische Bioinformatik
der Medizinischen Fakultät
der Universität des Saarlandes, Homburg, Saar

Von Plattformen zu miRNA-Biomarkern
-
Methoden zur miRNA-Molekulardiagnostik

Dissertationschrift

Zur Erlangung des Grades Dr. rer. med.
der Medizinischen Fakultät
der Universität des Saarlandes

vorgelegt von

Dipl. Ing. Cord Friedrich Stähler

Homburg, März
2018

Eidesstattliche Erklärung:

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Bei der Ausarbeitung haben mir Herr Prof. Dr. Andreas Keller und Herr Prof. Dr. Eckart Meese unentgeltlich geholfen, indem sie die vorliegende Schrift kritisch korrigiert haben. Weitere Personen waren an der inhaltlich-materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater/innen oder anderer Personen) in Anspruch genommen. Außer den Angegebenen hat niemand von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form in einem anderen Verfahren zur Erlangung des Doktorgrades einer anderen Prüfungsbehörde vorgelegt.

Ich versichere an Eides statt, dass ich nach bestem Wissen die Wahrheit gesagt und nichts verschwiegen habe.

Vor Aufnahme der vorstehenden Versicherung an Eides statt wurde ich über die Bedeutung einer eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung belehrt.

Cord F. Stähler,
Homburg, 10.03.2018

Danksagung:

Das Anfertigen einer Doktorarbeit erfordert viel Zeit und Ressourcen, gerade wenn die Doktorarbeit zusätzlich zur Ausübung eines Berufs geschrieben wird. Daher sind die Unterstützung und das Verständnis der Familie unverzichtbar.

Maßgeblich an der Betreuung der Doktorarbeit, genau wie an unzähligen gemeinsamen Projekten, war mein Doktorvater, Prof. Andreas Keller, beteiligt. Mit dem Abschluss dieser Doktorarbeit endet ein weiteres wichtiges gemeinsames Kapitel, aber unsere inspirierende and motivierende Zusammenarbeit geht sicherlich noch lange weiter.

Hoch komplexe Projekte im Umfeld der Molekulardiagnostik erfordern verschiedene interdisziplinäre Expertisen. Je komplexer die Projekte werden, um so wichtiger ist es, dass man mit den richtigen Kooperationspartnern arbeitet. Ohne deren Hilfe wäre es nicht möglich gewesen die in dieser Arbeit beschriebenen Projekte erfolgreich durchzuführen. Die Beiträge der einzelnen Personen sind in den entsprechenden Originalarbeiten gekennzeichnet, dennoch möchte ich die wichtigsten Personen, die mich begleitet und unterstüzt haben, hervorheben.

Bei der Firma febit war das seit der Gründung vor allem mein Bruder Peer Friedrich Stähler, dem ich mein Verständnis für die Molekulare Biologie und die Genetik verdanke. Gemeinsam mit meinem Bruder Peer sowie unserem Vater, PD Dr. med. Fritz Stähler, haben wir viele der frühen, grundlegenden Konzepte und Ideen der vorliegenden Arbeit formuliert und in Projekten umgesetzt. Während meiner anschließenden Zeit bei Siemens haben mich besonders die Diskussionen mit Prof. Dr. Hermann Requardt inspiriert und motiviert wichtige Aspekte der Diagnostik zu durchdringen, diese aber auch immer im Kontex der Anwendbarkeit im bestehenden Gesundheitssystem zu betrachten.

Meinen Arbeitgebern febit, Siemens und Merck möchte ich für Ihre generelle Bereitschaft die Forschung zu unterstützen, danken, da sie dadurch meine und unzählige andere wissenschaftliche Arbeiten ermöglichen, die wiederum die Grundlage für unsere immer bessere Gesundheitsversorgung bilden.

Die meiste Unterstützung habe ich in den vergangenen Jahren von den Arbeitsgruppen (AG) Humangenetik, Prof. Meese, und Klinische Bioinformatik, Prof. Keller, der UDS erhalten. Bei ihnen und den Teams der Arbeitsgruppen bedanke ich mich besonders.

Abstract – Englisch

An obvious way to improve human healthcare is to develop new and more effective drugs. Another opportunity is however to develop solutions that allow to utilize the available drugs better. This includes more accurate and early diagnosis of pathologies, improved therapy selection as well as digital and patient centric solutions in healthcare systems. Especially in molecular diagnostics new biomarkers have been developed and partially shown promising results in terms of improving patient care. In this work I describe the development of respective platform techniques, biomarkers and computational solutions during my PhD thesis.

First, I briefly introduce the concept of a flexible microarray platform and assays, such as the MPEA assay, tailored for the fast and efficient quantification of miRNA signatures. Then, I describe how we made use of respective platforms along with computational solutions to improve the understanding of physiological and pathophysiological processes. Further, I present results on my efforts to develop new molecular diagnostic biomarkers based on circulating miRNAs. Here, my special focus was in cancer (most importantly lung cancer) and diseases affecting the Central Nervous System (most importantly Multiple Sclerosis, Alzheimer's Disease and Parkinson's Disease). Together with the supervisors of my thesis I was among the first researchers worldwide to recognize that small non-coding RNAs (most importantly microRNAs) measured from body fluids have a great potential as biomarkers. An obvious advantage to messenger RNAs is the small length of the molecules of only 17-22 nucleotides. This makes microRNAs stable in vivo but also in vitro.

Finally, I will mention recent developments in patient care. The current trend is clearly the digitalization of central parts of healthcare. This affects all stakeholders in the healthcare system, most importantly medical doctors and patients. Especially patient empowerment and self-containment of medical data is becoming more important. Again, Multiple Sclerosis is used as an example. But also for physicians, computational tools have to be implemented to support them in making treatment decisions from highly complex data. In sum, my thesis describes the road from developing a molecular diagnostic platform over the research on biomarkers for detecting disease in time towards holistic computational solutions to improve patient care.

Zusammenfassung – Deutsch

Es ist offensichtlich, dass man Krankheiten besser behandeln kann, wenn man neue und effektivere Medikamente und Therapien entwickelt. Eine andere Möglichkeit ist es, Lösungen zu entwickeln, die es erlauben, vorhandene Medikamente besser einzusetzen. Das schließt die frühzeitige Diagnose von Erkrankungen, eine verbesserte Wahl der richtigen Therapie und die Entwicklung von patienten-zentrischen digitalisierten Lösungen mit ein. Insbesondere in der Molekulardiagnostik wurden neue vielversprechende Biomarker entwickelt. In dieser Arbeit führe ich meine Beiträge zur Entwicklung von Plattform Technologien zum Messen von Biomarkern aus, erläutere die Erforschung von Biomarkern selbst und beschreibe die Anwendung der dazugehörigen, computergestützten Methoden.

Beginnen möchte ich mit einer Beschreibung der Entwicklung einer flexiblen Mikroarray Plattform und Assays, wie zum Beispiel des MPEA Assays, die maßgeschneidert für die schnelle und effiziente Quantifizierung von miRNA Biomarkern sind. Dann gehe ich darauf ein, wie wir Plattformen, Assays und computergestützte Lösungen eingesetzt haben, um physiologische und pathologische Prozesse besser zu verstehen. Außerdem präsentiere ich Resultate meiner Bemühung, neue molekulardiagnostische Biomarker basierend auf zirkulierenden miRNA Mustern zu entwickeln. Hierbei habe ich mich auf Krebs (vornehmlich Lungentumore) und Erkrankungen, die das Zentrale Nervensystem betreffen (Multiple Sklerose und die Alzheimer Erkrankung), konzentriert. Gemeinsam mit meinen Betreuern war ich unter den ersten Forschern weltweit, die das große Potenzial kleiner nicht-kodierender RNAs (am wichtigsten dabei microRNAs), die aus Blut gemessen werden können, erkannt haben. Ein offensichtlicher Vorteil gegenüber mRNA Biomarkern ist die kurze Länge von nur 17-22 Nukleotiden. Diese macht miRNAs sowohl in-vivo als auch in-vitro stabil.

Letztlich gehe ich in meiner Arbeit auf momentane Entwicklungen in der Patientenversorgung ein. Ein klarer Trend ist die Digitalisierung zentraler Teile der Gesundheitsversorgung. Das betrifft alle Personen im Gesundheitswesen, allen voran Mediziner und Patienten. Selbstbestimmung des Patienten wird besonders wichtig werden. Hier dient mir wieder Multiple Sklerose als ein Beispiel. Auch für Ärzte müssen, angesichts der immer komplexeren Daten, computergestützte Lösungen entwickelt werden, die ihnen helfen, die richtige Therapieentscheidung zu treffen. Zusammenfassend halte ich fest, dass meine Arbeit den Weg von der Entwicklung einer molekulardiagnostischen Plattform über die Entwicklung von Biomarkern zur Frühdiagnose von Erkrankungen bis hin zu ganzheitlichen computergestützten Lösungen, die die Patientenversorgung verbessern, beschreibt.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Seit der Veröffentlichung des ersten menschlichen Genoms im Jahre 2002 durch das Human Genome Project (HGP) [1] und die Firma Celera werden stetig neue Technologien entwickelt, die es uns ermöglichen, molekulare Muster aus verschiedensten Organismen zu lesen. Entsprechende DNA oder RNA Muster werden in vielen verschiedenen Gebieten verwendet, neben der Erforschung von Krankheiten die Menschen betreffen sind unter anderem die Agrikultur und die Nutztierhaltung wichtige Anwendungsbeispiele für molekulare Analysen. Hier werden Gene, Genexpression, Methylierung oder nicht-kodierende RNA sowie Protein Muster erhoben, um Krankheiten des Menschen besser erforschen beziehungsweise besser zu verstehen, wie der Ertrag von Tieren erhöht werden kann und Tiere unter Umständen ohne den Einsatz von schädlichen Antibiotika gesünder leben können.

Die Technologien, die dabei entwickelt werden, sind zunehmend komplexer geworden. Während in den Anfängen zu Beginn des Jahrtausends noch sogenannte Mikroarrays eingesetzt wurden, hat die Firma Solexa im Jahr 2005 eine disruptive neue Technologie für den Massenmarkt vorgestellt, Next-Generation Sequencing oder Hochdurchsatz-Sequenzierung (HTS) [2]. Während Mikroarrays zunächst klar für das Auslesen des Transkriptoms (der zu einem bestimmten Zeitpunkt in einem bestimmten Zelltyp vorkommenden Menge aller Gene eines Organismus) verwendet wurden, wurde HTS vor allem für die Sequenzierung des Erbgutes (die Gesamtmenge der DNA, also Protein kodierende Gene, regulatorische Elemente und andere Teile, deren Bedeutung zum Teil noch nicht völlig klar ist) eingesetzt. In den vergangenen Jahren wurden jedoch mehr und mehr Assays entwickelt, die es auch erlauben, HTS zur quantitativen oder zumindest pseudo-quantitativen Messung des Transkriptoms und sogar weiterer Aspekte wie der Methylierung der DNA präzise zu messen. Auch zur Entschlüsselung nicht-kodierender Elemente (kleine nicht kodierende RNAs wie piRNAs oder microRNAs (miRNAs) sowie

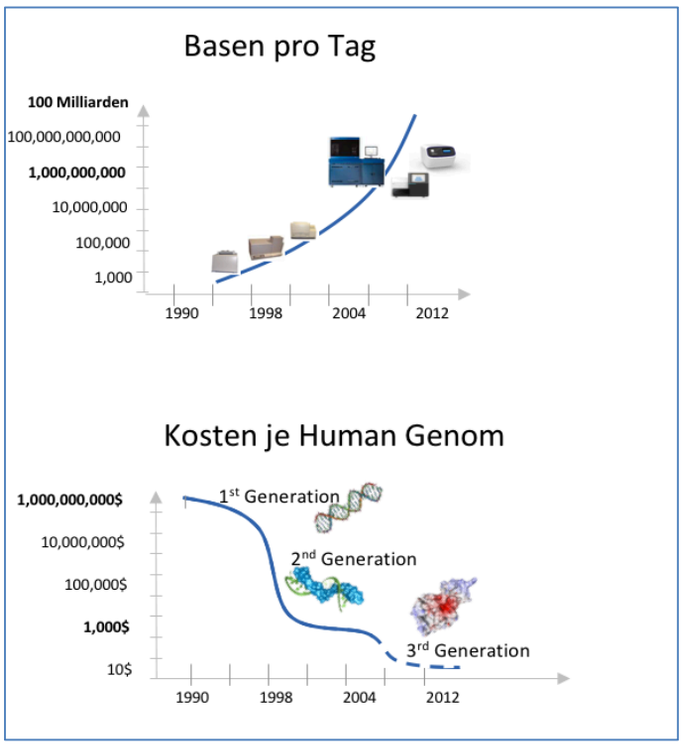


Abbildung 1: Kosten je Genom.

Der obere Teil der Abbildung zeigt den exponentiell wachsenden Durchsatz moderner Sequenzier-Technologien. Im unteren Teil wird schematisch dargestellt, dass die Kosten der Sequenzierung nicht nur evolutionär innerhalb der Technologie Generationen sinken, sondern auch revolutionären Charakter zwischen verschiedenen Generationen haben.

längerer Elemente wie zum Beispiel lincRNAs) wird immer mehr HTS Technologie verwendet. Die Entwicklung neuer Sequenzier-Technologien verläuft dabei nur teilweise evolutionär, oft gibt es Technologiesprünge, welche einen revolutionären Charakter haben. Ein Beispiel war der Sprung der ersten Generation Sequenzierer, wie sie zum Beispiel für das Humane Genom Projekt eingesetzt wurden, hin zur HTS (zweite Generation). Dadurch konnten die Kosten pro Humanes Genom von mehr als einer Milliarde Dollar auf wenige tausend, oft sogar weniger als tausend Dollar verringert werden. Bereits heute werden Sequenzierer der Dritten Generation (sogenannte Nanoporen Sequenzierung) erfolgreich eingesetzt. Die technische Weiterentwicklung verspricht dabei, dass in einigen Jahren Geräte der Größe eines USB Sticks ein Humanes Genom für weniger als 100 Dollar sequenzieren können. Die Kosten je Genom, der Durchsatz eines Sequenzierers je Tag und die zeitliche Abhängigkeit der Sequenzierer Generationen wird in Abbildung 1 und Abbildung 2 dargestellt. Wegen dieser rapiden technischen Entwicklung der Sequenzierung werden Mikroarrays heute sehr viel seltener angewendet. Hauptsächlich, wenn eine relativ präzise Quantifizierung von vielen Genen oder nicht-kodierenden Elementen gefragt ist, wird auf diese bewährte, hochparallele Technologie zurückgegriffen.

Eine Herausforderung entsprechender Technologien ist eine stetig anwachsende Komplexität und damit auch eine sehr viel höhere Bedeutung von Algorithmen, Datenstrukturen und Bioinformatik-Lösungen. Das gilt für alle molekularen Messungen, nicht nur HTS, sondern auch Proteinmuster, die mit Massenspektrometrie erhoben werden. Abbildung 3 zeigt dabei anschaulich, um wie viele Größenordnungen die Gesamtgröße von molekularen Datensätzen in den vergangenen Jahren zugenommen hat. Während für Mikroarrays noch wenige Megabyte erreicht bzw. benötigt wurden, sind heutige HTS Datensätze leicht viele Gigabyte oder sogar Terabyte groß. Entsprechend

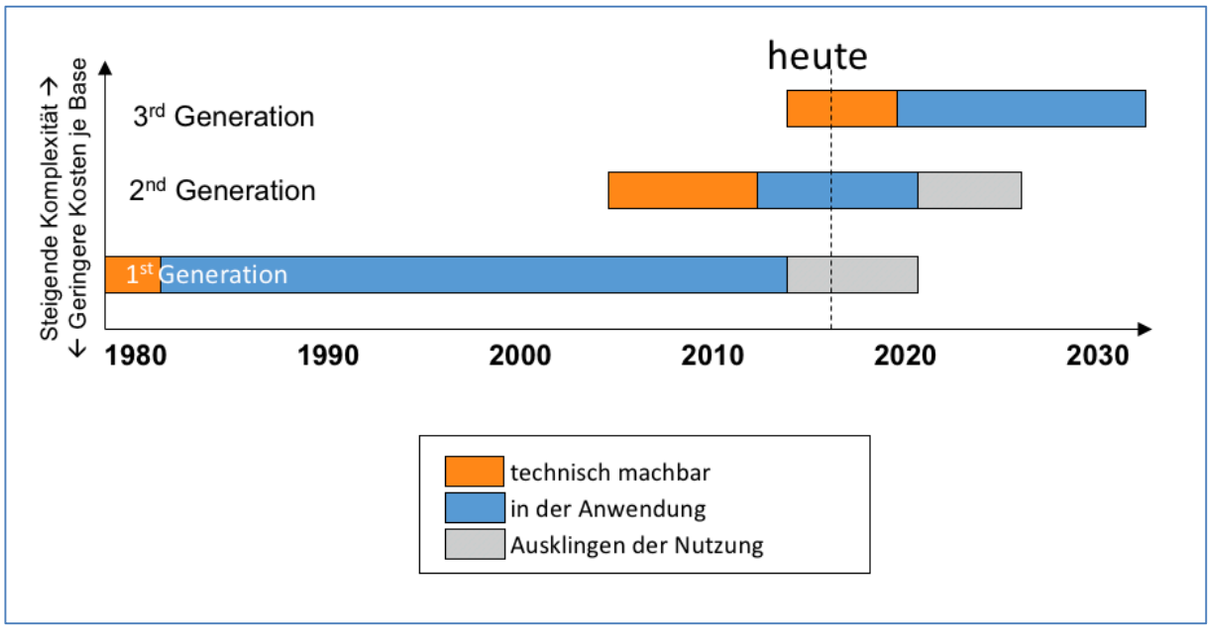


Abbildung 2: Zeitliche Überlappung und Reife der verschiedenen Sequenzier-Technologien.

erreichen Computer, die für das Speichern der Daten verwendet werden (sogenannte Fileserver) Kapazitäten die in den Petabyte Bereich gehen. Eine der größten Herausforderungen ist es, adäquate computer-gestützte Lösungen für diese Problematik zu entwickeln und die Ergebnisse solch komplexer Methoden für Ärzte aufzubereiten.

Der vorangegangene Absatz hat einen kurzen Überblick über die technische Entwicklung in der Molekularbiologie und Genetik gegeben. Sicherlich erhebt dieser kurze Überblick

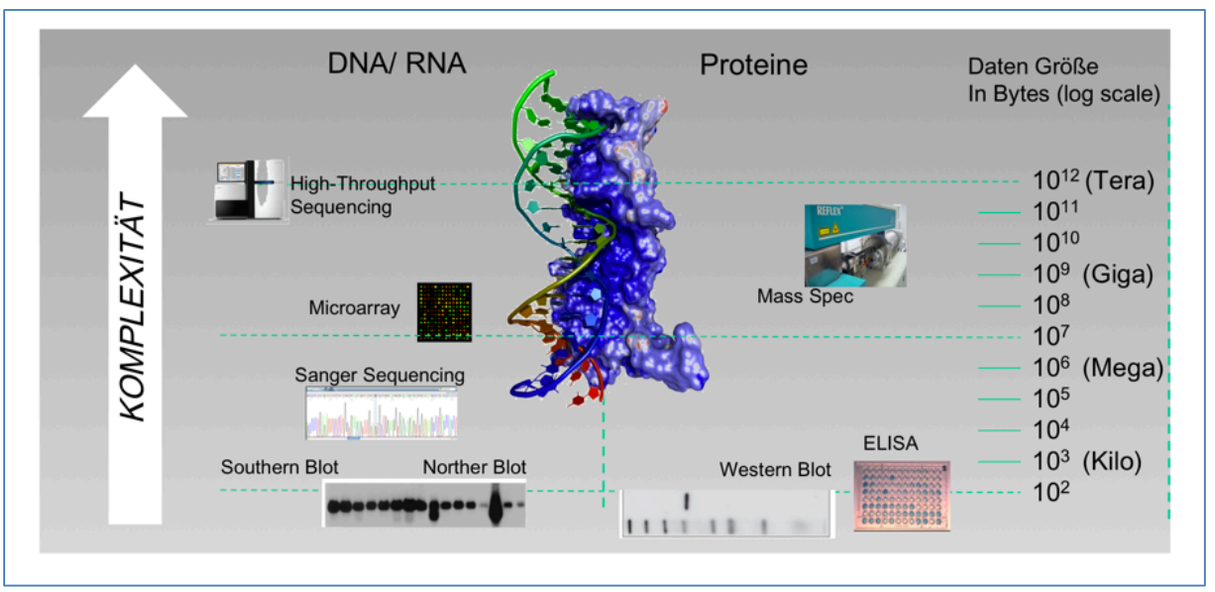


Abbildung 3: Die Komplexität verschiedener molekularer Technologien.

Hochdurchsatz Methoden (HTS, Massenspektrometrie) haben dazu geführt, dass Datensätze heute oft viele Giga- wenn nicht sogar Terabyte groß sind.

keinen Anspruch auf Vollständigkeit, wie es tiefgehende Übersichtsartikel in diesem Gebiet tun. Hierzu verweise ich auf geeignete Übersichtsartikel [3-6]. Der Überblick hilft jedoch, die technischen Entwicklungen, die in Kapitel 3 und 4 beschrieben werden, besser einzuordnen und er zeigt vor allem einen klaren Trend hin zu Hochdurchsatz-Plattformen. Diese bisher nie dagewesene Menge an Daten motiviert auch die Entwicklung hin zu computergestützten Analyse Methoden. Die technische Weiterentwicklung, zusammen mit neuen Analyseverfahren, hat auch dazu geführt, dass die Forschung in Lebenswissenschaften insgesamt signifikante Fortschritte gemacht hat. Über diese Entwicklung, die parallel zur technologischen Weiterentwicklung stattgefunden hat, wird im folgenden Abschnitt eingegangen.

Während in den letzten Jahrzehnten des vergangenen Jahrhunderts hauptsächlich Methoden wie PCR eingesetzt wurden, um einzelne Gene zu verstehen, hat es die Entwicklung von DNA und RNA Mikroarrays in den vergangenen 30 Jahren ermöglicht, das Verständnis von Genen im Menschen und vielen anderen Organismen auf einer systematischen Ebene intensiv voranzutreiben [7-10]. Der Schritt von wenigen einzelnen Genen hin zum Transkriptom, der Menge aller zu einem bestimmten Zeitpunkt exprimierten Gene, war einer der wesentlichen Fortschritte in der Molekularbiologie der vergangenen Jahrzehnte. Durch die oben beschriebene Entwicklung der Sequenziertechnologie wurde neben Genen, die für Proteine kodieren weitere Elemente identifiziert, die vom Genom abgeschrieben werden, aus denen aber keine Proteine

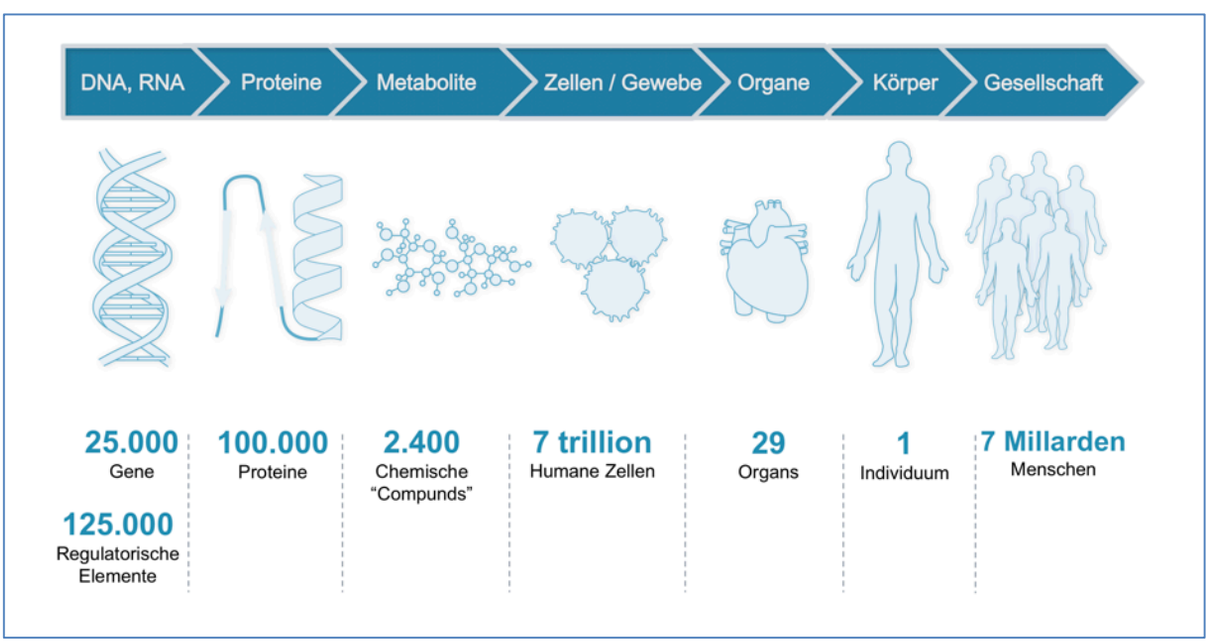


Abbildung 4: Multi-Skalen in der Biologie und Biomedizin.

Angefangen vom genetischen Code mit seinen 3.3 Milliarden Basen und 20.000-25.000 Genen, hunderttausenden regulatorischen Elementen und Proteinen in abermilliarden Zellen über Organe und Organsysteme hin bis zu einem Individuum und einer Gemeinschaft von Individuen. Um die Komplexität verstehen zu können sind präzise molekulare Methoden, Medizinische-Bildgebung, aber auch computer-gestützte Analysen notwendig.

gebildet werden. Diese Elemente werden nicht-kodierende RNA (ncRNA) genannt. Generell wird zwischen langen [11] und kurzen nicht-kodierenden RNAs unterschieden [12]. Kurze nicht-kodierende RNAs sind neben anderen miRNA, tRNAs, piRNA oder yRNAs. Die Bedeutung dieser nicht-kodierenden Elemente für die Organisation von Zellen wurde von Gomes bereits 2013 zusammengefasst [13]. Neben nicht-kodierenden RNAs wurden viele andere epigenetische Mechanismen wie die Methylierung der DNA zunehmend erforscht und durch Hochdurchsatz-Methoden zwischen 1993 und 2016 signifikante Fortschritte erzielt [14, 15]. Auch das Verständnis der Modulation des Chromatin Zustandes durch Histon-Modifikation hat das Verständnis über regulatorische Mechanismen der Genexpression verändert [16]. Da auch massenspektrometrische Verfahren weiterentwickelt wurden und sogar präzise quantitative Messungen von Protein Mengen im hohen Durchsatz möglich werden [17, 18] ist es momentaner Gegenstand vieler Forschungsprojekte, die verschiedenen -omics Technologien an Patienten integrativ zu messen. Sogenannte multi-omics Studien sind einer der momentanen Trends in der Molekularbiologie und der Biomedizin [19]. Solche hoch-komplexen Studien erfordern allerdings auch spezielle Analyse-Strategien und neue Algorithmen [20, 21]. Es ist sogar möglich, entsprechende molekulare Muster bis hin zu einzelnen Zellen zu messen [22].

Die Erfahrung hat jedoch gezeigt, dass multi-omics Daten-Analysen auch Herausforderungen bergen. Mehrere Milliarden Basen im Humanen Genom, 25.000 Gene, hunderttausende regulatorische Elemente und mindestens ebenso viele Proteine erlauben schier endlose Kombinationsmöglichkeiten. Während die Suche nach dys-regulierten Genen noch der Suche nach der Nadel im Heuhaufen entspricht, begegnen wir analog in multi-omics Analysen einer exponentiell größeren Herausforderung. Noch dazu kommt, dass es extrem wichtig, aber auch zeitaufwändig und weitaus schwieriger als gedacht ist, Hochdurchsatz-Datensätze manuell zu kurieren [23]. Nichtsdestotrotz werden entsprechende Ansätze für die verschiedensten Krankheiten angewendet, wie zum Beispiel Kolon Karzinome [24], Brustkrebs [25], Leberkrebs [26], Lungenkrebs [27], Kardiomyopathien [28] oder Alzheimer [29].

Die im ersten Teil der Einleitung beschriebenen Technologien dienen im weitesten Sinne dazu, den genetischen Code zu lesen. Sie werden mit Hilfe von Algorithmen dazu verwendet, den genetischen Code und was aus dem genetischen Code gemacht wird besser zu verstehen. Beides kann in der Zukunft noch besser angewendet werden, um den genetischen Code zu schreiben. Dabei spielen Gen Editing wie TALENs, ZNFs [30], oder CRISPR Cas [31] schon heute eine wichtige Rolle.

Im Folgenden möchte ich kurz den Aufbau der vorliegenden Arbeit zusammenfassen. Für meine Forschung spielt eine der oben genannten Molekülklassen eine besondere Rolle, microRNAs (miRNAs). Während Gene, lange nicht-kodierende RNAs oder andere Moleküle relativ empfindlich gegenüber äußeren Einflüssen sind, haben sich miRNAs als sehr stabil – in vivo und in vitro – gezeigt [32]. Gleichzeitig sind sie Masterregulatoren in der Genexpression [33] und als Biomarker für eine Vielzahl von Erkrankungen beschrieben [34]. Die Entwicklung von blutbasierten diagnostischen Tests basierend auf miRNA Mustern steht im Fokus des zweiten Teils meiner Arbeit. Da miRNAs hier eine so zentrale Bedeutung haben, möchte ich an dieser Stelle nicht nur auf die grundlegende Primärliteratur verweisen, in der die Entdeckung und Entwicklung von miRNAs sowie ihre biologische Funktion erklärt wird [35-38]. Ich habe das an die Einleitung anschließende Kapitel 2 den miRNAs gewidmet: Dort beschreibe ich die Hintergründe der Entdeckung, die Biogenese, die molekulare Funktion und den gegenwärtigen Stand der miRNA Forschung.

In Kapitel 3 gehe ich auf einige technische und Bioinformatik-Aspekte ein, die in der Arbeit angewendet wurden. Kapitel 3 ist entsprechend kurz gehalten, da die projekt-spezifischen Methoden in den einzelnen Kapiteln der Resultate detaillierter beschrieben sind.

Im Resultat Kapitel 4 beschreibe ich zunächst eigene technologische Entwicklungen in der Molekularbiologie (Kapitel 4.1). Das beinhaltet ein flexibles Mikroarray Instrument, das Geniom, das dezentral im Labor eingesetzt werden kann, um mittels eines Synthese-Verfahrens in Situ Mikroarrays über Nacht herzustellen [39-41]. Diese Arbeiten habe ich hauptsächlich aus Sicht eines Ingenieurs durchgeführt. Da in dieser Arbeit maßgeblich die Entwicklung der Biomarker beschrieben ist, dient Kapitel 4.1. hauptsächlich dazu, eine Gesamtübersicht und den Kontext zu bekommen. Für das Geniom Instrument wurden verschiedene Assays entwickelt, wie zum Beispiel eine Anreicherung für sogenanntes targeted-next-generation-sequencing tNGS oder der Microfluidic Primer Extension Assay MPEA, der es erlaubt, miRNAs besonders exakt zu quantifizieren [42]. Darüber hinaus wird die Entwicklung klinischer Assays beschrieben, um miRNAs möglichst kostengünstig und schnell direkt im Krankenhaus („point-of-care“) zu messen [43, 44]. Final befasse ich mich im Kapitel über Technologie-Entwicklung mit einer neuen Sequenzier-Technologie, die in China entwickelt wurde (cPAS) und die besonders zur Quantifizierung kleiner RNAs geeignet ist [45].

Die Erfahrung hat gezeigt, dass obwohl ein Wert in der Entwicklung von neuen Technologien besteht, der eigentliche Schlüssel zum Erfolg, der Einsatz der richtigen Technologie ist, um biologisches oder medizinisches Wissen zu erlangen. Der zweite Teil

der Arbeit befasst sich genau damit: Wie können wir vorhandene Technologie einsetzen, um Krankheiten früher zu erkennen und besser zu behandeln? Ausgehend von Ergebnissen in der genetischen Diagnostik, dem Messen der DNA hat sich gezeigt, dass RNA Muster, vor allem miRNA Muster, ein signifikantes Potenzial haben, Krankheiten früh zu erkennen [46-49]. Besonders das minimal-invasive Messen von miRNA Signaturen aus Blutproben ist ein vielversprechender Ansatz. Dabei ist es zunächst wichtig, die technische Stabilität der Marker nachzuweisen [50], aber auch biologische Einflussfaktoren wie das Alter oder das Geschlecht müssen verstanden werden [51]. Auch die organspezifische Komponente der zirkulierenden miRNA Muster ist von zentraler Bedeutung für die Entwicklung minimal-invasiver Biomarker [52]. Alle vorgenannten Aspekte bilden die Grundlage für die Erforschung von miRNA Mustern als Biomarker, hier am Beispiel von Lungenerkrankungen [53-55] und Erkrankungen des zentralen Nervensystems [56-58]. Der Vergleich von Mustern in Erkrankungen, die verschiedene Organe betreffen, hat zudem eine generell krankheitsspezifische Komponente ergeben, miRNAs, die unabhängig der Erkrankung höher oder tiefer exprimiert sind als in Kontrollprobanden [34]. Die Entwicklung der miRNA Biomarker, der zentrale Bestandteil meiner Arbeit, ist in Kapitel 4.2 beschrieben.

Ab Kapitel 4.3 werden Aspekte die über die miRNAs als Biomarker hinausgehen behandelt. Das betrifft zum Beispiel die co-expression von miRNAs, also verschiedene miRNAs die gegenseitig die Aufgabe der jeweilig anderen übernehmen können [59] und Algorithmen, die es erlauben den Einfluss der Genregulation der Biomarker zu verstehen [60]. Final werden in Kapitel 4.4 kurz Entwicklungen in der synthetischen Biologie erläutert [61, 62]. Im Ausblick wird motiviert, dass das Editieren oder die Modifikation von RNA hervorragende Therapieoptionen sind, die es wert sind weiter beleuchtet zu werden. Außerdem wird der stetig voranschreitende Trend der Digitalisierung im Gesundheitswesen am Beispiel der Multiplen Sklerose skizziert. Patientenzentrische Lösungen, die auch die Selbstbestimmung des Patienten über seine Daten besser ermöglichen, bilden die Grundlage, um vor allem chronische Krankheiten in Zukunft noch effektiver behandeln zu können.

Die Resultate, die in dieser Arbeit zusammengefasst sind, beruhen auf 21 Originalarbeiten, die ich in den folgenden Kapiteln vorstellen möchte. Alle Originalarbeiten finden sich im Anhang an die vorliegende Ausarbeitung. Wie in komplexen wissenschaftlichen Arbeiten üblich, sind diese Publikationen im Team entstanden. Obwohl ich mich in der Darstellung auf meine Beiträge fokussiere schmälert das nicht die Beiträge der Koautoren, die ich sehr zu schätzen weiß. Ihr Beitrag ist in den entsprechenden Originalarbeiten gekennzeichnet.

Kapitel 2

microRNAs

Viele Jahrzehnte wurde der Begriff „Junk DNA“ von Wissenschaftlern verwendet. Dieser bezeichnete ursprünglich die Teile der DNA, die scheinbar keine Funktion besitzen und wurde in den 1960er und 1970er Jahren durch Susumu Ohno geprägt [63]. Seit dieser Zeit wurde „Junk DNA“ in der wissenschaftlichen Welt zunehmend diskutiert [64-71]. Die in der Einleitung skizzierten Technologiesprünge haben es erlaubt, immer größeren Bereichen des Humanen Genoms Funktionen zuzuweisen. Daten des ENCODE Projekts

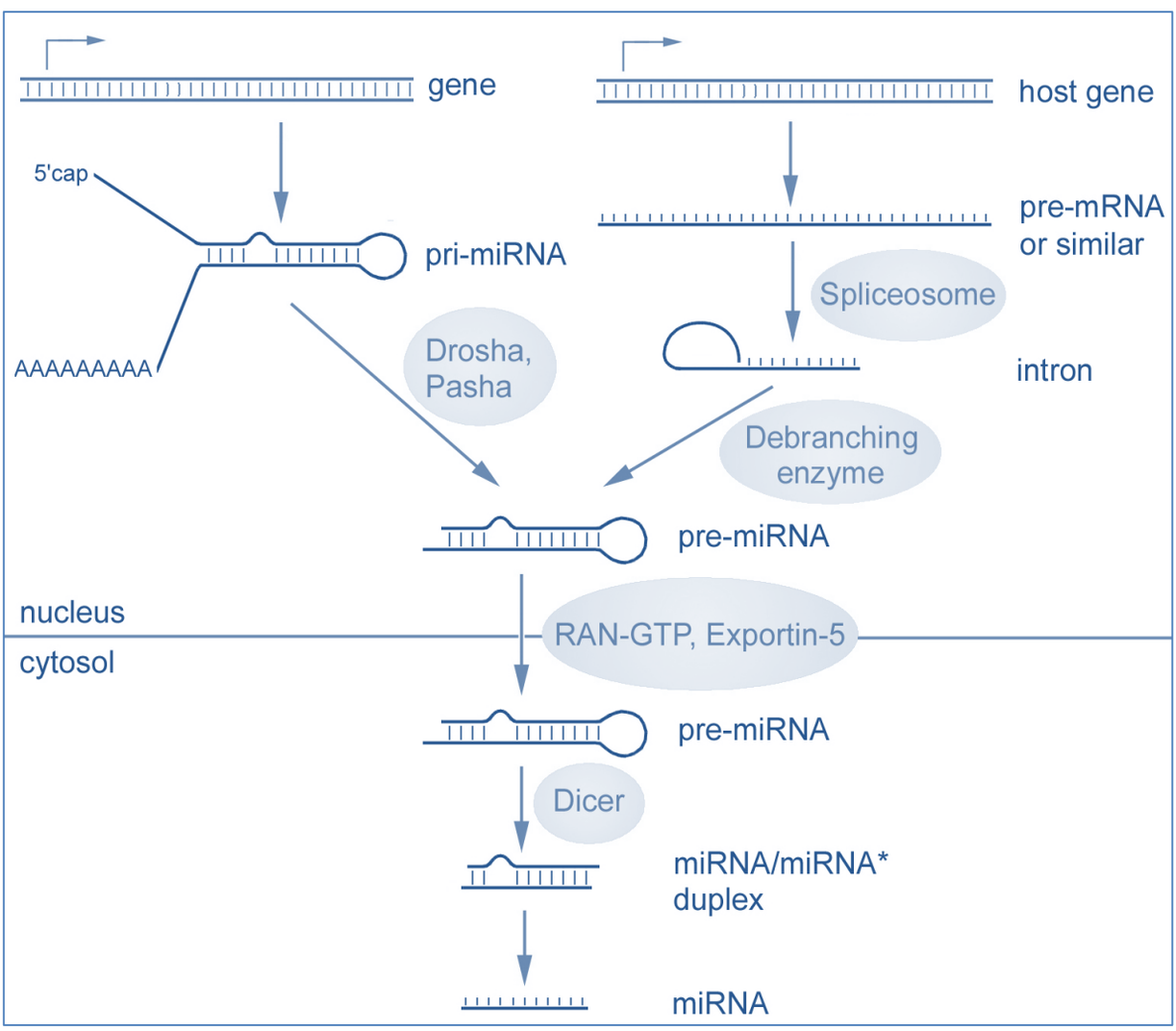


Abbildung 5: Die Biogenes von miRNAs.

Die Abbildung zeigt zusammenfassend wie miRNAs aus dem Genom abgeschrieben und prozessiert werden, bis hin zur reifen miRNA. Die Abbildung ist modifiziert von Narayanese übernommen.

(Encyclopedia of DNA Elements), die in mehreren Artikeln veröffentlicht wurden und in einem Science Editorial zusammengefasst worden sind, haben gezeigt, dass scheinbar 80% des Humanen Genoms eine Art der Funktion ausüben und daher kein „Junk“ sind [72]. Aber auch diese Arbeit wurde kontrovers diskutiert, so dass die Diskussion über Junk DNA bis heute nicht abschließend geklärt ist [73].

Für einen Teil des Genoms ist klar, dass er eine Funktion ausübt, obwohl keine Proteine in ihm kodiert werden: nicht-kodierende RNAs. Diese nicht-kodierenden RNAs, die in kurze und lange nicht-kodierende RNAs unterteilt werden, spielen eine essenzielle Rolle in der Genregulation. Eine der Unterklassen, die am meisten untersucht wurde, sind kleine nicht-kodierende RNAs, speziell miRNAs. In einem Artikel 2015 mit dem Titel „Junk DNA isn’t“ [74] hat Lin He die Rolle der nur 17-22 Nukleotide langen Moleküle in der Genregulation beschrieben und dabei die Bedeutung von miRNAs für die Biomedizinische Forschung zusammengefasst.

Entdeckt wurden miRNAs bereits zu Beginn der 1990er durch Lee, Feinbaum und Ambros [35]. Sie zeigten, dass es im Genom von *C. elegans* kurze „Gene“ gibt, die in RNA umgewandelt werden und die Expression anderer Gene unterdrücken. Das entsprechende Gen lin-4 wurde von den Autoren allerdings noch nicht mikroRNA genannt. Was Lee und seine Mitarbeiter herausgefunden haben, war dass es zwei Transkripte von lin-4 gibt, eines, dass 60 Nukleotide lang ist und eines, dass nur 20 Nukleotide lang ist. Diese entsprechen dem Precursor und reifer miRNA (siehe Abbildung 5 und Abbildung 6). Als Mechanismus wurde die Bindung an den 3’ untranslatierten Bereich von Genen (Untranslated Region; UTR) beschrieben. Außerdem wurde die typische Haarnadelstruktur für miRNAs veröffentlicht, wie sie auch in Abbildung 6 gezeigt ist. Der Begriff miRNA oder mikroRNA wurde erst 10 Jahre später geprägt [37]. Zu dieser Zeit war bereits viel über die Biogenese und Funktion bekannt. Aus dem Genom wird die sogenannte pri-miRNA abgeschrieben. Diese wird durch Drosha und Pasha zur pre-miRNA prozessiert. Mittels Exportin-5 wird diese aus dem Zellkern ausgeschleust. Das Enzym Dicer schneidet dann die zwei reifen Formen, die als -3p und -5p Form bezeichnet werden aus der pre-miRNA. Die Biogenese ist in der Übersicht in Abbildung 5 gezeigt. In der Terminologie werden die Precursor mit „mir-“ bezeichnet, während die reifen miRNAs mit „miR-“ gekennzeichnet sind. Da miRNAs sehr konserviert zwischen Organismen sind [75] werden sie üblicherweise noch mit drei Buchstaben, die den Organismus angeben, gekennzeichnet. Außerdem sind miRNAs in Klustern oder als Familien organisiert. Dabei enthält eine Familie mehrere sehr ähnliche Repräsentanten. Mitglieder einer miRNA Familie werden mit den Buchstaben „a“, „b“, ... voneinander abgegrenzt. Abbildung 6 zeigt ein Beispiel für eine der bekanntesten miRNAs aus der mir-34 Familie. Beim Menschen wird diese miRNA zum Beispiel mit „hsa-mir-34a“ bezeichnet

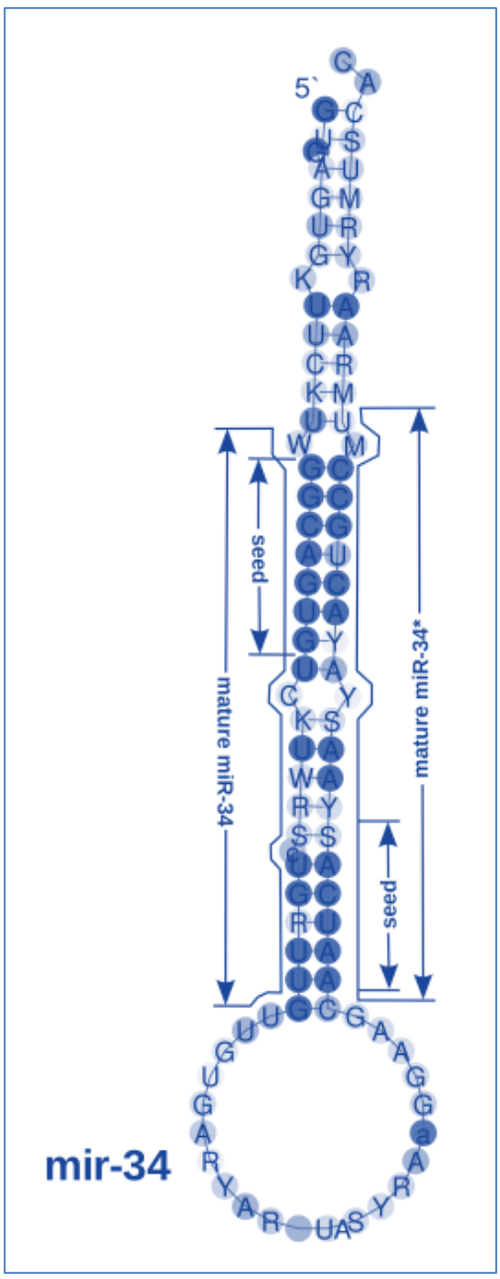


Abbildung 6: Haarnadel-Struktur der mir-34.

Die Abbildung zeigt die Haarnadel-Struktur der mir-34 gekennzeichnet sind die mature und die mature (*) Form der miRNA (entsprechend der -5p und -3p Form). Zusätzlich ist der Seed gekennzeichnet, die Region, die in der reifen miRNA maßgeblich für die Bindung an den UTR des Zielgenes ist. Die Abbildung wurde modifiziert von Paul Gardner übernommen.

und hat entsprechend die beiden reifen Formen „hsa-miR-34a-3p“ und hsa-miR-34a-5p. In Abbildung 6 ist die erste der beiden reifen Formen noch als (*) miRNA gekennzeichnet, diese Bezeichnung wird heute normalerweise nicht mehr verwendet. Abbildung 6 zeigt außerdem für die beiden reifen Formen noch ein wichtiges Detail, die sogenannte Seed Region. Diese Seed Region sind die 7 Basen, die für die Regulation der Genexpression am entscheidendsten sind [76]. Eine Übersicht über die Biogenese und die Funktion von miRNAs ist in Abbildung 7 gezeigt.

Die bekannten miRNA werden seit 2003 in der miRBase, die als Referenz Datenbank gilt, gespeichert [77]. Zwischen 2003 und 2014 wurden insgesamt 21 Versionen der miRBase veröffentlicht [78-83]. In diesen Versionen wurden – vor allem durch HTS Projekte – zunehmend größere Zahlen an miRNAs angegeben. Beim Menschen ist die Anzahl an reifen miRNAs zum Beispiel auf 2.500 angewachsen. Zusammengenommen entsprechen diese 2.500 miRNAs fast 0.002 Prozent des Humanen Genoms. Zusätzlich haben mehrerer Studien in HTS Experimenten mehrere tausend neue Kandidaten veröffentlicht [84-87], die noch nicht in der miRBase annotiert sind. Für viele der Kandidaten gibt es jedoch kaum eine oder gar keine Validierung, sodass Schätzungen davon ausgehen, dass bis zu 60% der Kandidaten in der miRBase und noch deutlich mehr in anderen Projekten auf Artefakte, zum Beispiel durch die Sequenzierung, zurückzuführen sind [88]. Besonders die späten Versionen der miRBase (allen voran Version 19, 20 und 21) sind für miRNAs angereichert, die eher in Proben von schlechter Qualität mit degradierter RNA gefunden wurden. In den frühen Versionen der miRBase (vor allem Version 1-12)

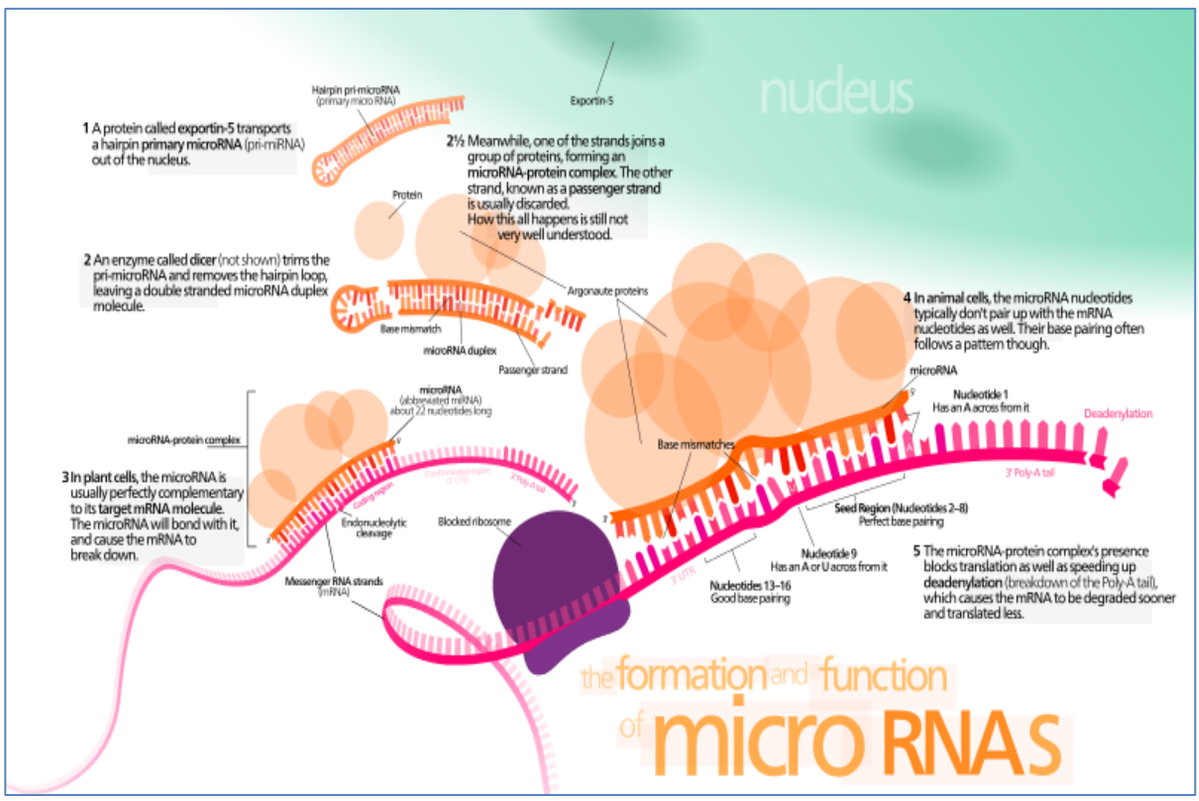


Abbildung 7: miRNA Biogenese und Funktion.

Die Abbildung die die Biogenese und die Funktion der Gen Regulation durch miRNAs übersichtlich darstellt wurde von https://upload.wikimedia.org/wikipedia/commons/thumb/a/a7/MiRNA.svg/800px-MiRNA.svg.png übernommen.

sind entsprechende, wahrscheinlich fehlerbehaftete, miRNAs zu einem deutlich geringeren Prozentsatz vertreten. Die miRNAs in diesen frühen Versionen wurden außerdem zu einem substanziell höheren Maß, mit anderen Techniken die auf Hybridisierung beruhen (siehe auch Kapitel 3), wie zum Beispiel Northern Blots, validiert.

Aufbauend auf diesen Erkenntnissen wurden in verschiedenen Ansätzen zum einen neue, sehr spezifische, aber auch sehr sensitive Datenbanken entwickelt. Die spezifischen Datenbanken zielen darauf ab, die wahrscheinlichsten Kandidaten mit dem höchsten Maß an Validität zu speichern. Das bekannteste Beispiel für eine entsprechende Datenbank ist die miRGeneDB von Fromm und Mitarbeitern [89]. Die sensitiven Datenbanken beinhalten neben den echt positiven miRNAs noch eine Vielzahl an potenziellen Kandidaten. Von diesen Kandidaten ist naturgemäß nur ein Bruchteil echt positiv, dennoch sind natürlich auch mehr valide Sequenzen enthalten als in den spezifischen Datenbanken. Entsprechende sensitive Lösungen sind noch aus einem anderen Grund wichtig: Da viele Kandidaten aus der Literatur nicht in Datenbanken abgelegt wurden, sind etliche doppelt oder dreifach publizierte Kandidaten mit verschiedenen Bezeichnungen bekannt. Die umfassendste Datenbank mit Sequenzen, die für kleine nicht-kodierende RNAs stehen, ist miRCarta [90].

In den vergangenen Jahren wurden mehr und mehr Datensätze veröffentlicht, die für den Menschen (und andere Organismen) kleine nicht-kodierende RNAs in verschiedenen Geweben, Zelltypen, Entwicklungsstadien und anderen Bedingungen nachweisen [91-101]. Ein Internetbasiertes Programm zur Auswertung von entsprechenden Datensätzen wurde in der AG von Prof Keller entwickelt, miRMaster [102]. Mit miRMaster wurden bisher 298 Experimente ausgewertet, insgesamt haben dies 27,344 Sequenzier-Proben enthalten und 345 Milliarden Reads wurden dabei prozessiert (Stand Februar 2018). Diese Menge an Sequenzier-Daten von kleinen nicht-kodierenden RNAs entspricht theoretisch der Masse an Nukleinsäuren die in 4.200 Humanen Genomen enthalten ist. Eine Meta-Analyse aller Datensätze hat ergeben, dass 874,123 Regionen über das Genom verteilt sind, die mit entsprechenden kurzen Fragmenten angereichert sind [103]. Diese Regionen mit einer mittleren Länge von 31 Nukleotiden entsprechen ungefähr 0.8% des Humanen Genoms und enthalten wahrscheinlich den Großteil aller existierenden miRNAs und weiterer regulatorischen Elemente.

In meiner Arbeit ziele ich auf die Entwicklung von Biomarkern zum Einsatz in der klinischen Diagnostik ab. Daher ist es notwendig sich von Anfang an auf valide Marker, am besten mit bekannter Funktion, zu konzentrieren. Wie im vorherigen Abschnitt beschrieben sind es vor allem die miRNAs aus den frühen Versionen der miRBase, die gut charakterisiert sind, wahrscheinlich am wenigsten Artefakte aufweisen und daher die geeignetsten Biomarker darstellen. Diese sind auch in Körperflüssigkeiten wie zum Beispiel Blut häufig vertreten [104] und bilden daher die bestmögliche Grundlage für die Entwicklung von nicht- oder minimal invasiven Markern zur Früherkennung von Erkankungen.

Kapitel 3

Methoden

In diesem Kapitel gehe ich auf die grundlegenden Methoden ein, die in den verschiedenen Forschungsprojekten in meiner Doktorarbeit angewendet wurden. Zunächst werden die experimentellen Techniken, die hauptsächlich zum Einsatz kommen, beschrieben: Mikroarrays und Hochdurchsatz-Sequenzierung (HTS). In diesem Abschnitt ist es vor allem wichtig die konzeptionellen Unterschiede und Einsatzgebiete der Technologien zu verstehen. Danach werden im zweiten Teil des Kapitels die grundsätzlich verwendeten Methoden der Biostatistik und Bioinformatik erwähnt.

In den einzelnen Resultatunterkapiteln sind spezielle Techniken, die spezifisch angewendet wurden, erläutert und es sind sowohl in diesem als auch im Resultatkapitel weiterführende Quellen mit Detailinformationen zu den jeweiligen Techniken angegeben.

3.1. Technologie

Mikroarrays: Mikroarrays sind üblicherweise zweidimensionale Träger aus Glas oder Silikon auf der in hoher Dichte Analyten aufgetragen sind. Zunächst wurden sie zum Messen von Antikörpern verwendet, die ersten Mikroarrays zu diesem Zweck wurden bereits 1983 vorgestellt [105]. Daneben gibt es viele verschiedene Arten von Mikroarrays, wie zum Beispiel Proteinarrays, Peptidarrays, Gewebearrays oder DNA Mikroarrays. Bei allen gängigen Mikroarrays ist der Inhalt, also die Analyten die gemessen werden sollen, vorher fest definiert. Im Falle von DNA Mikroarrays zum Messen der Genexpression bedeutet das, dass die Gene die nachgewiesen werden sollen bekannt sein müssen und komplementäre Fänger-Sonden zum Nachweis der Gene müssen auf dem Glas- oder Silikonträger an fest definierten Positionen immobilisiert werden. Üblicherweise werden mehrere zehntausend Transkripte gleichzeitig parallel nachgewiesen. Die Sonden können entweder auf der Oberfläche gespottet werden oder durch ein Synthese-Verfahren „in-situ“ aufgebracht werden. Das Mikroarray System, das ich konzipiert habe beruht auf der zweiten Technik und ist in Kapitel 4.1. im Detail beschrieben. Das Grundprinzip der Messung ist dann bei allen Methoden vergleichbar. RNA die mit einer

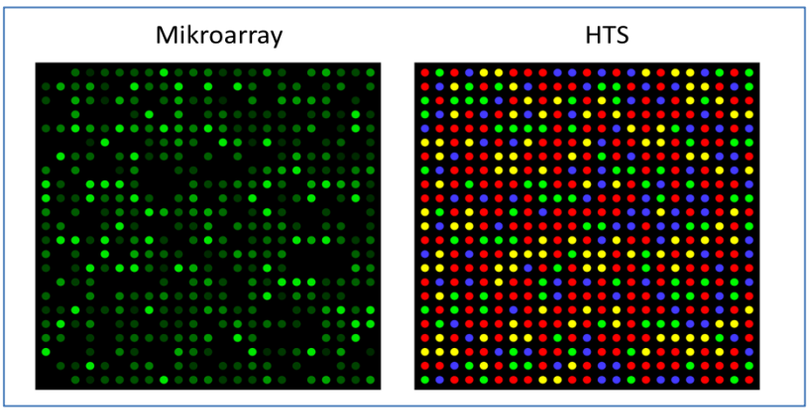


Abbildung 8: Mikroarray und HTS Flow Cell.

Die Abbildung zeigt links schematisch eine Oberfläche eines Mikroarrays (der Firma Affymetrix) und rechts eines Trägers (der Firma Illumina) wie er für HTS verwendet wird. Obwohl die Unterschiede auf den ersten Blick marginal aussehen, sind die Technologien prinzipiell unterschiedlich. Die Abbildung wurde modifiziert von Thomas Shafee übernommen.

Markierung versehen wurde oder mit einer Markierung auf dem Mikroarray versehen werden kann, wird auf den Array aufgebracht. In einem normalerweise mehrere Stunden dauernden Prozess binden die RNAs, die nachgewiesen werden sollen, an die Fänger Sonden, die vorher immobilisiert wurden. Dieser Schritt nennt sich Hybridisierung. Nach der Detektion mit Laser-Scannern oder CCD Kameras werden die Signale gemessen. Dabei ist eine der Kernherausforderungen, den dynamischen Bereich festzulegen. Wenn zu wenig Material gebunden wird, kann keine Intensität und kein Signal gemessen werden. Wenn zu viel eines bestimmten Genes vorhanden ist, ist die entsprechende Position auf dem Mikroarray gesättigt. Eine typische Aufnahme eines Mikroarrays, hier von der Firma Affymetrix, ist im linken Teil von Abbildung 8 gezeigt. Je intensiver ein Punkt, der einem bestimmten Gen entspricht, leuchtet, um so mehr des Genes war in der Ausgangsprobe vorhanden. Neben den oben beschriebenen technischen Herausforderungen kommen weitere hinzu, wie zum Beispiel die Vergleichbarkeit zwischen Experimenten die durch geeignete Normalisierungsmethoden sichergestellt werden muss.

Ein detaillierter Überblick über Mikroarray Technologie für verschiedene Analyten wurde 2006 von Barbulovic-Nad et al. publiziert [106]. Ein umfangreicher Übersichtsartikel zur bioinformatischen Auswertung von Mikroarrays wurde von Wang veröffentlicht [107].

HTS: Im Vergleich zu Mikroarrays ist es bei der Hochdurchsatz-Sequenzierung (HTS) nicht notwendig vor Versuchsdurchführung festzulegen, welche Gene gemessen werden sollen. Unabhängig ob DNA nachgewiesen wird, um Einzelbasenaustausche oder andere genetische Veränderungen zu entdecken oder ob die Expression von Genen gemessen werden sollen. Zur HTS gibt es grundlegend verschiedene Ansätze wie die Sequenzierung durch Hybridisierung, Sequenzierung durch Ligation, Sequenzierung durch Synthese oder Sequenzierung mittels Halbleitertechnologie. Die einzelnen Methoden im Detail zu beschreiben liegt nicht im Fokus dieser Arbeit; eine geeignete Übersicht wurde von Liu und Mitarbeitern publiziert [108]. In allen Fällen werden Methoden verwendet um parallel eine Vielzahl von Nukleinsäure-Ketten zu messen. Im Falle der am häufigsten verwendeten Technologie, der Sequenzierung durch

Synthese wie sie von der Firma Solexa entwickelt wurde und von Illumina seit mehr als einem Jahrzehnt eingesetzt wird, werden auf einem Glasträger gebundene Moleküle, die sequenziert werden sollen, schrittweise von Einzelsträngen in Doppelstränge umgewandelt.

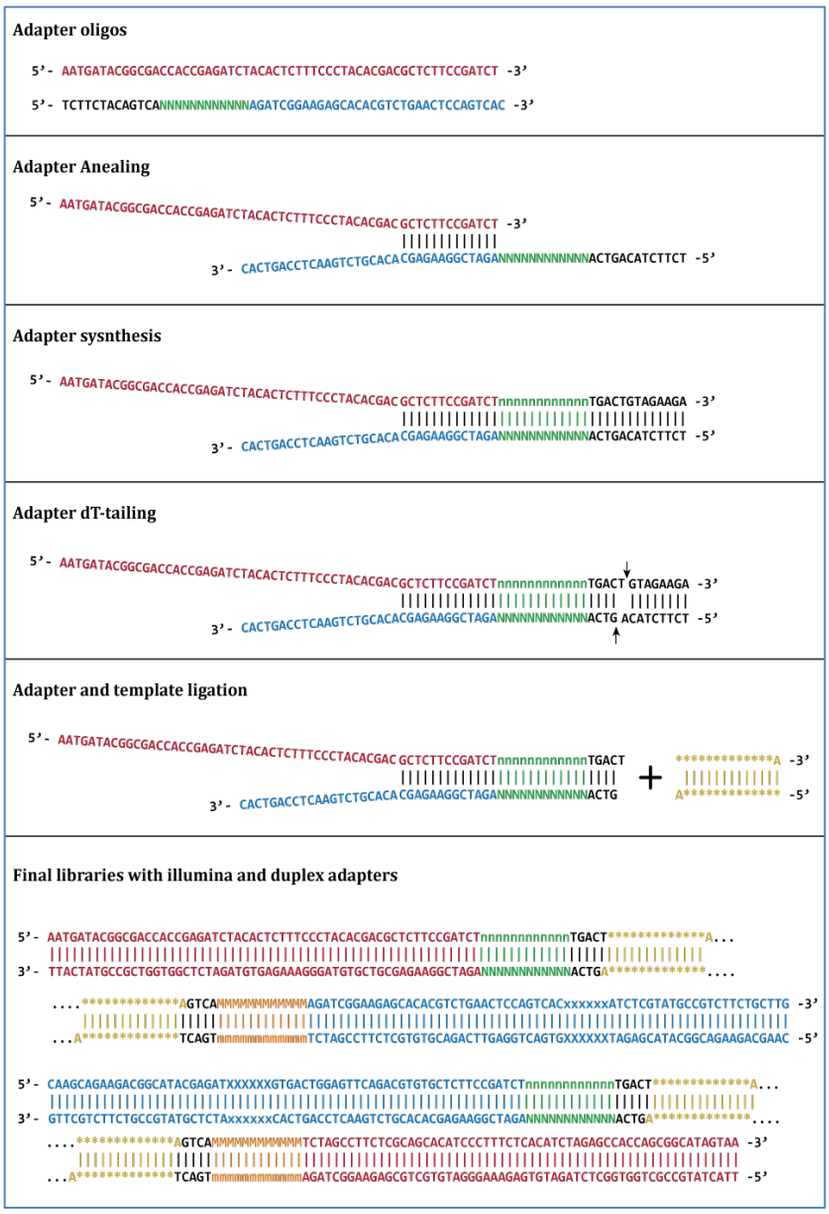


Abbildung 9: Schema des Herstellens der Sequenzier Bibliothek.

Die Abbildung zeigt schematisch, wie eine Sequenzier-Bibliothek für das besonders genaue Duplex-Sequenzier-Verfahren hergestellt wird. In Wirklichkeit besteht das experimentelle Protokoll von der Probe bis hin zur Ausgabedatei (fasta) aus mehreren Dutzend Schritten. Die Abbildung ist von https://upload.wikimedia.org/wikipedia/commons/2/23/Duplex_sequencing_library_preparation_procedure.svg entnommen.

Wann immer eine komplementäre Base eingebaut wird, wird ein Lichtsignal gemessen. Grundsätzlich ist die Idee den Mikroarrays sogar ähnlich, nur dass das nachzuweisende Molekül und nicht eine Fänger-Sonde auf dem Glasträger immobilisiert ist. Eine sogenannte Flow Cell wie sie zum Sequenzieren verwendet wird ist im Vergleich zum Mikroarray auf der rechten Seite von Abbildung 8 gezeigt.

Die Methodik die im Labor eingesetzt wird ist momentan noch komplexer als bei Mikroarrays. Ein Protokoll nur zur Herstellung der Sequenzier-Bibliothek, also der Nukleinsäuren und Adapter die nachgewiesen werden sollen, ist schematisch in Abbildung 9 gezeigt. Detaillierte Anwendungs-protokolle zur Sequenzierung bestehen normalerweise aus mehreren Dutzend Einzel-schritten. Eine weitere Herausforderung von HTS ist die wesentlich komplexere Datenauswertung. Wie schon in Abbildung 2 in der Einleitung dargestellt sind HTS Datensätze

in der Regel drei bis vier Größenordnungen umfangreicher als entsprechende Mikroarray Datensätze. Zur Analyse der Datensätze, die mehrere Giga- bzw. je nach Ausmaß der Studie etliche terabyte Daten umfassen, wurden sehr viele verschiedene Auswerte-Pipelines veröffentlicht [109-117]. Diese haben jedoch teilweise sehr unterschiedliche Resultate gezeigt, selbst wenn die selben Eingabedaten verwendet wurden. Die Heterogenität und Variabilität dieser Bioinformatik Analyse Pipelines hat die Gesellschaft für Pathologie der USA und die Gesellschaft für Medizininformatik in den USA veranlasst, einen „Best Practice Guide“ für entsprechende Software Tools zu veröffentlichen [118].

Ein Vergleich von Mikroarrays und HTS, der weit über die Hintergrundinformation in diesem Abschnitt hinaus geht, speziell im Umfeld Mikrobiologie, wurde von Roh et al. publiziert [119]. Für die in dieser Arbeit zentralen miRNAs haben Willenbrock und Mitarbeiter die beiden Methoden verglichen [120]. Neben diesen Reviews hat Mestdagh speziell für die Analyse von miRNAs 13 verschiedene Methoden systematisch evaluiert, darunter auch Mikroarrays und HTS [121]. Diese Arbeit bietet den momentan vollständigsten Überblick über verfügbare Methoden und Techniken zur Quantifizierung von miRNAs.

3.2. Bioinformatik und Biostatistik

In diesem Abschnitt stelle ich kurz grundlegende statistische Methoden vor. Diese dienen dazu einen Überblick zu geben und ersetzen keinesfalls Fachliteratur. Für detaillierte Beschreibungen so wie Formeln zu den verwendeten Tests habe ich mich am Fachbuch „Probability and Statistics“ von DeGroot und Schervish aus dem Adison Wesley Verlag orientiert. Alle statistischen Analysen sind in der frei verfügbaren ***R*** Entwicklungsumgebung durchgeführt worden.

Test auf Normalverteilung: In meiner Doktorarbeit wurden in mehreren Teilprojekten Daten erhoben die statistisch ausgewertet werden müssen. Oft sind es paarweise Gruppenvergleiche die zum Einsatz kommen, um beispielsweise die Hypothese zu testen, dass die Mittelwerte zweier Gruppen unterschiedlich sind. Ein Test der in der Biologie und Medizin oft zum Einsatz kommt, ist der im nächsten Absatz beschriebene T-test [122-126]. Dieser parametische Test wird allerdings oft falsch angewendet [127]. Eine der grundlegenden Annahmen ist, dass die Ausgangsdaten normalverteilt sind. Daher ist es zunächst notwendig einen Test auf Normalverteilung durchzuführen. Um beispielsweise bei Genexpressionsdaten auf Normalverteilung zu testen, existieren verschiedene Möglichkeiten [128]. Wir haben in der Regel den Shapiro-Wilk Test angewendet, der die Hypothese überprüft, dass die zugrunde liegende Grundgesamtheit einer Stichprobe

normalverteilt ist. Wenn der Test auf Normalverteilung positiv war wurde der T-Test angewendet, ansonsten wurde der nicht-parametrische Mann-Whitney-U Test angewendet, der ebenfalls unten beschrieben ist. Es wurde die Implementierung des Tests in ***R*** im „Stats“ Paket verwendet (Funktion *shapiro.test*).

T-Test: Als grundlegender Test auf Unterschiede im Mittelwert von zwei Gruppen wurde der t-Test eingesetzt, der häufig in biomedizinischen Fragestellungen angewendet wird [122-126]. Wenn nicht explizit erwähnt, wurde der T-test als zweiseitiger ungepaarter Test durchgeführt, unter der Annahme, dass die Standardabweichung beider Gruppen identisch ist. Bei ungleicher Varianz kann der Welch-Test als Alternative verwendet werden. Die Nullhypothese des t-Tests ist, dass die Mittelwerte der beiden zu testenden Grundgesamtheiten identisch sind. Die Alternativ Hypothese ist, dass die Mittelwerte der beiden Grundgesamtheiten voneinander abweichen. Es wurde die Implementierung des Tests in ***R*** im „Stats“ Paket verwendet (Funktion *t.test*).

Mann-Whitney-U-Test: Der Mann-Whitney-U-Test (Wilcoxon Rangsummen-Test, Wilcoxon-Mann-Whitney Test WMW) testet für unabhängige Stichproben, ob zwei Verteilungen übereinstimmen, also ob die beiden zugrundeliegenden Verteilungen zu derselben Grundgesamtheit gehören. Der Mann-Whitney-U-Test wird dann verwendet, wenn die Voraussetzungen für einen t-Test für unabhängige Stichproben nicht erfüllt sind. Er wird ebenfalls gängig in der Biostatistik eingesetzt [129]. Im Fall von Daten mit vielen „Ties“ wurde die am Lehrstuhl von Prof. Keller entwickelte exakte Lösung des WMW Tests angewendet, die auf dynamischer Programmierung beruht [130]. Wenn nicht explizit erwähnt, wurde als Standard ein zweiseitiger WMW Test für nicht gepaarte Analysen verwendet. Es wurde die Implementierung des Tests in ***R*** im „Stats“ Paket verwendet (Funktion *wilcox.test*).

Adjustieren für Multiples-Testen: P-Werte aus den oben genannten Tests basieren auf der Annahme, dass eine Hypothese getestet wurde. Der p-Wert ist dabei eine Wahrscheinlichkeit, die zwischen 0% und 100% (respektive zwischen 0 und 1) liegen kann. Der p-Wert gibt dabei an, wie wahrscheinlich es ist, ein Stichprobenergebnis wie das vorliegende oder ein noch extremeres Stichprobenereignis zu erhalten, wenn die Nullhypothese wahr ist. Im Falle von Hochdurchsatz-Methoden wird allerdings prinzipiell eine wesentlich höhere Zahl an Hypothesen getestet, für jedes Gen / Protein / miRNA eine eigene. In diesem Fall ist die Wahrscheinlichkeit, dass ein möglicher, aber tatsächlich nicht vorhandener Unterschied erkannt wird höher. Diese Fehler werden als Fehler 1. Art bezeichnet. Je mehr Hypothesen getestet werden, desto geringer wird gleichzeitig die Wahrscheinlichkeit, dass ein tatsächlicher vorhandener Unterschied erkannt wird, es entstehen Fehler 2. Art. Um Multiples-Testen zu korrigieren, können verschiedene

Ansätze gewählt werden. Der wohl einfachste ist die Bonferroni Korrektur. Dabei wird entweder das Alpha Fehlerniveau auf (0,05 / Anzahl an Tests) herabgesetzt oder alternativ die erhaltenen p-Werte mit der Anzahl an Tests multipliziert. Im zweiten Fall werden adjustierte p-Werte größer als eins auf eins gesetzt. Wenn nicht explizit erwähnt, sind p-Werte in meiner Arbeit für Multiples-Testen korrigiert. Allerdings wurde nicht die Bonferroni Korrektur verwendet, sondern der Ansatz zum Kontrollieren der False Discovery Rate (FDR) von Benjamini und Hochberg. Es wurde die Implementierung des Tests in ***R*** im „Stats" Paket verwendet (Funktion *p.adjust*).

AUC / ROC Analyse: Eine weitere Analyse die oft zur Analyse der Qualität von Biomarkern verwendet wird (zum Beispiel in [131-136]) ist die Interpretation der Receiver-Operating-Characteristic-Kurve (ROC Kurve). In einem Diagramm wird die Sensitivität (Richtig-Positiv-Rate) als Ordinate und die Falsch-Positiv-Rate als Abszisse aufgetragen. Das Gütemaß ist dann die Fläche unter der ROC-Kurve, die Area Under Curve (AUC). Der Wert der AUC kann wischen 0 und 1 liegen. Es ist wichtig hervorzuheben, dass 0,5 der schlechteste mögliche Wert ist, da dieser zu einer ROC Kurve nahe der Diagonale und daher nahe des erwarteten Ergebnisses eines Zufallsprozesses liegt. Die normalerweise als optimal beschriebene Kurve hat eine Fläche größer 0,5 und möglichst nahe an 1. Eine Kurve mit einer Fläche kleiner 0,5 und nahe an 0 ist vom Informationsgehalt her allerdings genauso gut. Ein Beispiel ist die Hoch- und Runterregulation von Genen. Ein perfekt hochreguliertes Gen hat eine AUC von 1, ein perfekt runterreguliertes Gen einen AUC Wert von 0. Interessant ist ebenfalls, dass sich der p-Wert des WMW Tests aus dem AUC Wert ableiten lässt. In meiner Arbeit wurde die AUC nicht nur verwendet, um die Güte von einzelnen miRNAs abzuschätzen, sondern auch, um die Performance von maschinellen Lernverfahren, speziell von den weiter unten beschriebenen Support Vector Machines, zu evaluieren. Es wurde die Implementierung der ROC Analyse in ***R*** im „ROC" Paket verwendet (Funktion *AUC*).

Varianzanalyse: Wenn mehr als nur zwei Gruppen miteinander verglichen werden, zum Beispiel Multiple Sklerose, Alzheimer, Parkinson und Kontroll-Probanden, kann eine Varianzanalyse (Analysis of Variance, ANOVA) angewendet werden. Auch die ANOVA wird seit mehreren Jahrzehnten in der Biostatistik zur Beurteilung von Biomarkern verwendet [137]. Generell ist die Varianzanalyse eine allgemeine Methode zur statistischen Bewertung von Unterschieden in Mittelwerten zwischen mehr als zwei Gruppen. In der einfachsten Form kann die ANOVA als Generalisierung des t-Tests angesehen werden. Es gelten die drei Grundannahmen, dass die Stichproben unabhängig sind, alle Stichproben sind normalverteilt und es herrscht Varianzhomogenität. Die Nullhypothese lautet, dass kein Unterschied zwischen den Mittelwerten der zu testenden Gruppen vorliegt, die Alternativhypothese besagt dementsprechend, dass zwischen

mindestens zwei Mittelwerten ein Unterschied besteht. In meiner Arbeit wurden one-way ANOVA in ***R*** im „stats“ Paket verwendet (Funktion *aov*).

Hierarchisches Clustern: Clusteranalysen werden eingesetzt, um Strukturen in Datensätzen zu erkennen. Eines der gängigsten Cluster Verfahren, das in der biomedizinischen Forschung eingesetzt wird und sich seit etlichen Jahrzehnten bewährt hat, ist hierarchisches Clustern [138-143]. Im Grunde bezeichnet hierarchisches Clustern eine Klasse von Verfahren die distanz- oder ähnlichkeitsbasiert sind. In meiner Arbeit habe ich hauptsächlich hierarchische Cluster Methoden, basierend auf der Euklidischen Distanz verwendet. Der verwendete Ansatz entspricht einem Bottom-Up Clustern. Jedes Objekt (als zu clusternde Objekte werden sowohl Gene / miRNAs als auch Probanden verwendet) bildet initial einen eigenen Cluster und ähnlichste Cluster werden in jedem Schritt iterativ zusammengefügt. Um die Ähnlichkeit zwischen zwei Clustern zu definieren wurde „complete linkage“ Clustering verwendet. Als grafische Ausgabe des Prozesses werden sogenannte Dendrogramme generiert. Sie verbinden in einer baumartigen Struktur die jeweils ähnlichsten Objekte. Je näher an der Wurzel des Baumes zwei Objekte zusammenkommen, um so unähnlicher sind sie. Umgekehrt, je näher an den Blättern Objekte zusammentreffen, um so ähnlicher sind sie. Als weitere grafische Darstellung werden Heat Maps generiert. Heat Maps sind Matrizen, die beispielsweise für jedes gemessene Gen oder jede gemessene miRNA eine eigene Zeile besitzen und für jeden Probanden eine eigene Spalte. Oft werden Dendrogramme für Patienten und Gene zusammen mit der Heat Map gezeigt. Durch das Clustering versucht man, Strukturen zu finden, die bei einer bestimmten Gruppe der Probanden (zum Beispiel den Patienten) anders sind als bei einer anderen Gruppe (zum Beispiel Kontrollen). Wenn nicht explizit erwähnt, wurde „unsupervised“ Clustering verwendet. Die Strukturen wurden gefunden, ohne dass man die Information verwendet hat welches Individuum zu den Patienten oder den Kontrollen gehörte. Um komplexe Signaturen grafisch darzustellen wurde „supervised“ Clustering verwendet. Dies ist in jedem Fall explizit erwähnt und dient wie beschrieben nur der grafischen Darstellung von Signaturen in meiner Arbeit. Um die Profile zu clustern habe ich in ***R*** das „stats“ Paket verwendet (Funktion *hclust*). Um die grafische Darstellung als Heat Map zu erzeugen wurde die *heatmap.2* Funktion verwendet.

Klassifikation: Um Patienten und Kontrollen basierend auf miRNA Mustern zu unterscheiden wurden zusätzlich „supervised“ Klassifikations Verfahren eingesetzt. In diesem Zusammenhang ist es wichtig zwei Komponenten zu erwähnen: „feature selection“ und „cross validation“. Mit Methoden des maschinellen Lernens wird versucht, Objekte basierend auf Eigenschaften in Klassen zuzuordnen. Wir haben uns mit einem vergleichsweise einfachen Fall beschäftigt, dem Aufteilen von Probanden in zwei Klassen,

basierend auf miRNA Mustern. Klassifikatoren haben zunächst basierend auf einem Datensatz Muster gelernt. Anschließend wurden andere Profile in den trainierten Klassifikator gegeben um für diese vorherzusagen zu welcher Klasse sie gehören. Da ich keinen eigenen Trainings- und Testdatensatz zur Verfügung hatte, habe ich sogenanntes „re-sampling“, genaugenommen Kreuzvalidierung („cross validation“), verwendet. Dabei wird der gesamte Datensatz zufällig in k gleichgroße disjunkte Mengen aufgeteilt (k wurde auf 10 gesetzt). K-1 Teile des Datensatzes wurden verwendet, um den Klassifikator zu trainieren und um den k-ten Teil, der nicht verwendet wurde, vorherzusagen. Jeder Proband wurde folglich in neun Trainings Sets verwendet und einmal selbst klassifiziert. Da der Prozess stochastisch ist und eine Zufallskomponente birgt, wurde das Verfahren für jede Klassifikation mindestens zehnmal wiederholt. Zusätzlich war es für den Klassifikator wichtig, die Parameter (hier die miRNAs) zu erkennen, die den größtmöglichen Nutzen haben und die beste Trennung erlauben. Dazu wurde eine „stepwise-forward“ Filter Subset Selektion angewendet. In jedem Schritt wurden innerhalb der Kreuzvalidierung die miRNAs gewählt, die die höchste Signifikanz auf dem momentanen Trainings-Set hatten. Beginnend mit zwei miRNAs wurde die Anzahl schrittweise erhöht und bis zu 250 miRNAs in die Klassifikation eingeschlossen. Um für potenzielles „overfitting“ zu testen wurden nicht-parametrische Permutationstests durchgeführt. Als Klassifikatoren wurden verschiedene Standard Lernverfahren getestet. Die besten Ergebnisse wurden generell mit Support Vector Machines mit Radialer Basis Funktion als Kernel erzielt. Die Methodik der Klassifikation ist in der Alzheimer miRNA Publikation ausgeführt [57] und ein exzellenter Hintergrund über die verwendeten Verfahren des maschinellen Lernens findet sich im Buch "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" von Trevor Hastie, Robert Tibshirani und Jerome Friedman.

Kapitel 4

Resultate: Von der Plattform zum Biomarker

In diesem Kapitel stelle ich die wesentlichen wissenschaftlichen Ergebnisse meiner Arbeit vor. Wie in der Einleitung skizziert befasse ich mich im ersten Teil mit der Entwicklung von Plattformen und Assays, die es uns erlauben Nukleinsäuren in hohem Durchsatz, parallel und sehr exakt zu messen (Kapitel 4.1). Ebenfalls im Kapitel über technologische

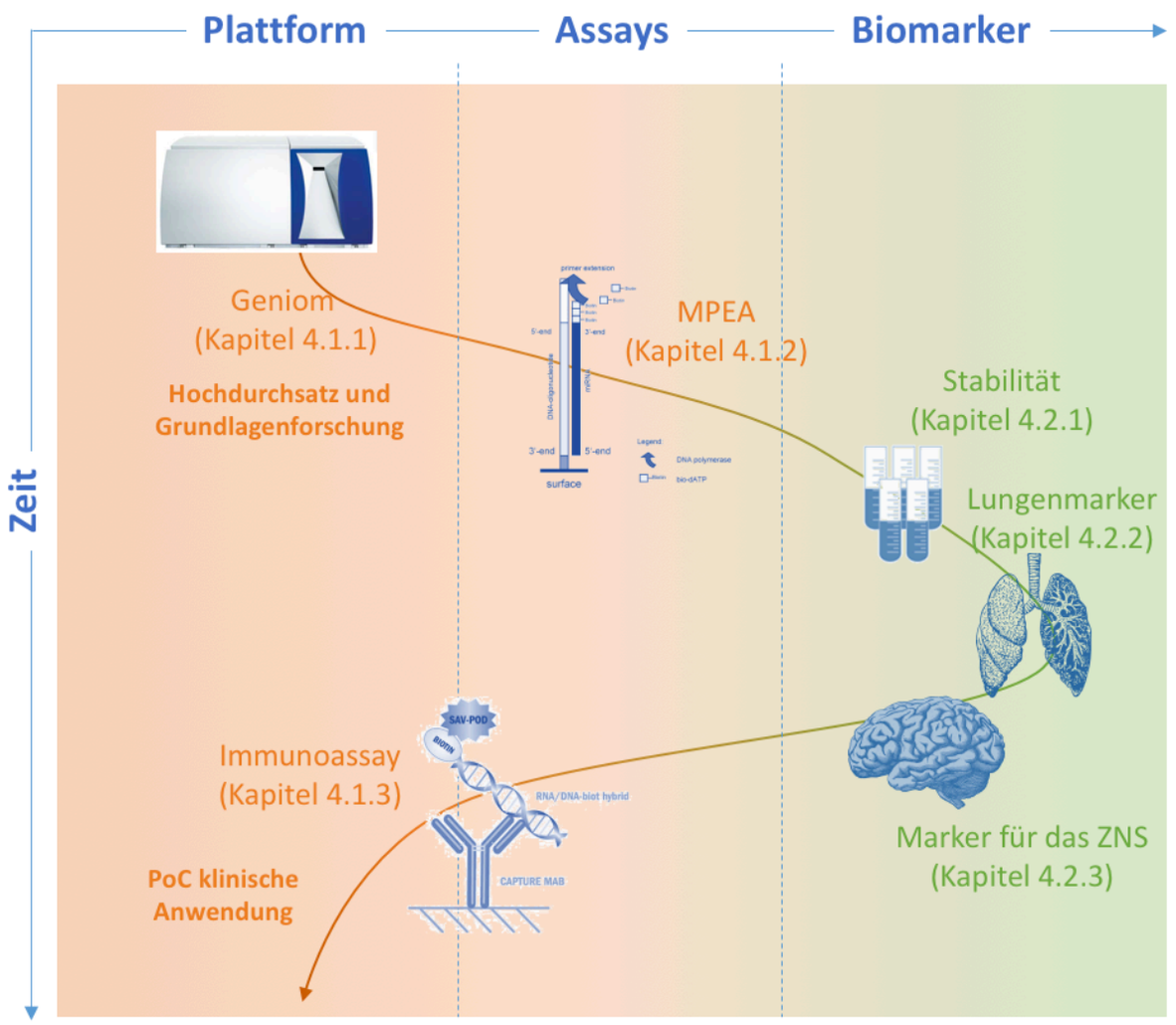


Abbildung 10: Übersicht über die Forschung.

Die Entwicklungen, die in dieser Arbeit beschrieben werden sind in obiger Abbildung übersichtlich dargestellt. Die technischen und Assay-Entwicklungen in Kapitel 4.1 sind in orange dargestellt, die Biomarker Entwicklung in grün. Auf der Assay Seite ist die Entwicklung vom grundlagenwissenschaftlichen Hochdurchsatz-Gerät bis hin zur patientennahen (Point-of-Care PoC) Testung fortgeschritten.

und Assay Entwicklung beschreibe ich meinen Ansatz molekulare Biomarker Signaturen mit Standard-Methodik, die in zehntausenden Krankenhäusern vorhanden ist, in der Routine zu messen. Obwohl die Techniken, die beschrieben werden, für alle Arten von Nukleinsäuren geeignet sind, habe ich in meiner Arbeit einen starken Fokus auf kleine nicht-kodierende RNAs, miRNAs gesetzt (siehe auch Kapitel 2). Daher skizziere ich im 2. Teil der Resultate, wie die verschiedenen Plattformen eingesetzt werden, um miRNA Biomarker zu finden (Kapitel 4.2). Danach befasse ich mich mit Eigenschaften von miRNAs, die über die rein deskriptive Korrelation der Biomarker mit Erkrankungen hinausgehen (Kapitel 4.3). Im letzten Abschnitt gehe ich noch auf Aspekte und Anwendungen in der Synthetischen Biologie ein (Kapitel 4.4). Die einzelnen Unterkapitel in der Abhängigkeit zueinander und in ihrer zeitlichen Entwicklung sind in Abbildung 10 zusammengefasst.

In den einzelnen Kapiteln verweise ich jeweils kurz auf Publikationen und Patente zu den jeweiligen Themen. Wie in der Einleitung beschrieben ist komplexe und interdisziplinäre Forschung nicht ohne entsprechend interdisziplinäre Kooperationspartner möglich. Die Beiträge der einzelnen Partner, die ich sehr zu schätzen weiß, sind in den entsprechenden Originalarbeiten gekennzeichnet.

4.1. Technsiche Plattformen und Assays

4.1.1. Das Geniom

Ein Nachteil der frühen Mikroarray Technologie, Ende der 1990 Jahre, war die geringe Flexibilität. Bevor ich die Idee zu einem flexiblen Mikroarray System hatte, hatten andere Firmen wie Affymetrix bereits kommerzielle Mikroarray Produkte auf dem Markt. Allerdings war der Inhalt der Arrays von den Firmen vorgegeben und „custom“ Produkte mit eigenem Inhalt waren teuer und haben mehrere Wochen bis Monate in der Herstellung benötigt. Mein Ziel war und ist es, Forschern eine höhere Flexibilität zu ermöglichen. Diese Flexibilität sollte es nicht nur erlauben eigene Gen Expressionsarrays über Nacht im eigenen Labor herzustellen, sondern auch viel weitreichendere Anwendungen zu ermöglichen. Beispiele dafür sind sogenanntes targeted Next-Generation-Sequencing, wie es heute in der Rutine Diagnostik eingesetzt wird. Dabei wird ein Set von Genen definiert, die auf einem Mikroarray oder in Lösung angereichert werden, sodass gerade die Fraktion der interessanten Gene sequenziert und ausgewertet werden kann.

Eine andere Anwendung, die in Kapitel 4.4. beschrieben wird, ist die Synthetische Biologie. Forschern sollte es ohne Weiteres möglich sein, möglichst fehlerfreie Oligonukleotide herzustellen. Die Herausforderungen an eine entsprechende Technologie waren dementsprechend groß. Standard Technologie zu verwenden, das Spotten von Mikroarrays oder aufwendige photolithografische Verfahren, wie es zum Beispiel von Affymetrix eingesetzt wurde, war entsprechend nicht möglich.

Die Lösung, die ich konzipiert habe basiert anstatt dessen auf einer in-situ Oligonukleotid Synthese die durch Licht aktiviert wird. In ein Glas-Silikon-Glas Sandwich werden dabei zunächst bis zu acht Kanäle – für maximal acht parallele Experimente – geätzt. Unter Verwendung von Standard-Synthese Chemikalien (Proligo) und 3' Phosporamidite mit einem photolabilen 5' Schutz wird parallel in allen Kanälen eine Synthese von vordefinierten Oligonukleotiden durchgeführt. Erwähnenswert ist, dass die Oberfläche durch einen sogenannten Spacer zugänglich gemacht werden muss. Dieser Set-up ermöglicht es, die Synthesezeit unabhängig von der Anzahl von Sonden zu machen. Jedoch wächst die Synthesezeit linear mit der Länge der Oligonukleotide an. Über Nacht ist es so möglich, fast 100,000 verschieden 25- bis 50-mere auf einem Mikroarray herzustellen. Diese können entweder identisch oder völlig unterschiedlich sein.

Das grundlegende Verfahren ist prinzipiell ähnlich zu der von Solexa entwickelten und von Illumina vermarkteten Sequenzierung durch Synthese. Zur Anwendung bei der Messung von Gen- oder miRNA Expression werden die Mikroarrays mit der entsprechenden RNA der Probe hybridisiert. Dazu wird die totale RNA mit oligo(dT) Primern in doppelsträngige cDNA umgeschrieben. Diese wird mit Phenol-Chloroform aufgereinigt und mit Ethanol gefällt. Mit T7 RNA Polymerase (Ambion) wird basierend auf der Vorlage der cDNA eine in-vitro

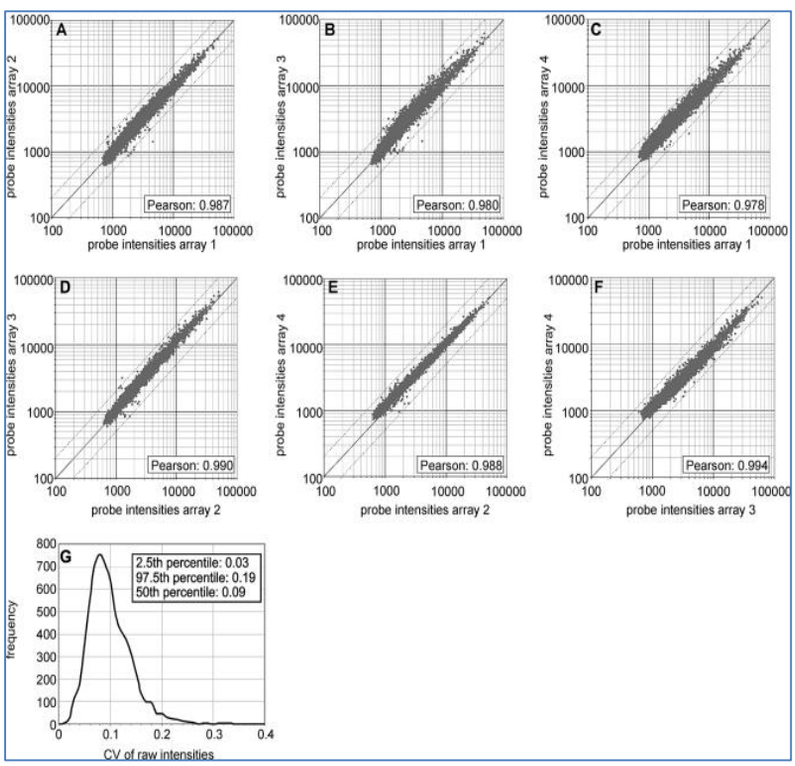


Abbildung 11: Reproduzierbarkeit der Geniom Array Technologie.

*Für 4 Mikroarrays werden alle 4*3/2 paarweisen Kombinationen von Scatter-Plots gezeigt (A-F). Panel G zeigt eine Dichteverteilung des Variationskoeffizienten. Die Abbildung entstammt aus unserem Manuskript Baum et al.*

Transkription durchgeführt. In der Lösung waren nicht markierte ATP, CTP, GTP und UTP sowie Biotin markierte CTP und UTP Moleküle.

Im Folgenden werden die eigens hergestellten Mikroarrays mit der so vorbereiteten Probe hybridisiert. Entscheidend ist, dass sowohl die Herstellung der Mikroarrays als auch die Hybridisierung der Probe im selben Gerät erfolgen. Die Mikroarrays werden dabei mit 15 Mikrogramm fragmentierter cRNA in 20 Mikroliter Lösung hybridisiert. Die Inkubationszeit beträgt dabei 16 Stunden bei konstant 45 Grad Celsius. Nach 20-minütigem Waschen mit Pufferlösung wird für 15 Minuten ein Streptavidin Fluoreszenzfarbstoff hinzugegeben. Die Signale werden mit einer CCD Kamera ausgelesen und mittels Bildverarbeitung quantifiziert. Als Signalintensität kann entweder die absolute Menge verwendet werden oder das Verhältnis von „perfect Match“ zu „miss Match“ Sonden, also Sonden, in die gezielt Veränderungen eingebaut wurden.

In einer ersten Studie konnte gezeigt werden, dass die Mikroarray Technologie, die ich maßgeblich entwickelt habe, sowohl sensitiv als auch reproduzierbar ist, einen hinreichend großen dynamischen Bereich (Dynamik Range) bietet und sich mit anderen Methoden gut vergleichen lässt. Spike In Experimente in Konzentrationsreihen beginnend bei 23 pikomolarer Lösung bis zu 100 nanomolarer Konzentration haben gezeigt, dass die gemessene Intensität mit der tatsächlichen Konzentration linear korreliert. Der dynamische Bereich war dabei drei Größenordnungen. Neben den Untersuchungen über die Sensitivität und den dynamischen Bereich zeigt der Mikroarray auch eine hohe technische Reproduzierbarkeit: der mittlere Korrelationskoeffizient bei wiederholten Messungen liegt bei 0.99 und der Variationskoeffizient liebt bei 9%. Selbst im Vergleich zu Affymetrix Mikroarrays wurden noch Korrelationswerte über 0.9 erreicht. Die hohe Reproduzierbarkeit und der geringe Variationskoeffizient sind in Abbildung 11 gezeigt.

Die Geniom Technologie ist in allen Kennzahlen gleichwertig zu anderen Array Technologien, sowohl von Affymetrix als auch von Agilent oder Illumina. Zur Messung von Genexpressionsmustern wird heute neben Mikroarrays auch HTS eingesetzt. Die HTS Methode bietet den Vorteil, dass Gene nicht nur quantifiziert werden können, sondern auch, dass verschiedene Splice Formen und Mutationen in Genen gefunden werden können. Daher wurde in den vergangenen Jahren zunehmend auf die HTS Technologie gesetzt und für das klassische Gen Expression Profilierung haben Mikroarrays an Bedeutung verloren. Exzellente Übersichtsartikel und direkte Vergleiche über die beiden teilweise konkurrierenden, teilweise aber auch komplementären Technologien wurden von Su und Mitarbeitern sowie Zhao und Mitarbeitern publiziert [144, 145].

Die Geniom Technologie, die ich in diesem Abschnitt beschrieben habe, ist aber keinesfalls überflüssig, sondern wurde für drei Anwendungen weiterentwickelt. Ich habe die Technologie verwendet, um kleine nicht-kodierende RNAs zu untersuchen. Im Gegensatz zu Genexpression ist es hier wesentlich weniger wichtig, Mutationen zu finden. Bei den im Mittel gerade 22 Basen langen RNAs, die nicht für Proteine kodieren, existieren außerdem keine Splicevarianten. Dafür ist eine genaue Quantifizierung, wie sie durch Mikroarrays ermöglicht wird, für die Diagnostik notwendig. Die entsprechenden Arbeiten werden im Kapitel 4.2.1. beschrieben und Ergebnisse in Kapitel 4.2.2. vorgestellt. Neben der Anwendung im Umfeld nicht-kodierender RNAs hat die Geniom Technologie eine weitere Anwendung gefunden. Sie wird eingesetzt, um fehlerfreie Oligonukleotide schnell zu synthetisieren, vom Glasträger abzulösen und in der Synthetischen Biologie anzuwenden (Kapitel 4.4). Diese Anwendungsvariante hat der Genomik-Pionier Graig Venter exklusiv für seine Firma SGI erworben. Zusätzlich wurde die entsprechende Technologie verwendet, um Anreicherung für Gen Panels und anschließende Sequenzierung zu ermöglichen [48]. Diese Methode, die heute in der genetischen Rutine-Diagnostik eingesetzt wird und auf die ich in der Auswertung nicht weiter eingehe, nennt sich targeted Next-Generation Sequencing. Es ist ebenfalls erwähnenswert, dass die Technologie nicht nur von mir und Kooperationen eingesetzt wurde, sondern auch von anderen Forschern weltweit verwendet wird. [146-151].

Publikationen: Die Arbeiten, die in diesem Abschnitt beschrieben werden basieren hauptsächlich auf den folgenden Publikationen [39, 41]. Außerdem wurde die Technologie patentiert (DE-19940750.9-52).

4.1.2. Microfuidic Primer Extension für miRNAs

Im vorangegangenen Abschnitt haben ich das Konzept eines sehr flexiblen Mikroarray Systems beschrieben, dass über Nacht eigene Mikroarrays mit neuem Inhalt kostengünstig herstellen kann. Insbesondere für die Quantifizierung von RNAs und hier wiederum von kleinen RNA Stücken ist die Technologie besonders geeignet.

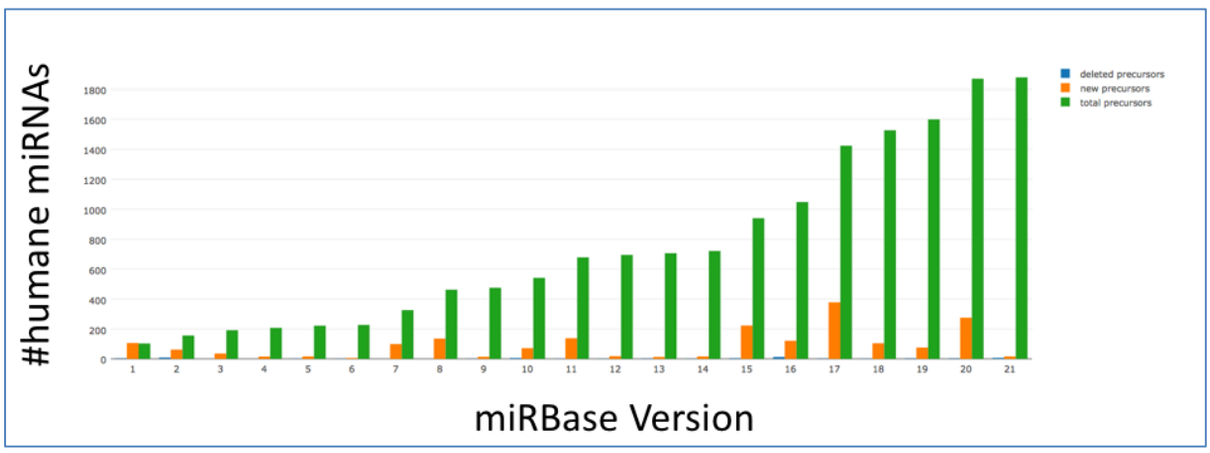


Abbildung 12: Die Entwicklung der miRBase.

Gezeigt sind die 21 miRBase Versionen die bisher veröffentlicht wurden und die Anzahl von humanen miRNA Precursor in der miRBase. Die Entwicklungs-Zyklen am Anfang der Datenbank waren oft deutlich weniger als ein Jahr. Die Abbildung oben ist mit Hilfe der miRCarta Datenbank erstellt worden.

Für miRNAs ist das schnelle Updaten der miRBase [79, 80] (siehe auch Kapitel 2), vor allem in den Anfängen ein Problem: Während große Hersteller von Mikroarrays mehrere Wochen bis Monate brauchen um einen entsprechend neuen Mirkroarray anzubieten, kann das Geniom diesen über Nacht herstellen. Die Update Zyklen der 21 Versionen waren oft aber deutlich geringer als ein Jahr (Abbildung 12).

Eine Herausforderung bei der Quantifizierung von miRNAs ist die Sequenz Homogenität – besonders am 3' Ende von miRNAs innerhalb von miRNA Familien. Eines der grundlegensten Beispiele ist die let-7 Familie. Einige Mitglieder der Familie sind in Abbildung 12 als Multiples Sequenz Alignement gezeigt. Nur an drei Basen können Unterschiede festgestellt werden. Ein häufig auftretendes Problem ist daher die Kreuzhybridisierung: let-7a miRNAs hybridisieren in konventionellen Assays oft mit let-7b Fänger Sonden.

let7-a-5p	UGAGGUAGUAGGUUGUAU**A**GUU
let7-b-5p	UGAGGUAGUAGGUUGU**G**UGGUU
let7-c-5p	UGAGGUAGUAGGUUGUAUGGUU
let7-f-5p	UGAGGUAGUAG**A**UUGUAU**A**GUU

Abbildung 13: Ausgewählte Beispiele der let-7 Familie beim Menschen.

Die Basen, die sich unterscheiden sind in fett hervorgehoben. Die Grafik ist in Anlehnung an Abbildung 2A aus Kappel et al. entstanden.

Eine weitere Herausforderung ist neben der Spezifität für die Familienmitglieder auch eine hinreichende analytische Sensitivität: Geringe RNA Mengen (wenige Nanogramm)

sollen mit möglichst wenig Präprozessierung und insbesondere ohne Amplifikation akkurat gemessen werden.

Das Prinzip des MPEA Assays ist die Fänger-Sonde die auf dem Mikroarray synthetisiert wird um einige Basen zu verlängern. So entsteht ein Überhang, die zu messende miRNA ist kürzer als die synthetisierte Sonde auf dem Mikroarray. DNA Polymerase kann dann verwendet werden, um mit Biotin markierte Basen einzubauen, die die Quantifizierung ohne vorherige Markierung der miRNAs erlauben. Verschiedene Parameter müssen getestet werden, um die optimalen Assay Bedingungen zu definieren. Das beinhaltet die Syntheserichtung (3'->5' oder 5'->3'), die Base die für die Verlängerung verwendet wird (bio-dATP, bio-dCTP, bio-dGTP oder bio-dUT) und die Anzahl der Nukleotide, um die die Fänger-Sonde verlängert wird.

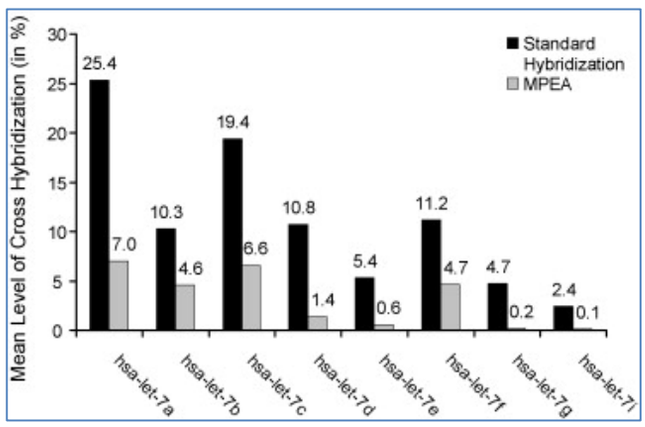


Abbildung 14: Kreuzhybridisierung im MPEA Assay.

MPEA im Vergleich zu Standard Mikroarray Assays für die let-7 Familie. Für den MPEA Assay wird eine signifikant niedrigere Kreuzhybridisierung und damit signifikant höhere Spezifität, vor allem am 3' Ende, erzielt. Die Abbildung entstammt aus Vorwerk et al.

Diese Parameter sind systematisch ausgetestet worden. Die besten Resultate werden mit bio-dATP erzielt. Dabei ist es entscheidend, dass die Synthese in Richtung von 3' nach 5' durchgeführt wird, das 3' Ende muss auf dem Array immobilisiert werden. Diese Syntheserichtung ist wichtig um die Spezifität am 3' Ende der miRNA zu erhöhen. Als letzter Parameter ist die Anzahl der Nukleotide getestet worden, die ein optimales Signal ergaben. Generell gilt, dass je mehr markierte Nukleotide eingebaut werden, um so sensitiver die Messmethode wird. Ab einer bestimmten Anzahl tritt aber eine steirische Hinderung ein. Getestet wurden alle Möglichkeiten von einer bis zu zwölf Nukleotiden. Die besten Resultate werden bei 5 Nukelotiden erzielt. Danach tritt der beschriebene Effekt der steirischen Hinderung auf. Der hier beschriebene Assay liefert reproduzierbare und spezifische Ergebnisse bis hin zu 50 Nanogramm totale RNA als Eingangsmaterial. Sogar bis zu 20 Nanogramm konnten noch verwertbare Ergebnisse erzielt werden.

Der MPEA Assay bietet in Kombination mit der Geniom Technologie eine sehr gute Möglichkeit, flexibel miRNAs, immer aus der jeweils neuesten Version der miRBase, zu messen. Besonders positiv neben der schnellen Durchführung von Experimenten sind die hohe Spezifität und die analytische Sensitivität. Daher bildet der MPEA Assay die Basis für den Großteil der Arbeiten, die im Kapitel 4.2. beschrieben sind. In den vergangenen drei

Jahren hat sich die miRBase kaum weiterentwickelt, das letzte Update datiert aus dem Juni 2014. Dadurch ist ein Vorteil der Geniom Technologie für miRNAs entfallen. Zusätzlich ist eine eigene Plattform für die Experimente die von vielen Technologien – anderen Mikroarrays, HTS, oder RT-qPCR – durchgeführt werden können nicht wirtschaftlich und die Geniom Technologie, die kommerziell nicht mehr verfügbar ist, wird nicht weiterverwendet.

Daher werden momentan hauptsächlich die Agilent Micro-Array Technologie und cPAS basierte Sequenzierung (Kapitel 4.1.4) eingesetzt um miRNA Profile zu generieren. Neben solchen dezentralen Plattformen gibt es momentan den generellen Trend in der Medizin, Messungen mindestens im Zentrallabor von Krankenhäusern, besser sogar direkt patientennah „Point-of-Care“ durchzuführen. Der hier von mir beschriebene MPEA Assay dient als Grundlage für den miRNA Immunoassay, der im folgenden Kapitel beschrieben ist.

Publikationen: Die Arbeiten, die in diesem Abschnitt beschrieben werden basieren hauptsächlich auf [42], beinhalten aber bereits Aspekte aus [43]. Außerdem ist die Idee im Kontext in zwei sehr umfangreichen Patenten im Detail beschrieben (EP2109499, DE102007018833).

4.1.3. Point-of-Care miRNA Testung

Wie im vorangegangenen Abschnitt beschrieben gibt es einen stetigen Trend hin zur Point-of-Care Testung. Das gilt im besonderen Maße für Infektionserkrankungen [152], aber auch für andere zeitkritische Tests wie die Messung von Troponin im Umfeld kardiologischer Diagnostik [153]. Eine hervorragende Übersicht bietet der Artikel von John und Price [154].

Ein erheblicher Anteil der In-Vitro Standard-Diagnostik in klinischen Laboren sind Immunoassays, sogenannte ELISA (Enzyme-linked Immunosorbent Assay) [155]. Diese kostengünstige (ein ELISA kostet oft weniger als ein Euro in der Herstellung) und schnelle Technologie (Zeit vom Probeneingang bis zum Testergebnis sind in der Regel nur 1-2 Stunden) wird zur Messung von Proteinen eingesetzt und ist aus der Routine Diagnostik nicht wegzudenken. Für die Messung von Nukleinsäuren, also DNA oder RNA, wurden ELISA bisher hingegen nur wenig verwendet.

Ich habe daher das Konzept eines Immunoassays ähnlich klassischer ELISA Tests aber zur spezifischen Messung von miRNAs entwickelt. Dabei ist die grundlegende Idee des Assays

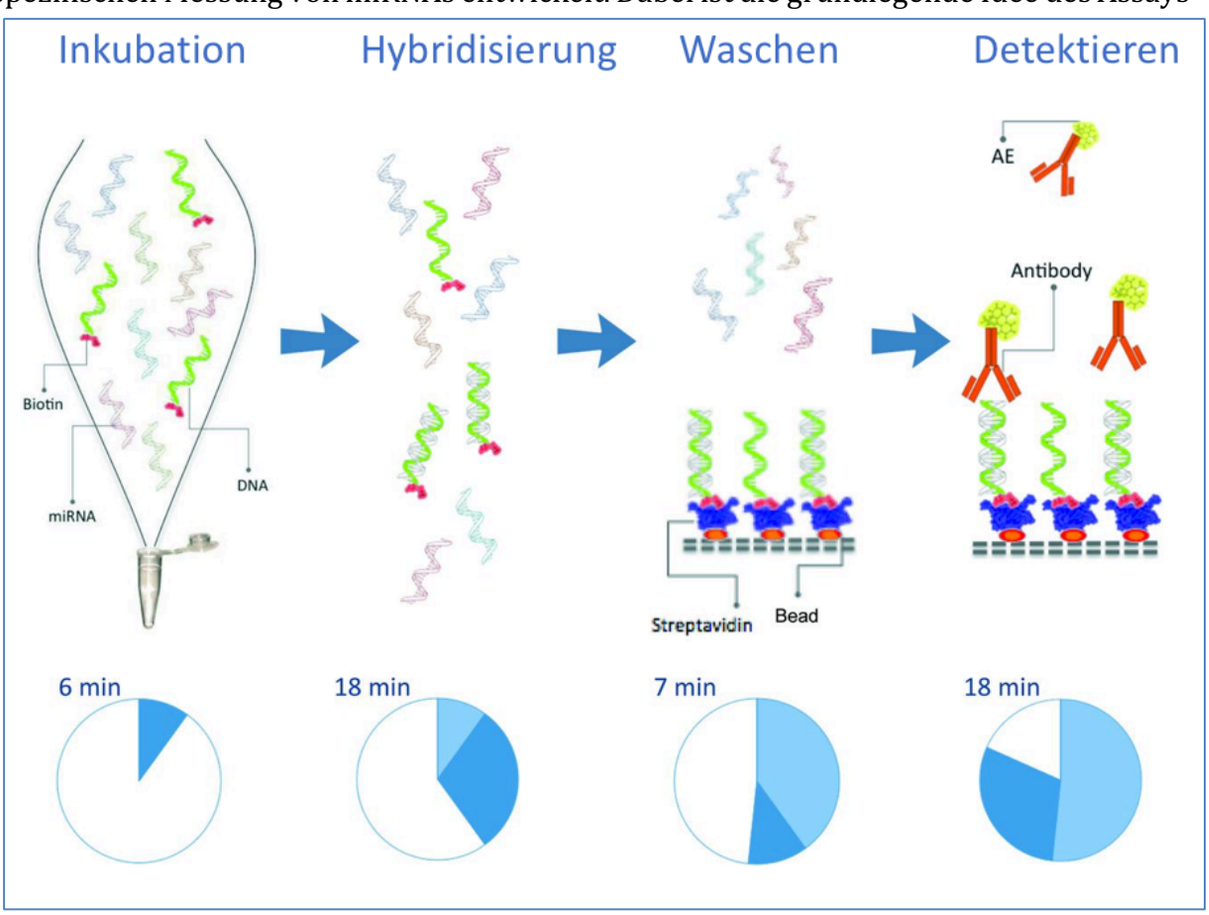


Abbildung 15: Prinzip des miRNA Immunoassays.

Die Gesamtzeit für alle Schritte, die vollautomatisch durchgeführt werden, beträgt 49 Minuten. Die Abbildung ist modifiziert aus Kappel et al. entnommen.

(Spezifität für miRNA Familien Mitglieder zu erreichen) und die Methodik der Auswertung ähnlich zum vorher beschriebenen MPEA Assay. Der Ansatz ist ein zweistufiger Test, der für das Siemens Centaur System entwickelt wurde, aber mit jedem anderen Analyzer im Zentrallabor oder Point-of-Care Gerät kompatibel ist. Die wichtigsten Komponenten sind Streptavidin markierte Mikropartikel, eine mit Biotin markierte Fänger-Sonde, die komplementär der nachzuweisenden miRNA ist sowie ein monoklonaler Antikörper, der spezifisch zur Detektion von DNA / RNA hybriden ist und der mit Acridinium Ester markiert ist.

Im ersten Schritt des Assays werden die miRNAs einer biologischen Probe, in diesem Fall einer Blutprobe, mit der Fänger-Sonde, die mit Biotin markiert ist, hybridisiert. Es bilden sich dabei perfekte Heterohybride aus der miRNA und der Fänger-Sonde, die eine einzelsträngige DNA ist. Im zweiten Schritt werden die Hybride mit der immobilisierten Streptavidin Phase gebunden. Final wird der Antikörper zur Detektion der DNA/RNA Hybride zugegeben. Der Antikörper ist sehr spezifisch, er erkennt nur perfekte Paare aus DNA und RNA, ein Missmatch wird nicht erlaubt. Daher ist die Menge an gebundenem Antikörper proportional zu der Menge an DNA/RNA Hybriden die wiederum proportional zu der Menge der miRNA, die detektiert werden soll, in der Blutprobe ist.

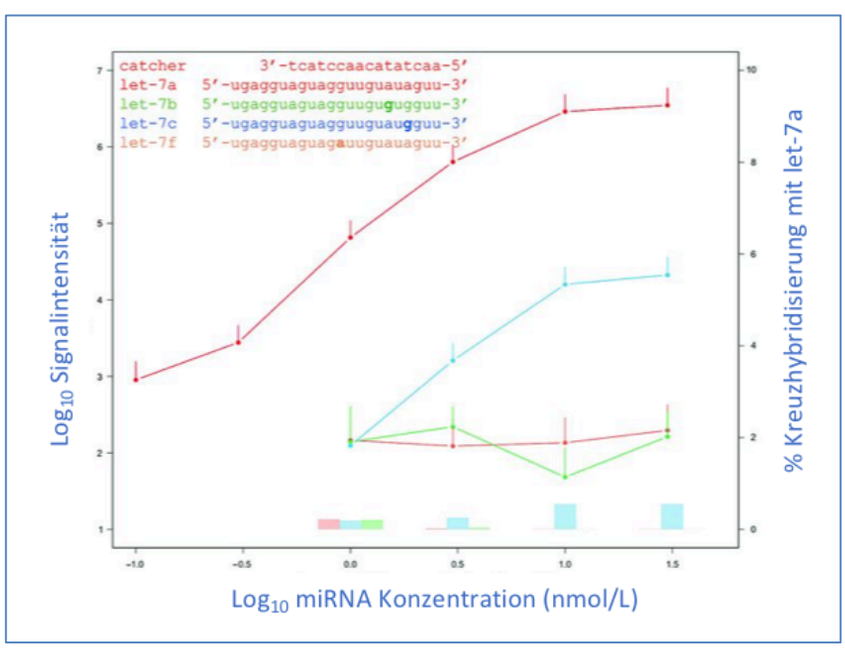


Abbildung 16: Analytische Sensitivität und Spezifität des Immunoassays.

Die Grafik zeigt für verschiedene Konzentrationen von let-7a, let-7b, let-7c und let-7f und eine Fänger-Sonde, die spezifisch für let-7a ist, die Signalintensität (linke Skala und Kurven) und die prozentuale Kreuzhybridisierung (rechte Skala und Balkenhöhe). Die Abbildung ist modifiziert aus Kappel et al. entnommen.

Der voll automatische Assay, der in Abbildung 15 übersichtlich und schematisch gezeigt ist, besteht im Detail aus den folgenden Schritten: Pipettiere 75 µL Probenmaterial in eine Küvette. Pipettiere 75 µL 20 mmol/L Sodium Phosphat, pH 7.2, 300 mmol/L NaCl, 0.1% Triton X-100, 0.5% Bovine Serum Albumin, 0.02% Sodium Azide und Biotin markiertes Oligonucleotid (10 nmol/L) und inkubiere es mit der Probe für 6 Minuten bei 37 °C. Pipettiere 150 µL der Solid Phase dazu und inkubiere dies für 18

Minuten bei 37 °C. Trenne die Solid Phase von der Lösung. Wasche die Küvette 6.75 min at 37 °C. Pipettiere 95 µL Antikörper und inkubiere dies für 18 Minuten bei 37 °C. Wasche die Küvette nach Separation der Solid Phase. Pipettiere 300 µL Säure und 300 µL Base, um die Chemolumineszenz zu erzeugen.

Der Assay wurde mit einem ähnlichen 2xperimentellen Set-Up, wie im vorherigen Abschnitt für den MPEA Assay beschrieben, getestet. Verschiedene Konzentrationen (0.1-30nmol/L) von Mitgliedern der let-7 Familie wurden gemischt. Im konkreten Beispiel, das in Abbildung 16 gezeigt ist, wurde eine Fänger-Sonde, die spezifisch für let-7a ist, zugegeben. Die maximale Kreuzhybridisierung, die beobachtet wurde liegt bei weniger als 0.6%, die technische Spezifität dementsprechend bei 99.4%. Weitere Konzentrationsreihen haben gezeigt, dass selbst Konzentrationen von 1 Pikomol je Liter stabil gemessen werden können. Die gemessene Konzentration hat mit der tatsächlichen Konzentration dabei sehr exakt übereingestimmt (Pearson Korrelation von 0.998). Nachdem die technische Spezifität und Sensitivität bestimmt worden ist, wurde der Assay auf biologische Proben angewendet. Als Beispiel dient das später in Kapitel 4.2.3 beschriebene Set an Alzheimer miRNAs. Für alle getesteten Marker (hsa-miR-5010–3p, hsa-miR-26a-5p, hsa-miR-151a-3p und hsa-let-7d-3p) wurden in 40 biologischen Replikaten stabile Signale nachgewiesen. Bemerkenswert war der Variationskoeffizient von miR-26a-5p, der nur 4% betragen hat. Auch Unterschiede zwischen Patienten und Kontrollen, wie sie sonst typisch für miRNAs sind, konnten detektiert werden. Final wurde der Immunoassay gegen RT-qPCR, als Gold Standard, getestet. In diesen Experimenten war die Korrelation zwischen den beiden Technologien enorm hoch (Pearson Korrelation 0.994) und zeigen dass der Immunoassay kompetitiv zur klassischen RT-qPCR basierten Detektion on miRNAs ist.

Mit dem Immunoassay habe ich einen entscheidenden Beitrag geleistet, dass miRNAs in Richtung klinische Testung weiterentwickelt werden können. Die Limitation liegt momentan in der Fähigkeit, mehrere miRNAs parallel aus der selben Probe zu messen. Die einzige Möglichkeit die nicht nur konzeptionell vielversprechend war, sondern auch verwirklicht werden konnte, ist ein serielles Multiplexing. Das bedeutet, dass der in Abbildung 15 gezeigte Ablauf für jede miRNA hintereinander und nicht parallel durchgeführt wird. Begonnen wird dabei mit der am niedrigst-konzentrierten miRNA in der Probe. Das zusammen mit der verfügbaren Menge an Ausgangsmaterial macht es bisher möglich, etwa 4-8 miRNAs von einem Patienten und aus einer Blutprobe zu messen. Für die meisten Anwendungen (siehe Kapitel 4.2.) ist diese Anzahl an Markern ausreichend.

Erwähnenswert ist, dass der Assay und die Lizenzen an dem Assay 2015 von der Firma Biovendor gekauft worden sind und seit Ende 2017 kommerziell als miREIA Assay angeboten werden (https://www.biovendor.com/mireia-breakthrough-assays). Die Weiterentwicklung der Firma Biovendor erlaubt dabei sogar Messungen bis zu Konzentrationen von 0.1 attomol/µl miRNA (0.1 Trillionstel Mol je Liter).

Publikationen: Die Arbeiten, die in diesem Abschnitt beschrieben werden basieren hauptsächlich auf [43]. Außerdem wurde ein ähnliches Konzept mit anderem Detektions-Mechanismus entwickelt und publiziert [44]. Details zu diesem parallelen Ansatz habe ich in der vorliegenden Ausarbeitung nicht beschrieben. Der Immunoassay wurde außerdem patentiert (EP20120159196).

4.1.4. cPAS Sequenzierung

Ich habe mich in meiner Forschung damit befasst, miRNAs in die klinische Routine, oder mindestens näher an die klinische Routine zu bringen. Dazu habe ich hauptsächlich die Kern-miRNAs aus den frühen Versionen der miRBase betrachtet. Neben diesen und weiteren miRNAs aus der miRBase sind jedoch Teile des humanen miRNomes unbekannt. Die bisher beschriebenen Technologien, Mikroarrays, RT-qPCR und die klinischen Assays die ich entwickelt habe, sind zur Detektion neuer miRNAs, die bisher nicht beschrieben sind, ungeeignet. Im Gegensatz dazu bietet Hochdurchsatz Sequenzierung (HTS) die Möglichkeit bisher noch nicht identifizierte miRNAs in speziellen Zelltypen, Geweben oder Körperflüssigkeiten zu finden. Die wohl am meisten eingesetzte Methode ist Sequenzierung durch Synthese, wie sie von der Firma Illumina eingesetzt wird.

Charakteristisch für diese Technologie ist normalerweise eine Amplifizierung des Ausgangsmaterials. Das führt zu mehreren möglichen Fehlerquellen. „Bias“ in HTS Datensätzen ist daher bekannt [156, 157] und weit verbreitet, insbesondere wenn es um die Quantifizierung von RNAs geht [158]. In einer Studie mit der chinesischen Firma BGI konnte jedoch gezeigt werden, dass eine Sequenzierung die nicht auf einer klassischen PCR, sondern auf einer linearen Amplifikation beruht, deutlich bessere Resultate liefert [45]. Das Prinzip der combinatorial probe-anchor synthesis (cPAS), wie sie auf dem BGISEQ-500 Sequenzierer etabliert wurde, funktioniert mit DNA Nanoball (DNB) Nanoarrays mit Hilfe einer schrittweisen Sequenzierung durch eine Polymerase. Diese Methode zeigt vor allen Dingen bei kurzen Reads eine sehr hohe Genauigkeit. Außerdem erlaubt es die Technologie, Milliarden von Molekülen parallel zu messen. Die Methode, die

von Complete Genomics / BGI hauptsächlich entwickelt wurde, um DNA zu sequenzieren, eignet sich daher besonders für miRNA Anwendungen.

In einer Proof-of-Concept Studie wurden sechs Gehirnproben, zwei Herzproben und zwei Blutproben sequenziert und insgesamt 300 Millionen Reads generiert. Technische Replikate der sechs Gehirnproben haben eine mittlere Korrelation von 0.98 ergeben. Mit anderen Technologien wie zum Beispiel der Sequenzierung mittels Illumina hat sich immer noch eine Korrelation von 0.75 ergeben. Da für meine Forschung die Anwendung als blutbasierte Biomarker besonders wichtig ist, möchte ich diesen Aspekt näher beleuchten und identische Blutproben gemessen auf Mikroarrays, mit Illumina Sequenzierung und cPAS Sequenzierung vergleichen. Abbildung 17 zeigt den relativen Anteil der 10 am häufigsten gefundenen miRNAs in Blutzellen, abhängig von den drei Technologien. Für Illumina Sequenzierung entsprechen 90.8% aller Reads einer einzigen miRNA, miR-486-5p, die als miRNA in roten Blutzellen bekannt ist [159]. Die anderen verwendeten Technologien haben ebenfalls eine Überrepresentation dieser miRNA gezeigt (7.7% aller Reads bei cPAS Sequenzierung und 17% der totalen Intensität auf Mikroarrays), jedoch war das Verhältnis in keinem Fall so extrem wie bei Illumina Sequenzierung. Validierung mittels RT-qPCR hat den Bias für diese miRNA in der Illumina

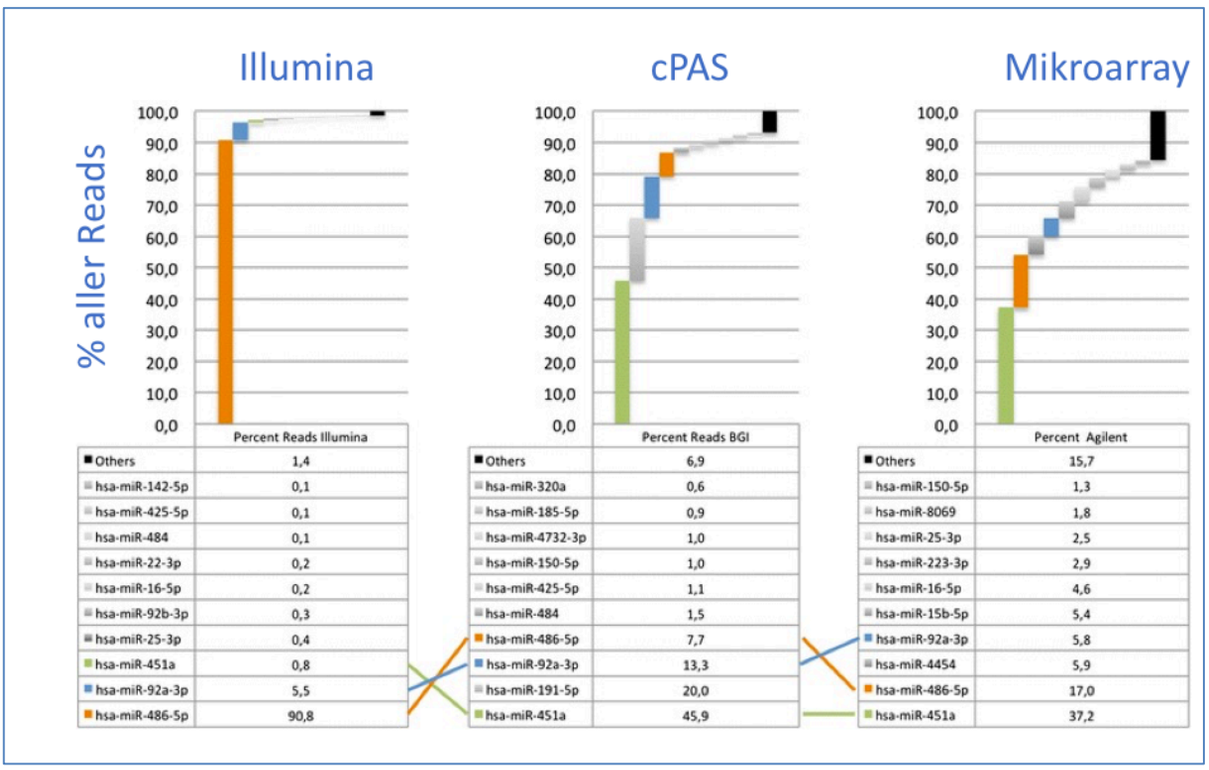


Abbildung 17: Verteilung der Signal Intensität von miRNAs.

Die Abbildung zeigt für die 10 häufigsten miRNAs wie viel % der totalen Signal Intensität je Technologie gemessen werden. Die Abbildung ist modifiziert aus Fehlmann et al. entnommen.

Technologie verifiziert. Im Umkehrschluss haben die Top-10 der miRNAs in der Illumina Sequenzierung 98.6% aller Reads ausgemacht. Für alle anderen, mehr als 2,000 miRNAs sowie potenziell neue Kandidaten bleiben zusammen gerade 1.4% der gesamten Sequenzier-Kapazität.

Der Bias in der gängigen Illumina Technologie zusammen mit den komplexen Anforderungen an Labore und den Zeitaufwand der nötig ist (immer noch mehrere Tage) und verhältnismäßig hohe Kosten von mehreren hundert Euro, machen einen Einsatz der entsprechenden Technik in der Standard-Diagnostik unwahrscheinlich. Im Gegensatz dazu bietet die cPAS basiere Sequenzierung einige Vorteile. Zumindest in Service Laboren ist ein Einsatz zur Diagnostik von Erkrankungen aus dem Blut möglich.

Publikationen: Die Arbeiten, die in diesem Abschnitt beschrieben werden basieren hauptsächlich auf [45].

4.1.5. Zusammenfassung der Technologien

Zusammenfassend habe ich in diesem Kapitel einen Überblick über Technologien und Assays gegeben, an deren Entwicklung ich beteiligt war und die zur Messung von miRNAs in der Diagnostik eingesetzt werden können. Die verschiedenen Technologien haben dabei verschiedene Vor- und Nachteile. Für die klinische Testung von kleinen miRNA Sets in der Routine-Diagnostik ist die Immunoassay Methode wahrscheinlich am besten geeignet, während für die Grundlagenforschung die Sequenzierung die meisten neuen Erkenntnisse verspricht. Mikroarrays liegen im Anwendungsspektrum zwischen diesen beiden Extremen.

Insgesamt hat die Forschung aber gezeigt, dass es vielversprechender ist sich auf Inhalte wie Biomarker in Krankheiten zu konzentrieren, statt auf die Entwicklung von Plattformen. Um miRNAs zu messen, können Forscher bereits heute aus mehreren Dutzend Technologien wählen [160], die sie dazu einsetzen können miRNA Biomarker zu detektieren und zu validieren. Mit diesem Thema beschäftige ich mich daher in den folgenden Abschnitten.

4.2. miRNAs als Biomarker

Im vorangehenden Abschnitt habe ich technologische Entwicklungen zur Messung von miRNAs beschrieben. Im Laufe meiner Forschung hat sich der Fokus allerdings schrittweise von der Technologieentwicklung, die ich als Ingenieur begonnen habe, über Assay Entwicklung bis hin zur Anwendung der Detektion von Biomarkern, verschoben. Grundlegend basiert der Forschungsansatz dabei auf drei wesentlichen Paradigmen:

1. Es sollen leicht zugängliche Biomarker gemessen werden. Hier bieten sich Körperflüssigkeiten wie zum Beispiel Blut an. Dieser Ansatz erlaubt eine breite Anwendung unabhängig des Organes oder der Erkrankung und ermöglicht gleichzeitig einfaches longitudinales Messen.

2. Da einzelne miRNAs nicht genügend Aussagekraft haben, sollen Sets von miRNAs gemessen werden. Je nach Komplexität der klinischen Fragestellung sind typischerweise 4-12 miRNAs notwendig, um hinreichende Genauigkeit zu erlangen.

3. Die Muster sollen nicht in einer einzelnen Erkrankung betrachtet werden, sondern über verschiedene Erkrankungen hinweg. Das ist notwendig, um die Spezifität einer miRNA Signatur für eine Erkrankung abschätzen zu können.

Im Folgenden werde ich in vier Unterabschnitten auf die Entwicklung von miRNA Biomarkern eingehen. Zunächst ist es essenziell, die technische und biologische Stabilität zu verstehen. Welche miRNAs können unabhängig äußerer Einflüsse reproduzierbar gemessen werden und sind am besten unabhängig von „Confoundern“, wie zum Beispiel dem Alter und dem Geschlecht (Abschnitt 4.2.1)? Anschließend beschreibe ich die Anwendung im Umfeld der Diagnose von Lungenerkrankungen (Lungentumore / Chronisch Obstruktive Pulmonary Disease COPD; Abschnitt 4.2.2) und Erkrankungen des Zentralen Nervensystems (Multiple Sklerose und Alzheimer; Kapitel 4.2.3). Im vierten Teil gehe ich dann auf den Aspekt des krankheitsübergreifenden miRNOmes ein, also welche miRNAs beispielsweise bei allen untersuchten Erkrankungen dysreguliert sind.

4.2.1. Technische und biologische Stabilität von miRNAs

Bevor miRNAs in der klinischen Diagnostik eingesetzt werden um Krankheiten zu erkennen, ist es notwendig, Detailwissen über die biologische und technische Stabilität zu erlangen. Im Bereich der technischen Stabilität ist es besonders wichtig, den Einfluss von Lagerungs- und Transportbedingungen zu kennen [50]. Außerdem haben Studien gezeigt, dass wiederholtes Einfrieren und Auftauen die Probenqualität beeinflusst. Die entsprechende Analyse für miRNAs aus Vollblutproben wurde in Analytical Chemistry veröffentlicht [50]. Das entsprechende Studien Set-Up ist in Abbildung 18 gezeigt.

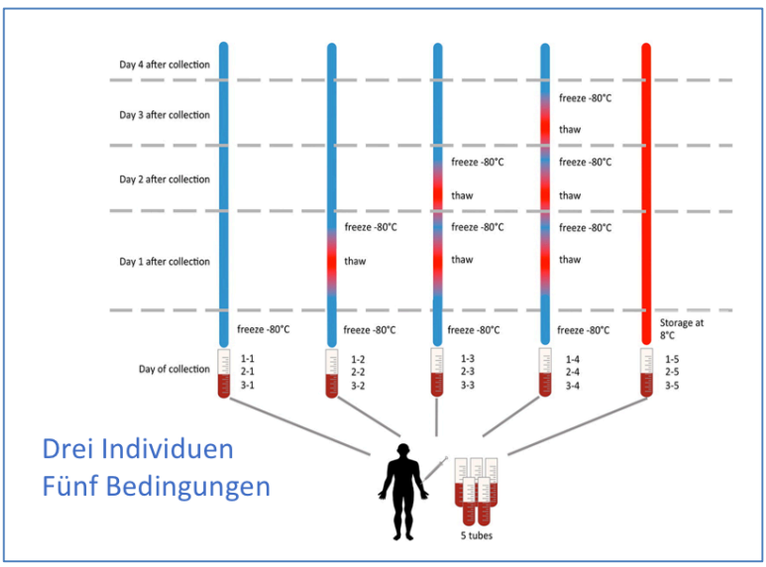


Abbildung 18: Stabilitätsanalyse von miRNAs.

Für drei Spender wurden fünf verschiedene Bedingungen getestet. Die Abbildung ist modifiziert aus Backes et al. entnommen.

Für drei Spender wurden 5 verschiedene Bedingungen getestet, die Lagerung bei Raumtemperatur, bei -80 Grad und bis zu dreimaliges Einfrieren und Auftauen. Diese Herangehensweise hat es erlaubt, abzuschätzen wie stark die technische Variabilität im Verhältnis zur intra-individuellen Variabilität schwank. Um zu verstehen, wie sich die Muster insgesamt zwischen den verschiedenen experimentellen Bedingungen verhalten haben, wurden multivariate statistische Methoden verwendet, das sind Methoden, die auf mehreren sogenannten „Features“ basieren. Im vorliegenden Fall ist jedes Feature die Expression einer miRNA. In der Studie wurden Signale von 455 miRNAs zugleich verwendet. Die Methoden, die die am besten interpretierbaren Ergebnisse gezeigt haben, waren bottom-up hierarchisches Clustern mit der Euklidischen Distanz als Abstandsmaß und die vorwiegend als Dimensions-Reduktion genutzte Principal Component Analyse (Hauptkomponenten Analyse). Generell haben die Proben, die gleichbehandelt wurden, auch ähnliche Muster gezeigt. Insgesamt waren die Effekte aber vergleichsweise gering. Dennoch zeigen die Resultate auch, dass es wichtig ist, Proben innerhalb einer Studie absolut gleich zu behandeln.

Neben den Effekten die auf eine Probe insgesamt einwirken, ist es fast noch wichtiger zu verstehen, auf welche Marker der maximale Einfluss besteht. Solche Marker können bei

der Entwicklung und der Translation von Biomarkern zur Anwendung hin zum Beispiel ausgeschlossen werden. Dazu wurde jede miRNA alleine in Varianzanalysen (ANOVA) und bezüglich des Variationskoeffizienten hin untersucht. Die Varianzanalyse hat gezeigt, dass fünf miRNAs nach Adjustierung für Multiples-Testen signifikant waren. Diese sind hsa-miR-320b (p = 0.0002), hsa-miR-320a (p = 0.001), hsa-miR-16-5p (0.018), hsa-miR-18b-5p (0.037) und hsa-miR-375 (0.0375). Die entsprechenden Biomarker sollten bei der klinischen Testung genauer beobachtet werden, da signifikante Schwankungen leicht auf technische Artefakte hindeuten können. Interessanterweise waren die Schwankungen zwischen Individuen normalerweise größer als die technischen Schwankungen. Beispiele – jeweils für technische Schwankungen und Schwankungen zwischen Individuen – sind in Abbildung 19 gezeigt.

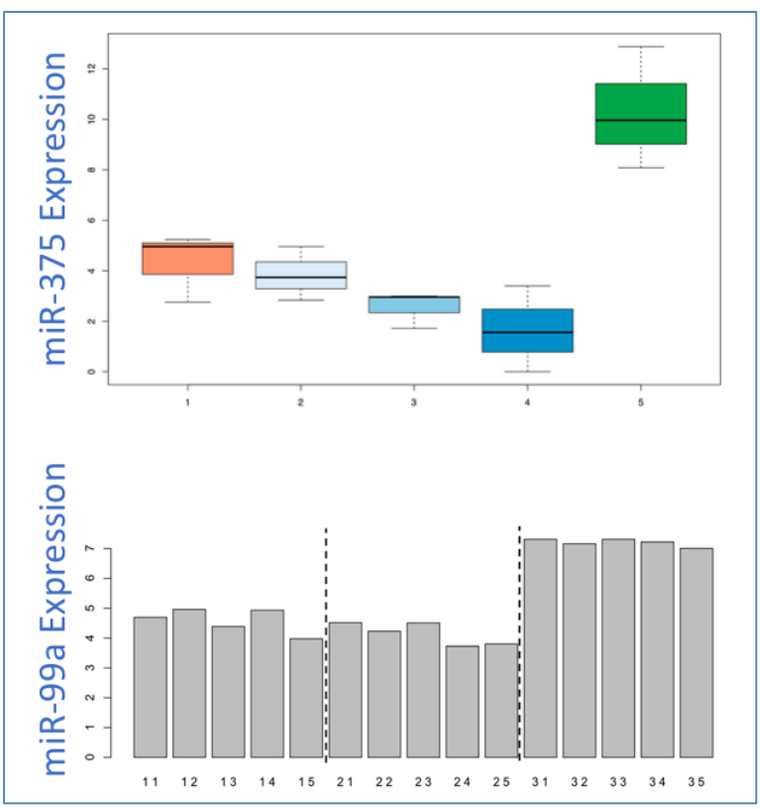


Abbildung 19: Technische und interindividuelle Variabilität von miRNAs.

Der obere Teil der Abbildung zeigt die Variabilität zwischen den verschiedenen Experimenten für eine der variabelsten miRNAs. Hier zeigten die Individuen jeweils ähnliche Expressionswerte. Im Vergleich dazu ist unten die Schwankung einer miRNA, je nach Individuum gezeigt. Hier ist die Expression zwischen den Experimenten etwa gleich, aber der dritte Proband hatte signifikant höhere Level der miRNA. Die Abbildung ist modifiziert aus Backes et al. entnommen.

Die Frage, die sich als nächste stellt, ist die Ursache nach der Schwankung in der Expression zwischen verschiedenen Personen. Hierfür kann es verschiedene Gründe geben, entweder schwanken die miRNAs tatsächlich so stark zwischen beliebigen Individuen oder es gibt generelle Einfluss-faktoren die den Level einzelner miRNAs beeinflussen. Die wohl klassischsten Beispiele dafür sind das Alter und das Geschlecht. Um ein Verständnis dafür zu erlangen, ist es notwendig, gesunde Probanden verschiedener Altersgruppen und sowohl Männer als auch Frauen zu vergleichen [51]. Die Analyse von 167 gesunden Probanden mittels Mikroarray und HTS hat einen deutlicheren Einfluss des Alters auf die miRNA Muster gezeigt als es das Geschlecht gezeigt hat. Bezüglich des Geschlechts waren 144 miRNAs

signifikant unterschiedlich reguliert, nach der notwendigen Adjustierung für Multiples-Testen war jedoch keine miRNA mehr signifikant. Mit dem Alter der Probanden waren insgesamt 318 miRNAs signifikant verknüpft. Nach der Adjustierung für Multiples-Testen waren davon immerhin noch 35 Marker signifikant mit dem Alter korreliert. Entsprechende miRNAs müssen nicht zwangsläufig als Biomarker ausgeschlossen werden. Es kann jedoch – je nach dem Umfang der Schwankung – sinnvoll sein, für verschiedene Altersgruppen verschiedene Grenzwerte in einem klinischen Test einzuführen. Konkret könnte das heißen, dass im Falle von miR-34a ein Mann im Alter von 35 Jahren bei 12 Nanogramm je Milliliter eine andere Diagnose erhält als ein Mann von 65 Jahren mit der gleichen absoluten Menge dieser miRNA oder eine zwanzigjährige Frau, die ebenfalls die gleiche Expressionsstärke der miRNA zeigt.

MirNaCon, ein internetbasiertes Software-Tool (frei verfügbar: http://www.ccb.uni-saarland.de/mirnacon), erlaubt es anderen Forschern diese Betrachtung in ihrer Forschung und der Translation der miRNA Biomarker zu berücksichtigen. Sie können eine Liste von miRNAs eingeben und bekommen innerhalb weniger Sekunden angezeigt, welche der miRNAs weder vom Alter noch vom Geschlecht abhängig sind. Diese haben eine höhere Chance in der Translation zur klinischen Testung.

Da die hier beschriebenen Muster auf Blut basieren stellt sich außerdem die Frage, ob die entsprechenden Marker auch im Gewebe gefunden werden. Gerade bei miRNAs ist eine hohe Spezifität für Gewebe bekannt. Da miRNAs aber wie oben aufgeführt auch von Person zu Person schwanken, ist es notwendig, Organmuster von verschiedenen Organen der selben Person zu messen, damit solche Schwankungen zwischen Personen nicht zu artifiziellen organspezifischen Befunden führen [52]. Wie im frei verfügbaren internet-basierten Tool TissueAtlas gezeigt (https://ccb-web.cs.uni-saarland.de/ tissueatlas/), gilt die Beobachtung, dass miRNAs sehr gewebsspezifisch exprimiert werden, nur bedingt. In der Tat ist es so, dass 82.9% aller getesteten miRNAs einen mittleren Spezifität-Index hatten, also weder in allen Organen vorkamen noch spezifisch für einzelne Organe waren. Dennoch waren insgesamt 143 miRNAs in allen getesteten Organen vorhanden. Das Blut zeigte in erstaunlich vielen Fällen Expression für eher spezifische miRNAs und die im Rahmen dieser Arbeit gemessenen PAXgene Muster, die auf Blutzellen basieren, scheinen generell viele organtypische miRNAs zu enthalten. Der wohl interessanteste Aspekt dieser Arbeit war, dass in einem Vergleich zwischen Spezies nicht etwa Mensch- und Rattenmuster zusammen passten, sondern in fast allen Fällen die Gewebe des Menschen mit den entsprechenden Geweben der Ratte am besten übereinstimmten.

Zusammenfassend haben die Ergebnisse in der grundlegenden Forschung über miRNAs ergeben, dass die kleinen nicht kodierenden RNAs eher stabil sind und nur teilweise von

äußeren technischen und biologischen Faktoren beeinflusst werden. Diese zu kennen ist jedoch für die Entwicklung von Biomarkern wie sie in Kapitel 4.2.2 und 4.2.3 beschrieben sind, unerlässlich.

Publikationen: Die Arbeiten, die in diesem Abschnitt beschrieben werden basieren hauptsächlich auf* [50-52]. *Zusätzlich wurde ein Patent angemeldet, um miRNAs aus dem Blut besser quantifizieren zu können (US2015184223).

4.2.2. Anwendungen im Bereich Lungentumore

Parallel zu den Untersuchungen der Stabilität, die in Kapitel 4.2.1 beschrieben sind, habe ich mit Prof. Keller und Prof. Meese an Lungentumormarkern, basierend auf miRNAs als Anwendungsbeispiel für die Technologien und Assays, die in Kapitel 4.1. beschrieben sind, gearbeitet. Lungentumore bzw. Lungenerkrankungen sind eines der Hauptforschungsfelder in den AGs „Klinische Bioinformatik“ und „Humangenetik“ in Saarbrücken und Homburg [53-55, 104, 161-166]. Aus dieser Vielzahl an Studien war ich an zwei Arbeiten, auf die ich mich im Folgenden konzentrieren werde, beteiligt [54, 55].

Ziel der Untersuchungen war es sowohl neue blutbasierte Frühdiagnosemarker als auch prognostische Marker für nicht kleinzellige Lungentumore (NSCLC) zu detektieren und zu validieren. Zu Beginn der Studien waren deutlich weniger miRNAs bekannt als momentan. Die rapide Entwicklung und die Entdeckung neuer miRNAs ist in Kapitel 2 und Kapitel 4.1. erläutert und in Abbildung 12 grafisch dargestellt. Zur Zeit der ersten Studien über Lungentumore waren etwa 1,000 reife miRNAs beim Menschen bekannt – zum Vergleich: heute sind es bereits 2,500 die in der miRBase stehen und etwa weitere 10,000 Kandidaten in miRCarta. Wie in Abschnitt 4.2.1. beschrieben, ist Blut ein ausgezeichnetes Ausgansmaterial um miRNAs zu messen [104]. Nicht nur, dass das Probenmaterial leicht zugänglich ist, außerdem enthält das hier verwendete Vollblut viele verschiedene Blutzelltypen, jeder mit einem eigenen komplexen miRNome [167-170]. Daher erlaubt es Blut, nicht nur Unterschiede zwischen gesunden und erkrankten Probanden zu identifizieren, sondern auch neue miRNAs zu entdecken.

Tabelle 1: 30 neue miRNAs in Lungentumorpatienten und gesunden Kontrollen

miRNA	Counts controls	Counts patients	Major sequence	Minor sequence
hsa-can-miR-163	4370	3189	TCGCATTGAACCTGAGAGGCA	CCTCCGGTATTCAAGCGATT
hsa-can-miR-277	514	493	GCCCGCCCCAGCCGAGGTT	CCCCGGCGCGGGCGGGTTC
hsa-can-miR-811	688	262	GGGCCGTGGAGGTGGACTG	GTGCACAACTGCAGGGGTGTG
hsa-can-miR-915	64	86	CTCTTCATCTACCCCCCAG	GGAGGGTGTGGAAGACAT
hsa-can-miR-49	53	91	CGTTGCCATGTCTAAGAAGAA	CTTCTTAGACATGGCAGCTTC
hsa-can-miR-473	49	60	GTCAGTTTGTCAAACTCTTT	GGAGTTGTGATCCTTTGGAGA
hsa-can-miR-571	27	74	CGCAACCCACACACGGTCTCA	AGACCGTGTGTGGGTTGCTGAG
hsa-can-miR-346	25	70	TTGGAATCCTCGCTAGAGCGT	GCTCTAGCGGGGATTCCAATA
hsa-can-miR-675	49	27	CCACAAACCTGCCAGCCCTG	GGGCGGCTATTGTGGGG
hsa-can-miR-275	46	30	TGGGTGTGGGCAGTGGGCGGGCCAAGGACA	GCAGTTGGCACCGTCCCCTGCGCCTACCCACT
hsa-can-miR-385	60	5	GGCGGGCAGCGGGTGAGGGGGTGG	GCGGGGCCCCGGACAAGGGTCCGCAGA
hsa-can-miR-213	28	33	TGCTCTTACATCTCAAACGAT	CGGTTGAGATGCAAGGGCTGC
hsa-can-miR-881	48	10	GCCCCTTTCTCAGACCCCCA	GGCCCTGGAAAGGGTCAG
hsa-can-miR-358	19	32	GCCCAGAGGATCACGGAGCCA	GCTCCTTGCACCTGTGGCTGC
hsa-can-miR-480	1	47	CTAGCAGTCTCAGGACACA	TGCCCTGAGACTGCTAAGT
hsa-can-miR-56	20	25	ATCACCACCAAACCTGTTCTTC	AGAACAGGTTTGGTGGGGATTC
hsa-can-miR-1040	20	19	GATTTCAGCGCTCTGCCCCT	GGGCAGAGCACTGTGTGTGG
hsa-can-miR-288	13	20	GGGGCAGCAGAGGACCTGGGC	CCTGATCCTCAGCTGCCCTCTC
hsa-can-miR-1011	17	15	GTCTTTTGCCCTTTCAGCT	CTGGAAGGGCAAAAGACTG
hsa-can-miR-839	14	16	GTGCCTGTGCAGAGGGAGCT	CCCCCTCCGAGCAGGCACTG
hsa-can-miR-1065	10	19	TTGGCCACCACACCTACCCCTT	GGGTGATGGGTGTGGTGTCCACAGG
hsa-can-miR-454	4	24	CCACCTTCAAAGGCACTCCG	GAGGCCTCTGCTGGTGCTG
hsa-can-miR-390	17	11	TCCTCTCCTCCCTGTGCCGAC	AAGCGCGGGGAGGGAGGATA
hsa-can-miR-23	14	14	ACCCACCTGATGCCCCGTCCCA	GGGAGGGGCAGGAGGGGTGGAATG
hsa-can-miR-152	25	2	CCTCTTCCCAGCACTCCCCT	GAGGGTTGCGGGAAGGGGGA
hsa-can-miR-555	15	11	AAAACAGGATAGGCACTAAA	TAGAGCCTATCCTGTTTTGC
hsa-can-miR-678	7	19	CGGTCCCTAACCCCCTCCGGA	CAGGGGAGGGAAGGGGAGCCGAG
hsa-can-miR-963	19	7	AGAAATTGGTTAAATTGGAGGG	GACCAATTTAACCAATTACTAT
hsa-can-miR-942	18	7	CTCTCCCCGCTTTTAACCCTA	GGGTTAAGAGTGGGGAGAAGA
hsa-can-miR-308	17	8	ACACCAAAACAATGAAAAC	TATCATTGTTTTAGTGTTT

Wenn diese miRNAs noch unterschiedlich zwischen zwei Gruppen von Probanden exprimiert sind, ist die Wahrscheinlichkeit einer biologischen Funktion und der Validität von entsprechenden miRNAs höher. In der vorliegenden Studie wurde Blut von Lungentumorpatienten (Adeno Karzinome und Plattenepithel Karzinome, Stage IA bis IIIA, alle nicht therapiert) und Kontrollen verglichen. Aus dem PAXgene Blut wurden 1.5 Mikrogramm totale RNA für kleine RNAs angereichert (Ambion's flashPAGE Fractionator) und gefällt.

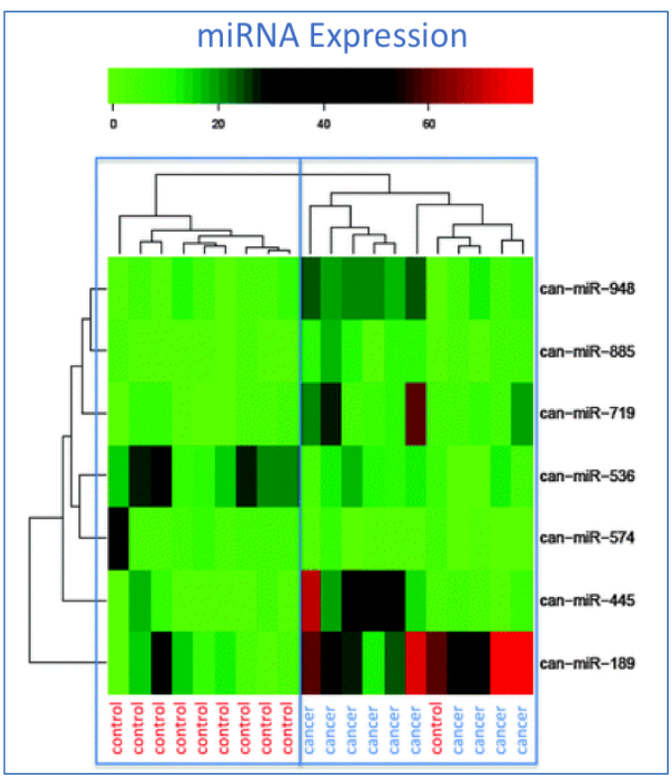


Abbildung 20: Clustering in Tumorpatienten und Kontrollen.

Die Heat Map zeigt die Expression der sieben neuen miRNAs in Kontrollprobanden und Lungentumorpatienten. Rot bedeutet hohe Expression und Grün niedrige Expression. Die vertikale blaue Linie trennt das Dendrogram in zwei Gruppen. Nur ein Tumorpatient passt von seinem Profil her zu den Kontrollen. Die Abbildung ist modifiziert aus Keller et al. entnommen.

Die Sequenzier Library wurde aus 100 Nanogramm angereicherter RNA erstellt. Nachdem die Sequenzier Adapter angefügt wurden, wurde die RNA in cDNA umgeschrieben. Die Fragmente waren dabei 60-80 Basen lang, bestehend aus den miRNAs und den Sequenzier Adaptern. Nach 15 PCR Zyklen zur Amplifikation wurde das Standard - Sequenzierprotokoll angewendet. Die Daten wurden mit RNA2MAP und eigenen Skripten in R prozessiert und neue Marker wurden gezielt mit RT-qPCR validiert. Insgesamt wurden in der Sequenzierung 530 Millionen kleine RNAs sequenziert. Von diesen konnten 352 Millionen auf das Humane Genom gemappt werden. Nur 38 Millionen dieser Reads mappten auf zu dieser Zeit bekannte miRNAs. Immerhin konnten damit 64% der miRBase abgedeckt werden.

Die hohe Zahl von Reads, die auf das Humane Genom aber nicht auf die miRBase mappen, erlaubt den Schluss, dass eine signifikante Anzahl neuer miRNAs in den Proben vorhanden sein muss. Diese Vorhersage mit dem Tool miRDeep [85, 86] hat insgesamt 210 neue miRNA Kandidaten ergeben. Da solche neuen miRNAs oft falschpositive Kandidaten enthalten, die auf Artefakte zurückgeführt werden können, [87, 171] ist

weitere Prozessierung notwendig. Eigenes Filtern der Sequenzen hat über 80 % der Kandidaten als niedrig exprimiert und wahrscheinlich falsch Positive markiert. Insgesamt sind 30 neue miRNA Kandidaten nach der Filterung als wahrscheinlich echte miRNAs verblieben. Diese sind in Tabelle 1 gezeigt, zusammen mit der Anzahl an Reads in Kontrollen und Patienten und der -3p und der -5p reifen Form. Von den 30 miRNAs, die in Tabelle 1 gezeigt sind, wurden 5 zufällig ausgewählt und mittels RT-qPCR in Blut- und Gewebeproben validiert. In allen Fällen konnten die ursprünglichen Befunde validiert werden.

Die Hauptfragestellung war jedoch, ob miRNAs zwischen Patienten und Kontrollen unterschiedlich exprimiert bzw. reguliert sind. Dazu wurden nicht nur die oben beschriebenen neuen, sondern auch die bekannten miRNAs analysiert. Da die miRNAs an sich nicht normalverteilt waren, sind die Ergebnisse des gängig verwendeten t-Tests in diesem Fall möglicherweise irreführend. Daher wurde der nicht-parametrische Wilcoxon-Mann-Whitney Test verwendet. Nach der Adjustierung für Multiples-Testen waren 70 miRNAs signifikant unterschiedlich zwischen den beiden Gruppen exprimiert. 71.4 % davon waren höher in Lungentumorpatienten. Interessanterweise waren auch 7 neue miRNAs aus Tabelle 1 signifikant unterschiedlich zwischen den Kontrollen und Patienten. Die Expressionswerte dieser 7 miRNAs in Kontrollen und Patienten sind in Abbildung 15 gezeigt. Diese Abbildung zeigt auch, dass nur ein Patient von seinem Profil her zu den Kontrollen passt. Die Genauigkeit der Zuordnung war folglich 95 %. Nicht-parametrische Permutationstests haben gezeigt, dass man mittels weiterer statistischer Lernverfahren die Genauigkeit sogar noch weiter erhöhen kann.

Die Studie hat den Nachweis erbracht, dass miRNAs aus dem Blut von Patienten und Kontrollen ein enormes Potenzial besitzen, eine Diagnose von Tumoren sogar in frühen Stadien zu erlauben. Selbst die niedrigen Grade (T1bN0) wurden korrekt erkannt. Die vielversprechenden Ergebnisse wurden inzwischen in den AGs Keller und Meese ohne meine Mitarbeit verifiziert [165]. Interessant ist auch, dass 70 % der 30 neu entdeckten miRNAs in den folgenden Versionen der miRBase annotiert wurden. In allen Fällen haben unabhängige Forscher und Arbeitsgruppen entsprechende miRNAs gefunden und in die Datenbank übernommen. In miRCarta sind bis auf wenige reife miRNAs alle Repräsentanten aus Tabelle 1 enthalten. Daher hatte die Studie, eine der ersten überhaupt, die miRNAs aus Blut bei Lungentumoren untersucht hat, doppelte Bedeutung für das Erkennen neuer miRNAs generell und für die Erkenntnis, dass miRNAs zwischen Probanden und Kontrollen unterschiedlich reguliert und damit gute Biomarker zur Früherkennung von Tumorerkrankungen sind.

Neben der Früh-Diagnostik haben miRNAs allerdings auch erhebliches Potenzial zur Prognostik bei Lungentumorpatienten gezeigt [172-177]. Hier waren es jedoch weniger Profile von Blutzellen, sondern vielmehr Serum und Plasma Profile von miRNAs [178, 179]. Im Gegensatz zu den Blutzellprofilen spiegeln Serum und Plasma Profile die Tumore direkter wieder, da auch vom Tumor sekrierte miRNAs gemessen werden können. Dabei ist zu beachten, dass das Messen von miRNAs aus Serum und Plasma kontrovers diskutiert wird [180, 181]. Entsprechend ist es wichtig die möglichen vorhandenen Quellen für Fehler und Artefakte zu kennen, um ihnen bestmöglich vorzubeugen.

Um zu untersuchen, inwieweit miRNAs prognostische Information untersuchen und ob sie parallel und longitudinal zu einer Therapie gemessen werden können, wurden 26 Patienten über einen Zeitraum von bis zu 18 Monaten nach Tumordiagnose und Resektion untersucht [55]. Insgesamt wurde den Probanden zu 8 Zeitpunkten innerhalb dieser 18 Monate Blut abgenommen. Als Kontrolle wurden Patienten selektiert, die an anderen Erkrankungen der Lunge leiden. Zunächst wurde die Komplexität des miRNomes untersucht, also die Anzahl an detektierten miRNAs in Kontrollen und Tumorpatienten zu den jeweiligen Zeitpunkten. Insgesamt hat sich gezeigt, dass Tumorpatienten ein reduziertes miRNome im Vergleich zu Patienten mit anderen Lungenerkrankungen haben. Die Zahl der miRNAs hat dabei im Verlauf der Therapie stark geschwankt.

Das miRNA Repertoire wurde auch mit der Entwicklung von Metastasen korreliert, um prognostische Information zu erhalten. Hier hat sich gezeigt, dass Patienten die Metastasen entwickeln ein deutlich komplexeres miRNome hatten, im Vergleich zu Personen die keine Metastasen entwickelt haben. Die Unterschiede in der Komplexität des miRNomes waren statistisch signifikant, ein ungepaarter t-Test hat einen p-Wert von 0.0096 ergeben. Neben dieser eher qualitativen Analyse wurde eine eher quantitative Analyse der miRNA Expressionswerte durchgeführt. Zunächst wurden paarweise t-Tests zwischen Kontrollen und Tumorpatienten zu jedem der 8 Zeitpunkte durchgeführt. Anschließend wurden die p-Werte logarithmiert mit dem Rang der 8 Zeitpunkte verglichen.

Eine negative Korrelation bedeutet dabei, dass eine miRNA im Verlauf der 18 Monate kontinuierlich stärker dysreguliert wird. Eine positive Korrelation bedeutet hingegen, dass die miRNA sich analog kontinuierlich in der Expression den Kontrollen angleicht. Die Analyse hat insgesamt 6 negativ korrelierte und 28 positiv korrelierte miRNAs ergeben. Folglich hat sich die Mehrzahl der Marker im Verlauf der Therapie an das Kontrollniveau angeglichen. Die miRNAs, die im Verlauf der 18 Monate zunehmend stärker vom Kontrolllevel abgewichen sind, sind hsa-miR-181d, hsa-miR-670, hsa-miR-196b, hsa-miR-3148, hsa-miR-762 und hsa-miR-539. Die miRNAs, die sich im Verlauf der

longitudinalen Analyse am stärksten dem Level der Kontrollen angepasst haben, waren hsa-miR-184, hsa-miR-141, hsa-miR-4281, hsa-miR-454 und hsa-miR-301a.

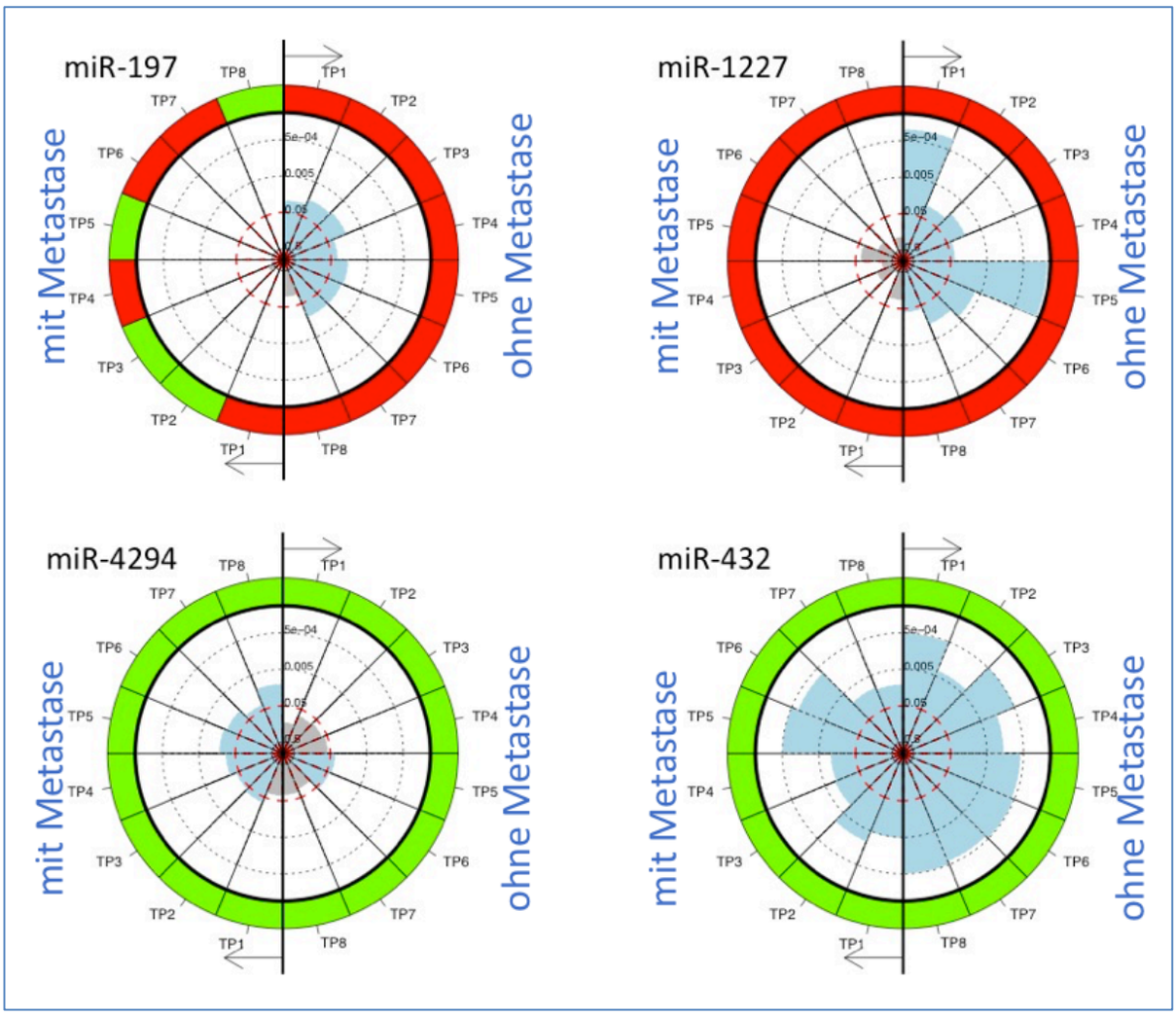


Abbildung 21: Zeit / Metastasendiagramme für 4 ausgewählte miRNAs.

Die Abbildungen enthalten im inneren Kreis die Information wie signifikant die miRNA zum jeweiligen Zeitpunkt war (je größer der blaue Teil desto signifikanter, alles außerhalb der rot gestrichelten Linie war signifikanter als der Alpha Level von 0.05). Der äußere Kreis zeigt die Richtung der Regulation (grün bedeutet weniger exprimiert im Vergleich zu Kontrollen, rot bedeutet höher exprimiert). Der rechte Teil der Kreise entspricht den 8 Zeitpunkten bei Patienten die keine Metastasen entwickelt haben, der linke Teil den 8 Zeitpunkten bei Patienten die Metastasen entwickelt haben. Die Abbildungen wurden modifiziert aus Leidinger et al. übernommen.

Der primäre Endpunkt in der Studie war das Entwickeln einer Metastase. Daher wurden die Profile der 8 Probanden die eine Metastase entwickelt haben zu jedem der 8 Zeitpunkte zu den 18 Patienten verglichen, die keine Metastase entwickelt haben. Um die Vergleiche besser durchführen zu können wurden dabei die unadjustierten p-Werte verwendet. Zum Ausganszeitpunkt (vor der Resektion der Tumore) waren insgesamt 25 miRNAs signifikant unterschiedlich exprimiert. Direkt nach der Operation ist die Anzahl auf 18 miRNAs gesunken (davon 4 überlappend). Zum Zeitpunkt drei waren die Unterschiede am deutlichsten, 40 miRNAs waren signifikant unterschiedlich. Im

Folgenden hat sich das Niveau zwischen Patienten die Metastasen entwickeln und solchen die keine Metastasen entwickeln wieder angeglichen. Insgesamt waren unabhängig der Zeitpunkte 131 miRNAs signifikant unterschiedlich zwischen den beiden Gruppen, nach Adjustierung für Multiples-Testen waren noch 38 davon signifikant. Der geringste p-Wert wurde für hsa-miR-197 berechnet ($p = 3\times10-7$). Diese miRNA war zu drei der Zeitpunkte signifikant (TP2, TP3, TP5). Eine weitere miRNA, hsa-miR-630 war sogar zu vier Zeitpunkten signifikant (TP1, TP2, TP4 und TP6), zu den jeweiligen Zeitpunkten allerdings etwas schwächer als miR-197.

Das verhältnismäßig komplexe Studien Set-Up das viele Analysen ermöglicht, macht es schwer, Abbildungen zu generieren die intuitiv und gleichzeitig interpretierbar sind ohne relevante Information zu verlieren. Konkret werden Patienten die Metastasen entwickeln zu solchen verglichen die keine Metastasen entwickeln und zu Kontrollen die keine Tumore haben. Die Vergleiche wurden zu acht Zeitpunkten durchgeführt; miRNAs können also in bis zu acht Zeitpunkten in Patienten mit und ohne Metastasen jeweils signifikant hoch- oder runterreguliert sein. Um diese Information übersichtlich darzustellen wurden spezielle Abbildungen basierend auf Kreisdiagrammen entwickelt. Der innere Teil des Kreises repräsentiert den negativen Logarithmus des p-Wertes für jeden Vergleich. Die farbliche Darstellung im äußeren Kreis repräsentiert die Richtung der Regulation (grün bedeutet runterreguliert und rot bedeutet hochreguliert). Der jeweils rechte Teil der Kreise entspricht den 8 Zeitpunkten bei Patienten die keine Metastasen entwickelt haben und der linkte Teil, den jeweiligen Zeitpunkt der Patienten, die Metastasen entwickelt haben. Durch diese Darstellung kann für jeweils eine miRNA der fast komplette Informationsgehalt grafisch dargestellt werden.

Abbildung 21 zeigt für vier miRNAs die entsprechenden Diagramme. Die vorher erwähnte miR-197 ist in Patienten ohne Metastasen fast zu allen Zeitpunkten signifikant höher vorhanden, im Vergleich zu Probanden ohne Tumorerkrankungen. Im Falle von Patienten mit Metastasen schwankt die Richtung der Regulation, ist allerdings zu keinem Zeitpunkt signifikant. Ähnliches gilt für miR-1227, hier schwankt die Regulationsrichtung nicht bei Patienten mit Metastasen, ansonsten verhält die miRNA sich aber gleich wie miR-197. Analog ist miR-4294 bei Patienten die Metastasen entwickeln oft weniger exprimiert. Eine weitere Beobachtung ist, dass miR-432, unabhängig ob Patienten Metastasen entwickeln oder nicht, in Tumoren immer niedriger exprimiert als in Kontrollen.

Auch die Ergebnisse dieser Studie wurden mittels RT-qPCR validiert. Obwohl die Ergebnisse an verhältnismäßig kleinen Kohorten durchgeführt wurden, scheinen die Resultate vielversprechend. Insbesondere das longitudinale Studien Set-Up, welches auch gepaarte statistische Analysen erlaubt, trägt dazu bei, dass überzeugende Ergebnisse

erzielt wurden. Natürlich bedarf es hier, wie auch in der diagnostischen Studie, weiterer unabhängiger Validierung, bis ein entsprechender Test im klinischen Kontext eingesetzt werden kann.

Vor allem die Ergebnisse im Umfeld der frühen Diagnostik von Lungentumoren waren bisher so positiv, dass eine weitere Validierung in den AGs Humangenetik und Klinische Bioinformatik durchgeführt wurde. Diese Validierung der ursprünglichen Ergebnisse, die ohne meine direkte Mitarbeit erfolgte, hat gezeigt, dass die Ergebnisse selbst an einem Kollektiv von 3,000 Probanden (Lungentumorpatienten, gesunde Kontrollen und Patienten mit anderen Erkrankungen wie COPD) Bestand haben. Ein entsprechendes Manuskript wird zur Publikation vorbereitet.

Publikationen: Die Arbeiten, die in diesem Abschnitt beschrieben werden basieren hauptsächlich auf [54, 55]. Daneben wurde basierend auf der Methodik im ersten Manuskript ein Patent angemeldet, das zeigt wie neue miRNAs im Vergleich von Probanden und Kontrollen besser gefunden werden können (US201314442858 20131104).

4.2.3. Diagnose von Multipler Sklerose & Alzheimer

Wie im einleitenden Absatz zu Kapitel 4.2 beschrieben und später im Abschnitt 4.2.4. im Detail diskutiert wird, ist es wichtig, nicht nur zu verstehen, ob eine miRNA zwischen Patienten einer Erkrankung und Kontrollen unterschiedlich exprimiert ist, sondern es ist gleichermaßen wichtig zu verstehen, wie spezifisch einzelne miRNAs oder miRNA Signaturen für eine bestimmte Erkrankung sind. Dabei ist es sinnvoll, die richtigen Kontrollen zu wählen. Bei Lungentumorpatienten macht es zum Beispiel Sinn, nicht nur gesunde Probanden als Kontrollen zu betrachten, sondern auch Patienten mit anderen Lungenerkrankungen wie COPD. Darüber hinaus kann es außerdem nützlich sein, andere Erkrankungen mit in die Betrachtung einzubeziehen. Neben dem Punkt der Spezifität der Signaturen dient dies auch dem Verständnis, ob und wie weit miRNA Signaturen aus dem Blut über ein bestimmtes Krankheitsbild hinaus generalisiert werden können.

Neben Lungentumoren habe ich mich mit Erkrankungen des Zentralen Nervensystems befasst und sowohl Multiple Sklerose als auch Alzheimer in meine Forschung eingeschlossen [56-58].

Multiple Sklerose: Bei der Multiplen Sklerose (MS) wurden sowohl Kontrollen (n=50) als auch Patienten mit klinisch isoliertem Syndrom (CIS, N=25) und Relapsing-Remitting Multiple Sclerosis (RRMS, n=25) Patienten eingeschlossen. Ziel der Studie war es, molekulare Marker zu finden, die gängige und etablierte Kriterien [182, 183] in der Diagnose von MS verbessern. Vor allem bei CIS und bei atypischen Formen der MS kann ein entsprechender Marker von entscheidendem Vorteil sein [184]. Da wie in Absatz 4.1. beschrieben, verschiedene Plattformen ihre jeweiligen Stärken und Schwächen haben, wurde sowohl Mikroarray Technologie als auch HTS verwendet, um ein möglichst umfangreiches Bild zu erhalten. Die Signaturen wurden außerdem mittels RT-qPCR validiert. Als Ausgangsmaterial wurden PAXgene Blutproben von den 100 Individuen verwendet. Die experimentellen Methoden sind analog zu den vorher beschriebenen Studien durchgeführt worden.

In den HTS Experimenten wurden insgesamt 835 miRNAs detektiert. 38 dieser miRNAs waren signifikant unterschiedlich zwischen MS Patienten und Kontrollen exprimiert. Darunter waren 16 mit geringeren Levels in MS Patienten und 22 mit höheren Levels. Die acht am stärksten dysregulierten miRNAs hatten besonders hohe Effektgrößen. Sie enthalten die fünf geringer exprimierten hsa-miR-361-5p, hsa-miR-7-1-3p, hsa-miR-548o-3p, hsa-miR-151a-3p, und hsa-miR-548am-3p sowie die drei höher exprimierten hsa-miR-22-5p, hsa-miR-27a-5p und hsa-miR-4677-3p. In der mikroarraybasierten Analyse wurden deutlich weniger miRNAs gefunden: nur etwa jede zweite miRNA aus den HTS Experimenten konnte gemessen werden. Insgesamt waren im Gruppenvergleich acht miRNAs signifikant. Fünf miRNAs die schwächer in MS Patienten waren (hsa-miR-146b-5p, hsa-miR-7-1-3p, hsa-miR-20a-5p, hsa-miR-3653, hsa-miR-20b) und drei, die stärker waren (hsa-miR-16-2-3p, hsa-miR-574-5p, hsa-miR-1202).

Wenn man davon ausgeht, dass im HTS Experiment 1.9 % der miRNAs signifikant waren und im Mikroarray Experiment 0.7 %, ist eine zufällige Überlappung der beiden Sets an miRNAs relativ unwahrscheinlich. Trotzdem stimmten drei miRNAs zwischen beiden Sets überein (hsa-miR-16-2-3p, hsa-miR-20a-5p und hsa-miR-7-1-3p; p-Wert für die Überlappung entspricht 0.004). Keine der miRNAs zeigte eine Korrelation mit der Form der MS, die Level zwischen CIS und RRMS Patienten waren nicht signifikant unterschiedlich. Um Informationen über die Spezifität der miRNAs für Erkrankungen zu erlangen, haben wir eine öffentliche Datenbank, die HMDD (Human miRNA Disease Database), abgefragt. Diese Analyse hat gezeigt, dass die oben beschriebenen miRNAs tatsächlich gehäuft in anderen Erkrankungen vorkommen. Für alle bis auf eine miRNA konnte eine Korrelation zu mehr als acht verschiedenen Erkrankungen hergestellt werden. Die miRNA, die daher am spezifischsten für MS ist, ist miR-16-2.

Zusammenfassend wurden Signaturen für MS gefunden, die es erlauben die bisherige Diagnose nach jetzigem Kenntnisstand mit Hilfe von miRNAs zu verbessern. Erstaunlicherweise waren die Unterschiede zwischen der CIS und der RRMS Form der MS relativ gering und statistisch nicht signifikant. Das kann allerdings an den verhältnismäßig kleinen Gruppen je untersuchter MS Art liegen. Während für MS insgesamt 50 Fälle untersucht wurden, waren es je Subgruppe nur 25 Fälle. Ein weiterer Punkt, der genauer untersucht werden muss, ist die Spezifität der Signatur für Multiple Sklerose. Viele der gefundenen miRNA Marker wurden auch in anderen Krankheiten entdeckt. Alleine miR-16-2 war sehr spezifisch und die Signatur die identifiziert wurde, wurde als solche in keiner anderen Erkrankung in ähnlicher Form gefunden.

Alzheimer (AD): Als zweite Erkrankung des Zentralen Nervensystems habe ich die Alzheimer Erkrankung näher betrachtet. Alzheimer ist eine Volkskrankheit, die uns in den nächsten Jahren und Jahrzehnten noch deutlich stärker betreffen wird. Bereits 2015 lebten weltweit etwa 47 Millionen Menschen mit Alzheimer. Bis 2050 wird sich diese Zahl laut aktueller Progonsen etwa verdreifachen. Die Entwicklung der Therapien für Alzheimer ist in den vergangenen Jahren fast stagniert [185-188]. Führende Pharmafirmen ziehen sich teilweise sogar aus der Forschung an Medikamenten für Alzheimer oder sogar Neurodegeneration insgesamt zurück. Ein Beispiel ist das Pharmaunternehmen Eli Lilly, das im Januar 2017 verkündet hat, seine Bemühungen in diesem Umfeld weitestgehend einzustellen. Eine der Hauptherausforderungen ist es dabei, dass die Patienten oft zu spät erkannt werden und mittels bildgebender Verfahren wie MRT diagnostiziert werden. Ein Überblick über die momentanen Diagnoseverfahren mit Schwerpunkt auf molekulare Diagnostik und zirkulierende Biomarker Panels ist von Zafari publiziert worden [189]. Auch bei der Alzheimer Erkrankung wurde das Potenzial zirkulierender miRNA Muster, analog der Verfahren bei Lungentumoren und Multipler Sklerose, getestet [56, 57].

In einer ersten Studie wurden Alzheimer Samples und Kontrollen, die in Bezug auf Alter und Geschlecht zugeordnet waren, auf ihr Repertoire an miRNAs im Blut mit HTS charakterisiert. Die experimentellen Methoden waren dabei wieder identisch zu der oben beschriebenen MS Studie. Die Patientenproben wurden von der SAMPLE Studie (Serial Alzheimer Disease and MCI Prospective Longitudinal Evaluation) von PrecisionMed (San Diego, CA, USA) erhalten und einer ausführlichen Standard Diagnostik (inklusive MRT und Mini-Mental State Exam MMSE) unterzogen. In dieser Studie wurden 416 reife miRNAs detektiert. Da die Read Zahlen wieder nicht normalverteilt waren, wurde der nicht-parametrische Wilcoxon-Mann-Whitney (WMW) Test angewendet und die p-Werte

wurden mittels der Benjamini-Hochberg (BH) Methode adjustiert. Die Analysen haben 180 dysregulierte miRNAs zwischen Patienten und Kontrollen ergeben, davon waren 90 jeweils höher bzw. tiefer bei Alzheimer Patienten im Vergleich zu Kontrollen. Eine netzwerkbasierte Analyse hat ergeben, dass unter den 180 miRNAs auch die Krankheitskategorie „Alzheimer miRNAs“ stark überrepräsentiert war (p=0.01). Die sechs miRNAs hsa-miR-21, hsa-miR-17, hsa-miR-29a, hsa-miR-29b, hsa-miR-106b und hsa-miR-107 gehörten alle dieser Kategorie an. Neben den p-Werten wurde auch die AUC, die Area Under The Curve, als weiteres Kriterium für die diagnostische Qualität der Biomarker berechnet. Bereits einzelne miRNAs hatten ausgezeichnete Werte mit AUC's über 0.91 und daher nahe dem Optimum von 1. Die am meisten überexprimierte miRNA war miR-30d-5p, die am meisten nach unten regulierte miRNA war miR-144-5p. Die p-Werte waren dabei jeweils $8x10^{-6}$. Auch diese beiden miRNAs sind nicht spezifisch, neben AD wurden sie in vielen anderen Erkrankungen (auch in MS, siehe oben) beschrieben.

Diese Resultate werfen zwei Fragen auf: Kann die Genauigkeit der Vorhersage durch die Kombination von miRNAs in Signaturen verbessert werden und sind diese Signaturen dann auch spezifisch für AD. Um diese Frage zu beantworten wurden zunächst maschinelle Lernverfahren (ML) verwendet. Die besten Ergebnisse haben Support Vector Machines (SVM) mit Radialer Basis Funktion als Kernel gezeigt. Die Ergebnisse der Klassifikation mittels SVM ist in Abhängigkeit der Anzahl an miRNAs in der Signatur in Abbildung 22 gezeigt. Mit steigender Anzahl erhöht sich auch die Genauigkeit der Vorhersage. Mit 250 miRNAs werden 90 % aller Proben korrekt zugeordnet.

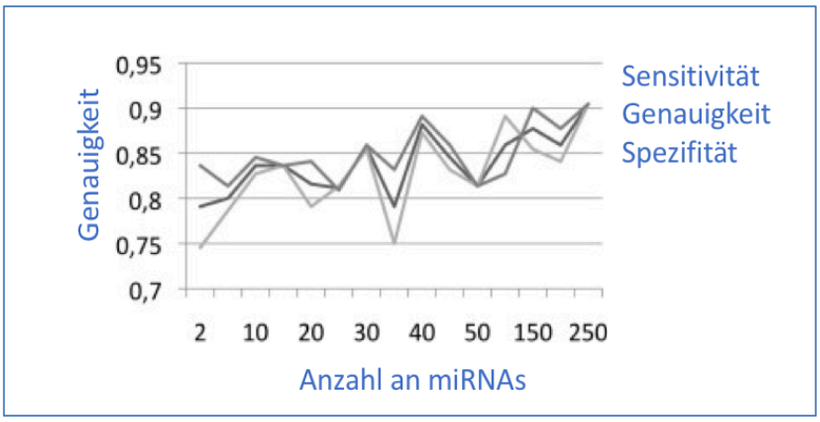


Abbildung 22: Genauigkeit in Abhängigkeit der Anzahl an miRNAs in der AD Signatur.

Die Abbildung zeigt die Spezifität, Sensitivität und die Testgenauigkeit des AD Tests in Abhängigkeit der Anzahl an miRNAs in der Signatur. Die Abbildung ist modifiziert aus Leidinger et al. entnommen.

Viele der miRNAs in der 250 Marker-Signatur haben jedoch eine hohe Redundanz gezeigt. Daher war es möglich mit substanziell kleineren Sets bereits ähnlich gute Resultate zu erzielen und dabei gleichzeitig die Gefahr des Overfittings zu reduzieren. Bereits mit 12 Markern war es möglich, eine Spezifität und Sensitivität von 85 % zu erzielen. Diese 12-Marker-Signatur besteht aus brain-miR-112, brain-miR-161, hsa-let-7d-3p, hsa-miR-5010-3p, hsa-miR-26a-5p, hsa-miR-1285-5p und hsa-miR-151a-3p (höher exprimiert bei

AD Patienten) und hsa-miR-103a-3p, hsa-miR-107, hsa-miR-532-5p, hsa-miR-26b-5p, und hsa-let-7f-5p (niedriger exprimiert in AD Patienten). Neben 10 bekannten miRNAs waren auch zwei bisher nicht bekannte miRNAs in der Signatur, die anstatt des typischen Vorsatzes „hsa-miR“ mit „brain-miR“ gekennzeichnet sind.

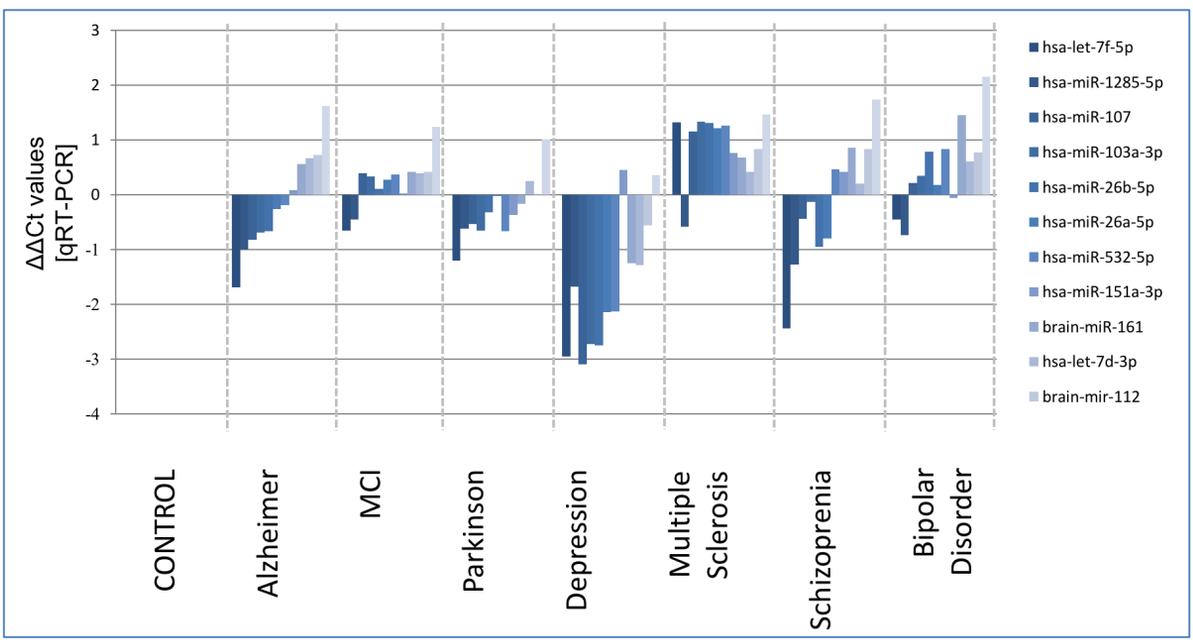


Abbildung 23: AD Signatur in anderen Erkrankungen.

Die Abbildung zeigt die 12-miRNA Signatur die bei AD Patienten gefunden wurde im Verhältnis zu anderen Erkrankungen. Die Balkenhöhe entspricht dabei der Expression der miRNAs. Jede Erkrankung hat dabei deutlich sichtbar ihre eigene Signatur. Die Abbildung ist modifiziert aus Leidinger et al. entnommen.

Um zu verstehen, wie spezifisch die Signatur für AD ist, wurde eine Kohorte von 202 Patienten verschiedener Erkrankungen mittels RT-qPCR auf diese Signatur hin untersucht. Die Erkrankungen die dabei betrachtet wurden, waren neben AD auch Parkinson (PD) Schizophrenie (Shiz), Bipolare Störung (BD), Mild Cognitive Impairment (MCI) und MS. Für Patienten aller Erkrankungen wurden die miRNA-Signaturen der 12 oben genannten miRNAs erhoben. Diese sind in Abbildung 23 gezeigt. Hier ist anzumerken, dass die Werte für gesunde Probanden verwendet wurden, um eine Normalisierung auf ein Ausgangsniveau zu ermöglichen. Wie in Abbildung 23 gezeigt ist, haben alle Erkrankungen spezifische Muster der 12-Marker miRNA-Signatur. Speziell in MS und Depression wurden signifikant andere Muster nachgewiesen, entweder waren alle miRNAs deutlich höher oder deutlich niedriger exprimiert. Besonders bei Patienten die unter Depression leiden wurden deutlich niedrigere Werte der miRNAs aus der 12-er Signatur nachgewiesen. Erstaunlich war auch, dass die AD Patienten fast genauso gut von MCI Patienten getrennt werden konnten wie von gesunden Kontrollen. Zwischen verschiedenen Graden der AD Erkrankung (Patienten mit MMSE >19 wurden als milde Form und Patienten mit MMSE 12-19 wurden als moderate Form betrachtet) zeigten sich

hingegen keine unterschiedlichen Signaturen.

Insgesamt wurde mit der 12-Marker-Signatur eine relativ genaue Diagnostik von Alzheimer in frühen Stadien ermöglicht. Da die Signatur es erlaubt MCI von AD Patienten abzugrenzen, scheint sie sehr spezifisch für AD zu sein.

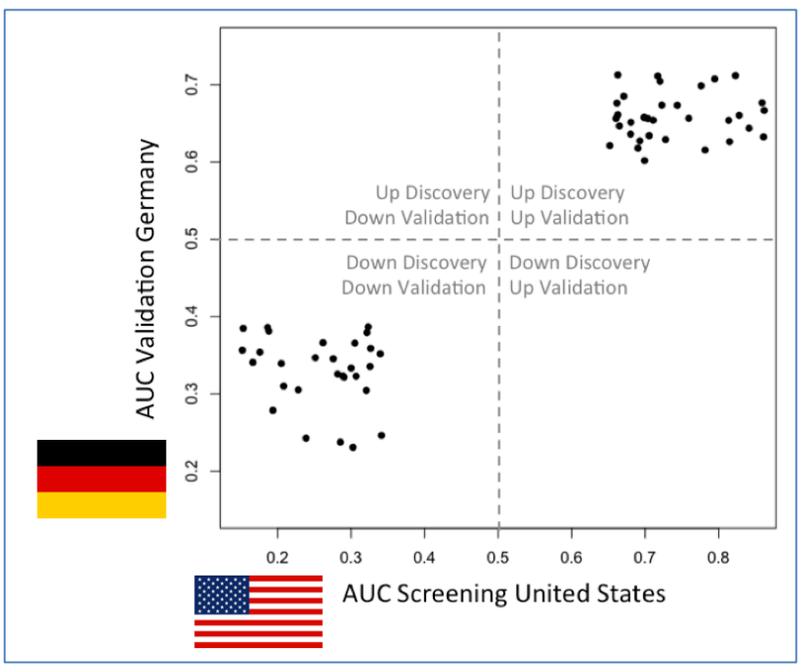


Abbildung 24: AD miRNAs in den USA und Deutschland.

Die Abbildung zeigt die AUC Werte der miRNAs in den USA (X-Achse) und Deutschland (Y-Achse). Die AUC Werte sind so gewählt dass AUC > 0.5 Hochregulation und AUC < 0.5 nach unten Regulation entspricht. AUC Werte von genau 0.5 bedeuten, dass die miRNAs nicht unterschiedlich zwischen Patienten und Kontrollen sind (horizontale und vertikale gestrichelte Linie). Alle 68 miRNAs waren konkordant, entweder hoch-exprimiert sowohl in AD Patienten in den USA und Deutschland (oberer rechter Quadrant) oder niedrig-exprimiert sowohl in AD Patienten in den USA und Deutschland (unterer linker Quadrant). Diskordante miRNAs (die in beiden Kohorten in jeweils die entgegengesetzte Richtung exprimiert wären) wären in den oberen linken bzw. unteren rechten Quadranten zu finden. Die Abbildung ist modifiziert aus Keller et al. entnommen.

Natürlich ist es wichtig, entsprechende komplexe Signaturen auch in einer weiteren Kohorte zu validieren. Da die Patienten in der ursprünglich verwendeten Kohorte aus den USA stammen, war es sinnvoll, eine nicht aus Amerika stammende Validierungskohorte zu wählen. Als zweite Kohorte wurden Patienten und Kontrollen die in Deutschland gesammelt wurden gemessen und mit den ursprünglichen Signaturen verglichen. Insgesamt wurden 290 HTS miRNA-Profile in diese Analyse einbezogen. Die Methodik war dabei wieder identisch zu der vorher beschriebenen MS und initialen AD-Studie.

Zusammengenommen wurden 3.85 Milliarden Reads in der Studie analysiert und dabei 580 miRNAs detektiert. Die gemeinsame Analyse der Daten hat ergeben, dass in der ersten Kohorte (USA) 203 dysregulierte miRNAs vor und 127 dysregulierte miRNAs nach Adjustierung für Multiples-Testen gefunden wurden. In der Kohorte aus Deutschland waren 146 miRNAs dysreguliert bevor adjustiert wurde, 49 nach der Adjustierung. Von den 203 und respektive 146 miRNAs stimmten 68 überein. Die Gesamtzahl an exprimierten miRNAs betrug 580. Ein hypergeometrischer Test hat ergeben, dass die

Überlappung statistisch sehr signifikant war (p=0.0003). Noch entscheidender war, dass wie in Abbildung 24 gezeigt alle miRNAs konkordant waren. Sie waren sowohl in den USA als auch in Deutschland entweder niedriger oder höher exprimiert, bei AD Patienten verglichen zu Kontrollen.

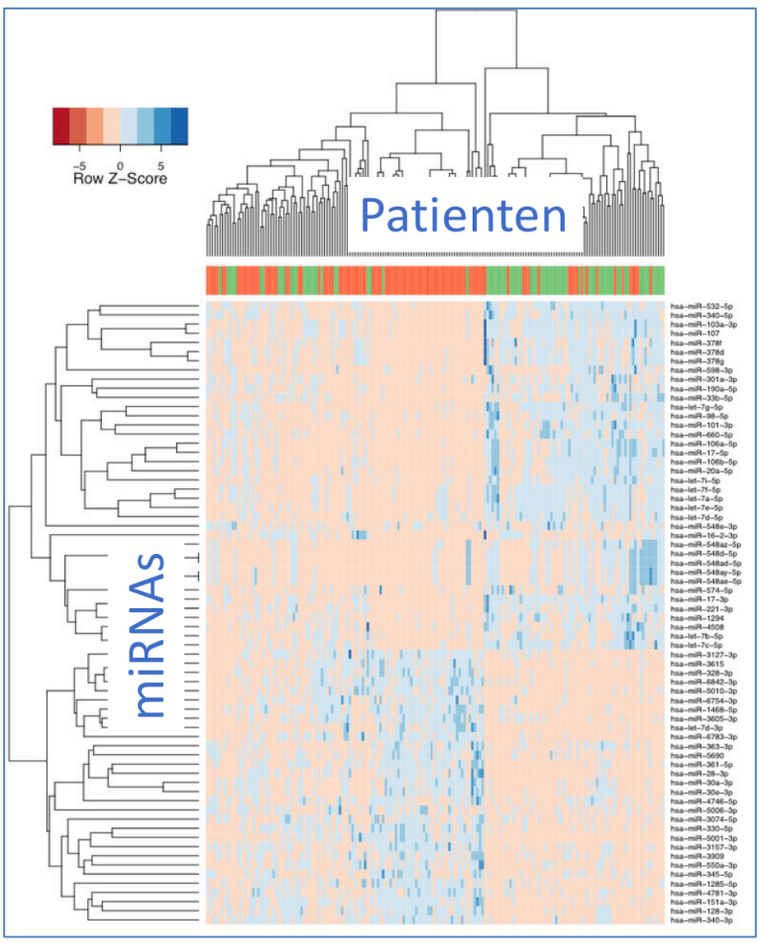


Abbildung 25: Clustering der 69 AD miRNAs.

Blaue Werte entsprechen hoher Expression, orangene Werte entsprechend niedriger Expression. Über der Matrix und auf der linken Seite der Matrix ist jeweils ein Dendrogram gezeichnet, das zeigt, wie gut miRNAs (Zeilen) und Patienten (Spalten) zusammen clustern. Die Abbildung ist modifiziert aus Keller et al. entnommen.

Insgesamt konnten die miRNA Muster verwendet werden, um AD Patienten von Kontrollen mit einer Genauigkeit von etwa 90 % zu trennen. Grafisch ist diese Trennung in Abbildung 25 gezeigt. Für die signifikanten miRNAs in beiden Studien wurde eine sogenannte Heat Map erzeugt. Sie ist das Ergebnis eines hierarchischen Clusterings unter Verwendung der Euklidischen Distanz. Analog zur ersten Studie über Alzheimer miRNAs war auch hier die Genauigkeit deutlich besser, wenn Signaturen anstatt von einzelnen miRNAs verwendet wurden. Während einzelne Marker eine AUC von 0.75 gezeigt haben, war es möglich unter Verwendung von SVM Klassifikation eine AUC von 0.842 zu erreichen. Mit verbesserter statistischer Analyse war es sogar möglich, die Klassifikatoren an Hand von Daten aus den USA zu trainieren und AD an deutschen Patienten vorherzusagen. Hier lag die Genauigkeit bei immerhin noch 73 % (Daten nicht gezeigt). Dieser erste größere Datensatz hat es außerdem ermöglicht, Effekte der miRNAs auf ihre Zielgene und Zielnetzwerke abzuschätzen. Siehe auch nachfolgendes Kapitel 4.3.

Zusammenfassend kann man sagen, dass es gelungen ist eine miRNA Signatur die für Alzheimer spezifisch ist zu detektieren und an Hand von zwei diversen Kollektiven auf zwei verschiedenen Kontinenten zu validieren.

Publikationen: Die Arbeiten, die in diesem Abschnitt beschrieben werden basieren hauptsächlich auf den beiden Publikationen [56, 57]. Zusätzlich wurden zwei Patente angemeldet, um miRNA basierte Diagnostik aus dem Blut bei AD Patienten durchzuführen (WO2017108535, US2016273040).

4.2.4. Das „Disease miRNome“

Zentraler Aspekt in allen vorherigen Abschnitten war die Spezifität von miRNAs für Erkrankungen. In Kapitel 4.2.2 habe ich Signaturen in Lungentumoren und COPD beschrieben und in Kapitel 4.2.3 von Erkrankungen des Zentralen Nervensystems. Schon dabei wurden Überlappungen in der miRNA-Expression gefunden, obwohl die betrachteten Pathologien deutlich unterschiedlich voneinander sind. Das lässt auf eine gemeinsame, unspezifische Komponente von miRNAs im Blut von Patienten im Vergleich zu Kontrollen schließen. Direkte Vergleiche basierend auf Daten aus der Literatur in sogenannten Metaanalysen sind schwierig. Oft werden unterschiedliche Systeme zum Sammeln des Blutes verwendet (EDTA / PAXgene / ...), verschiedene RNA Aufreinigungs-methoden werden eingesetzt, andere analytische Methoden (HTS / Mikroarray / RT-qPCT / ...) werden angewendet und die erhobenen Daten mit unterschiedlichen Methoden ausgewertet. Um einen besseren Vergleich zu ermöglichen ist die beste Alternative eine Studie aufzusetzen, die auf exakten „Standard Operating Procedures (SOPs)“ basiert. Ich war Teilnehmer eines großen Konsortiums von über 50 Wissenschaftlern, das SOPs zum Sammeln, Messen und Auswerten von Blut basierten miRNA-Signaturen entwickelt und an über 30 Erkrankungen getestet hat [34].

Insgesamt wurden in der entsprechenden Studie 454 Blutproben aus fünf Zentren eingeschlossen. Die Kohorten beinhalten Lungentumore, Prostatatumore, Pankreastumore, Melanom, Eierstockkrebs, Magentumore, Wilmstumore,

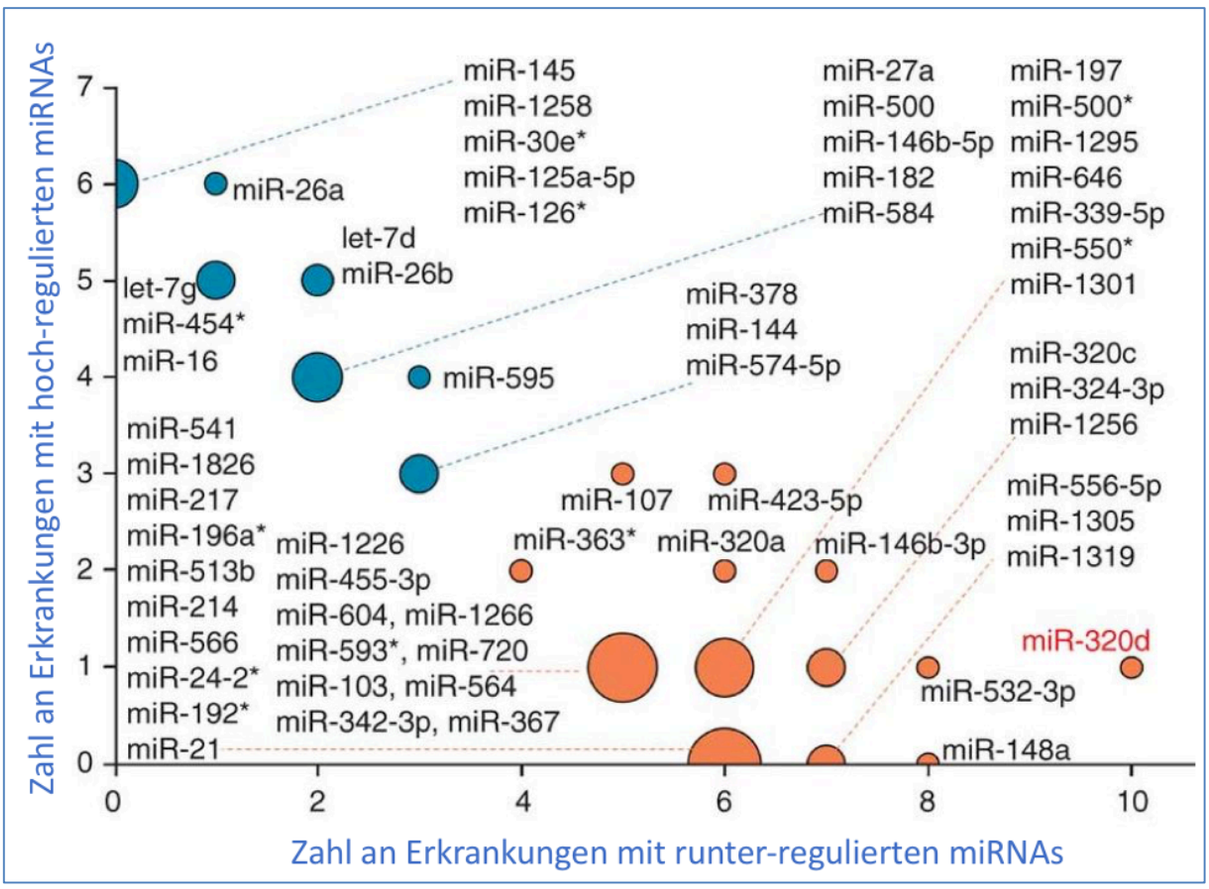


Abbildung 26 : Das Disease miRNome.

Die Abbildung zeigt für jede mögliche Kombination von höher zu niedriger exprimierten miRNAs in wie vielen Erkrankungen die miRNAs entsprechend reguliert waren. Die Größe der Bubbles entspricht dabei der Anzahl der miRNAs. Die Abbildung ist modifiziert aus Keller et al. entnommen.

Pankreaskrebs, Multiple Sklerose, COPD, Sarkoidose, Periodontitis, Pankreatitis und Herzinfarktpatienten. Jedes der teilnehmenden Zentren musste neben den Patienten auch Kontrollen ohne bekannte Erkrankung zur Verfügung stellen.

Für alle Patienten und Kontrollen wurden 863 miRNAs aus Vollblut mittels Mikroarrays gemessen. Die vergleichende Analyse der Erkrankungen hat ergeben, dass im Durchschnitt 103 miRNAs je Erkrankung nach Adjustierung für Multiples-Testen signifikant waren. 62 miRNAs waren dabei in mindestens sechs verschiedenen Erkrankungen signifikant. Drei miRNAs wurden in neun Vergleichen gefunden (hsa-miR-423-5p, hsa-miR-146b-3p und hsa-miR-532-3p), eine miRNA sogar in 11 (hsa-miR-320d). Gerade einmal 121 miRNAs waren mit keiner einzigen Erkrankung signifikant assoziiert. Erstaunlich war außerdem, dass die Regulationsrichtung oft konkordant zwischen den verschiedenen Erkrankungen war, miRNAs waren entweder generell höher bei Patienten als in Kontrollen vorhanden oder generell niedriger bei Patienten im Verhältnis zu Kontrollen. Das „Disease miRNome“, das diesen generellen Bezug darstellt, ist in Abbildung 26 gezeigt.

Um die notwendige Spezifität für die Diagnose von Erkrankungen zu bekommen wurden Maschine-Learning-Verfahren analog zu den in der Alzheimer und Multiplen Sklerose Studie vorgestellten Methoden verwendet: Klassifizierung mittels Support Vektor Machines. Für die 14 Erkrankungen wurde eine diagnostische Genauigkeit von mindestens 81 % erlangt, teilweise bis zu 100 %, zum Beispiel für bösartigen Hautkrebs. Die mittlere Genauigkeit über alle Erkrankungen die getestet wurden hinweg, lag bei 89 %. Schon Subsets von nur 10 miRNAs haben ausgereicht, um eine mittlere Genauigkeit von 81 % zu erzielen. Eine unabhängige, gezielte Validierung mittels RT-qPCR, unter Verwendung des WaferGen Systems, hat die Ergebnisse in dieser Studie am Beispiel von Lungentumoren und COPD verifiziert. Insgesamt bestätigen die Ergebnisse die Resultate aus den vorangegangenen Abschnitten. Während einzelne miRNAs diagnostisches Potenzial besitzen aber nicht spezifisch sind, erlaubt es die Kombination von miRNAs zu kleineren Sets von etwa 10 Markern, sowohl die Genauigkeit zu verbessern als auch die Spezifität zu erhöhen.

Publikationen: Die Arbeiten, die in diesem Abschnitt beschrieben werden basieren hauptsächlich auf [34] und wurden in einer weiteren Publikation verifiziert [190].

4.2.5. Zusammenfassung Biomarker Entwicklung

Die in vielen Originalarbeiten beschriebenen und in dieser Ausarbeitung zusammengefassten miRNA-Signaturen sind vielversprechende Biomarker für Erkrankungen. Nach der Charakterisierung der Signaturen in der Grundlagenforschung und dem Gewährleisten der Stabilität sind jetzt zwei weitere maßgebliche Schritte notwendig. Der erste ist es, die biologische Funktion und Wirkungsweise der miRNAs und miRNA-Signaturen besser zu verstehen. Erste kleine Schritte dazu sind im nächsten Abschnitt gezeigt. Daneben muss eine größere multizentrische Validierung durchgeführt werden. Im Falle von Alzheimer wurde die Kohorte auf 500 Patienten erhöht, im Umfeld Lungenerkrankungen sogar auf fast 3,000 Patienten und Kontrollen. Bisher scheinen sich die Ergebnisse der kleinen Kohorten zu bestätigen, sodass die berechtigte Hoffnung besteht, dass entsprechende miRNA Signaturen ihren Weg in die klinische Praxis finden. Wichtig ist herauszustellen, dass für den klinischen Einsatz die Marker selbst entscheidend sind und nicht so sehr die Technologie mit der sie gemessen werden. Sowohl für Immunoassays als auch Mikroarrays, HTS und RT-qPCR gibt es bereits klinisch zugelassene Tests. Die Erkenntnisse, die aus diesen und anderen Studien gezogen wurden, sind von Fehlmann et al. kürzlich in einer umfassenden Arbeit zum humanen nichtkodierenden Transkriptom zusammengefasst worden [103].

4.3. Die Komplexität und wechselseitige Wirkung von miRNAs

Stetig wachsende Datensätze ermöglichen es, auch die Effekte von miRNAs auf ihre Zielgene besser abzuschätzen. Wie in der Einleitung und in Kapitel 2 beschrieben ist es eine der Hauptfunktionen von miRNAs die Expression von Genen zu unterdrücken oder entsprechend die mRNAs der Gene abzubauen. Eine Vielzahl von experimentell validierten oder vorhergesagten Zielgenen von miRNAs ist in Datenbanken wie der miRTarBase [191-194] oder der StarBase [195, 196] hinterlegt. Darüber hinaus sind miRNAs Teil eines komplexen Netzwerkes, zu dem auch andere Transkriptionsfaktoren beitragen. Bereits in den Arbeiten über Alzheimer (Kapitel 4.2) hat sich gezeigt, dass miRNAs gezielt Netzwerke modulieren und so wahrscheinlich einen substanziellen Einfluss auf die Entstehung oder das Voranschreiten der Erkrankung haben können.

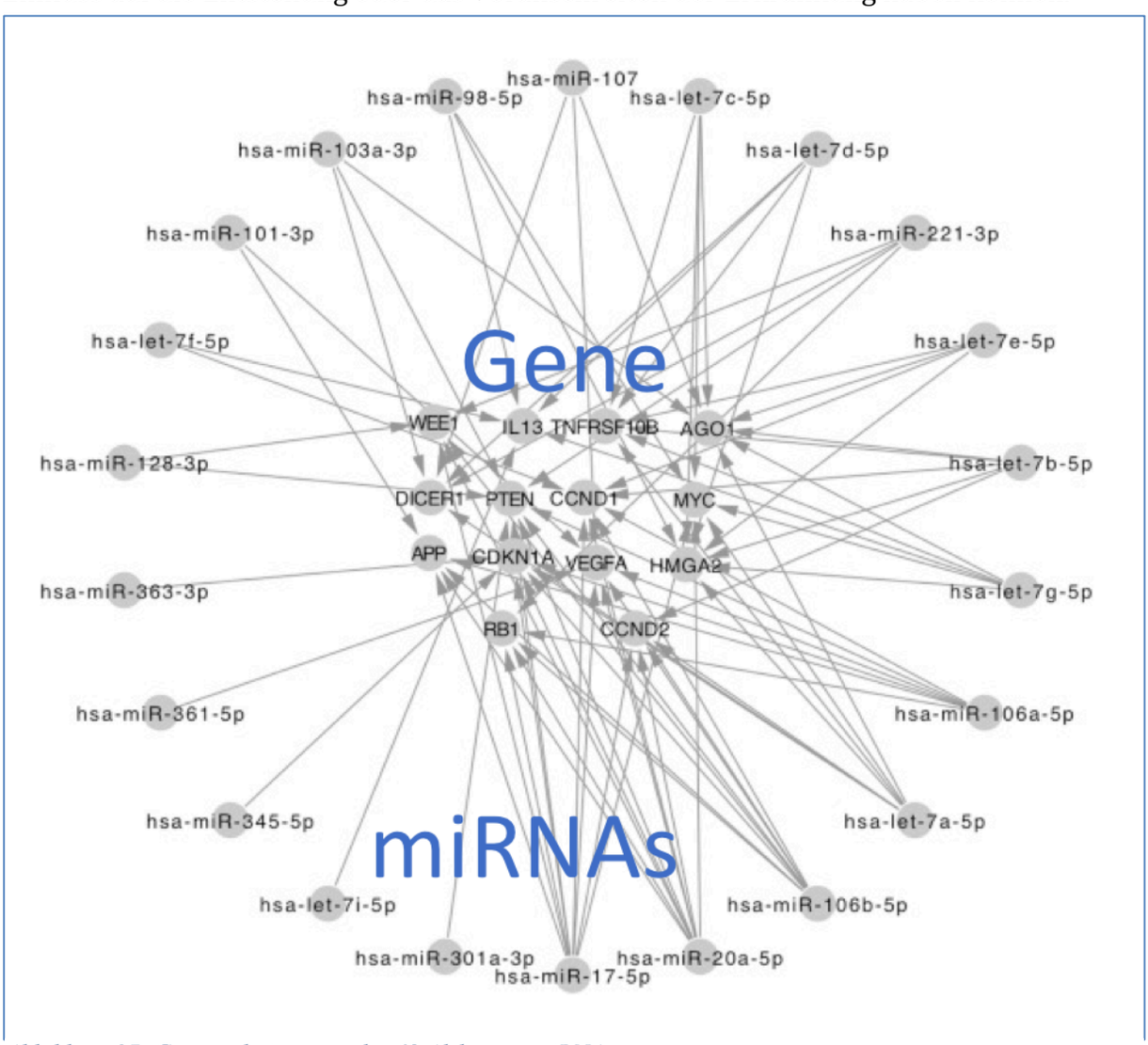


Abbildung 27: Genregulationsnetz der 68 Alzheimer miRNAs.

Das Netzwerk zeigt die signifikanten miRNAs aus beiden Alzheimer Studien die mindestens 5 Gene regulieren. Dieses Genset, welches in der Mitte dargestellt ist, beinhaltet viele Schlüsselgene für Alzheimer, zum Beispiel APP. Die Abbildung ist modifiziert aus Keller et al. entnommen.

In den Studien über Alzheimer wurden insgesamt 68 miRNAs als dysreguliert erkannt. Für 33 dieser miRNAs sind in der miRTarBase sowohl vorhergesagte als auch validierte Zielgene enthalten. Insgesamt wurden 563 Interaktionen zwischen den 33 miRNAs und 349 Genen detektiert. Das Kernnetzwerk, das aus den Genen besteht die von mindestens fünf miRNAs reguliert werden und den entsprechenden miRNAs ist in Abbildung 27 gezeigt. Viele der 14 Gene, die in Abbildung 27 enthalten sind, sind für ihre Bedeutung in Alzheimer bekannt. Das wahrscheinlich bekannteste davon ist APP, das für das Amyloid Precursor Protein codiert und eine Schlüsselrolle in der Entstehung von Neurodegeneration inne hat [197, 198].

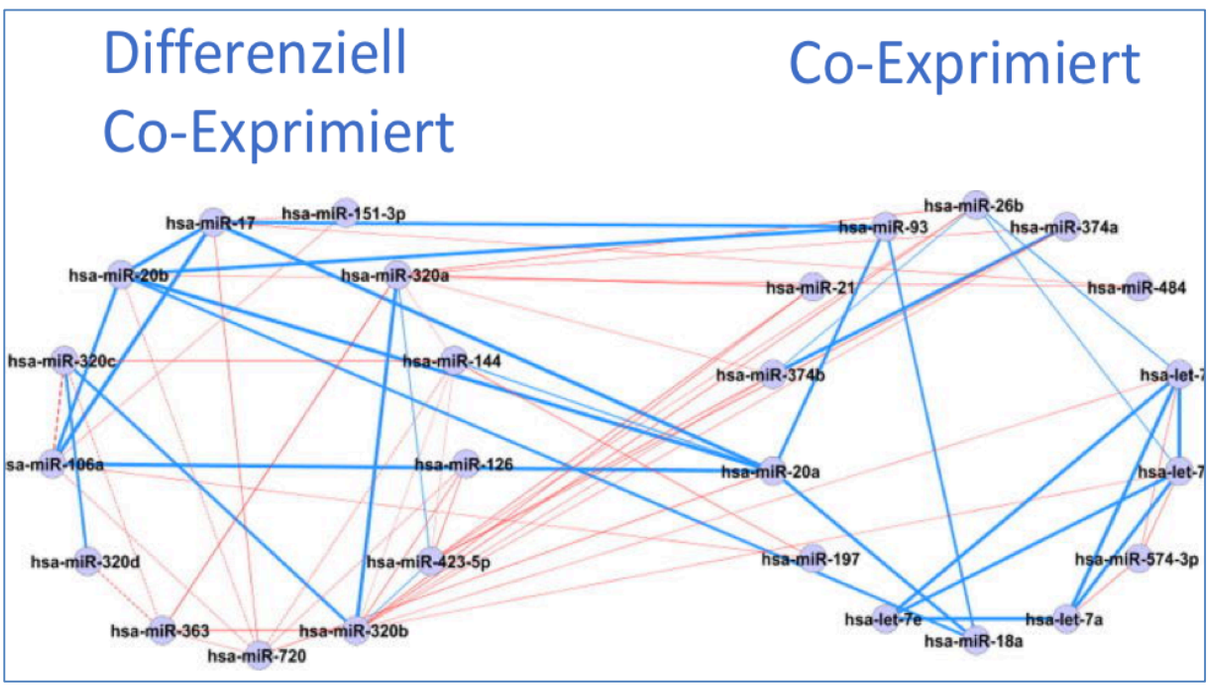


Abbildung 28: Kern-Co-Expressions Netzwerk.

Die Abbildung zeigt zwei Cluster von miRNAs. Die auf der linken Seite stehenden miRNAs sind differenziell co-exprimiert, die auf der rechten Seite stehenden miRNAs sind co-exprimiert. Da die Sequenzähnlichkeit einen Einfluss hat, ist die Kantendicke proportional zur Ähnlichkeit der Sequenzen gewählt. Blaue Kanten entsprechen positiver Korrelation, rote Kanten negativer. Gepunktete Kanten zeigen zudem die Paare an, die differenziell co-exprimiert sind. Die Abbildung ist modifiziert aus Stähler et al. entnommen.

Es ist wichtig zu erwähnen, dass die Interaktionen in Abbildung 27 aus der miRTarBase entnommen wurden und alle Paare von miRNAs und Zielgenen experimentell gefunden worden sind. Das ist in der Mehrzahl der Fälle aber wiederum nur eine indirekte Evidenz die auf Korrelationen beruht und aus Hochdurchsatzexperimenten abgeleitet wurde. Prinzipiell sollte für jede einzelne dieser Interaktionen ein spezifisches Experiment, wie zum Beispiel ein Luciferase Reporter Assay, durchgeführt werden [163]. Dies bedeutet jedoch einen enormen experimentellen Aufwand, der nicht im Rahmen einer theoretischen Arbeit möglich ist, sondern dedizierte Expertise im Labor erfordert.

Die Mechanismen der Regulation von Genen durch miRNAs scheinen insgesamt wesentlich komplexer als ursprünglich angenommen. Oft ist es nicht eine einzelne miRNA die ein isoliertes Gen reguliert. Vielmehr arbeiten miRNAs, teilweise sogar miRNA Familien kooperativ um komplette biochemische Pfade gezielt zu steuern [162, 163, 199-201]. Dabei hat sich auch die co-Expression von miRNAs entwickelt. Um zu verstehen, welche generelle Rolle miRNAs in Krankheiten einnehmen und wie sie gemeinsam agieren, habe ich mir gezielt miRNA co-Expression und differenzielle miRNA co-Expression angeschaut [59]. Basierend auf dem in Abschnitt 3.2.4. eingeführten Datensatz über mehrere hundert Patienten und 863 miRNAs wurden alle paarweisen Kombinationen von miRNAs untersucht (863*862/2 = 371.953 Kombinationen). Durch stringentes Filtern nach der absoluten Korrelation und dem p-Wert wurden 184 Paare von miRNAs, die entweder korreliert (118) oder antikorreliert waren (66), abgeleitet. Eine Detailanalyse hat dabei ergeben, dass diese Paare zum Teil differenziell co-exprimiert waren.

Konkret bedeutet das, dass die Korrelation entweder nur bei Kontrollprobanden oder nur bei Patienten vorhanden war. Das auffälligste Beispiel war das Paar hsa-miR-23a/hsa-miR-23b. Sowohl bei Tumorpatienten als auch bei Patienten die nicht an Krebserkrankungen litten, waren diese beiden miRNAs stark korreliert. Bei gesunden Probanden war die Korrelation hingegen fast nicht mehr sichtbar. Detaillierte Analysen haben ein Netzwerk ergeben, das aus zwei Komponenten besteht und in Abbildung 28 gezeigt ist. Auf der linken Seite des Netzwerkes in Abbildung 28 sind differenziell co-exprimierte miRNA Paare gezeigt, also solche die bei Erkrankungen ihre co-Expression gewinnen oder verlieren, während der rechte Teil die co-exprimierten miRNAs zeigt.

Auf der Suche nach Gründen für die co-Expression wurden verschiedene Hypothesen aufgestellt und getestet. Eine mögliche Erklärung ist Sequenzähnlichkeit. miRNAs mit ähnlicher Sequenz können beispielsweise evolutionär konserviert sein und die selbe Funktion ausüben. Tatsächlich hat die Sequenzähnlichkeit eine signifikante Rolle gespielt. Oft waren miRNAs mit ähnlicher Sequenz auch co-exprimiert. Allerdings gab es viele Fälle, bei denen sehr hohe co-Expression bestand ($p < 10^{-16}$) aber keinerlei Ähnlichkeit in der Sequenz festgestellt werden konnte. Ein weiterer Faktor, der signifikant dazu beigetragen hat, dass miRNAs co-exprimiert waren, war die chromosomale Lokalisation. Oft waren solche miRNA-Paare, die auf dem selben Chromosom teilweise direkt in miRNA Clustern co-lokalisiert waren, auch sehr stark miteinander korreliert. Insbesondere miRNAs der selben miRNA-Familien haben eine starke Tendenz zur co-Expression gezeigt. Eine Cluster Analyse hat ergeben, dass es co-exprimierte miRNA-Cluster gibt, deren Ursache weder auf Sequenzähnlichkeit noch auf Familienzugehörigkeit oder chromosomale Lokalisation beruhen.

Meine Ergebnisse deuten darauf hin, dass miRNAs generell ein kooperatives Verhalten zeigen, das weit über das bekannte Maß hinausgeht. Sie können teilweise ihre Funktion gegenseitig übernehmen und im Fall von Erkrankungen scheinen gezielt Teile des eher homöostatischen miRNA Regulationsnetzwerkes zusammenbrechen.

Publikationen: Die Arbeiten, die in diesem Abschnitt beschrieben werden basieren hauptsächlich auf [59].

4.4. Anwendungen in der „Synthetischen Biologie“

In meiner Doktorarbeit habe ich Plattformen und Assays entwickelt die zum Einsatz in der Molekulardiagnostik geeignet sind. In Kapitel 2.1 habe ich ausgeführt, dass die Technologien und Erkenntnisse aber auch darüber hinaus eingesetzt werden können. Eine Entwicklung, die in den vergangenen Jahren rasant Fahrt aufgenommen hat, ist die Synthetische Biologie. Im diesem Fachgebiet arbeiten Wissenschaftler interdisziplinär zusammen (Biologen, Chemiker, Ingenieure, Informatiker), um biologische Systeme zu erzeugen, die es so in der Natur nicht gibt. Angefangen mit neuen DANN-Oligonukleotid-ketten können so biologische Systeme mit neuen Eigenschaften erschaffen werden [202, 203]. Der Begriff der Synthetischen Biologie ist bereits seit mehreren Jahrzehnten geprägt und seit den 1980er Jahren werden entsprechende Systeme entwickelt [204-206]. Bereits seit fast einem Jahrzehnt ist es möglich, komplette Bakteriengenome zu synthetisieren und in lebende Bakterien einzubringen [207]. Die Forschung insgesamt hat sehr viele verschiedene Anwendungsfelder [208-217], von denen die meisten eines gemeinsam haben: sie benötigen synthetische Oligonukleotide als grundlegende Bausteine. Diese können erzeugt werden, indem gezielt in vorhandene DNA neue Mutationen eingebracht werden, vorhandene natürliche Code Stücke ohne Veränderung neu kombiniert werden oder man durch Syntheseverfahren beliebige Sequenzen herstellt.

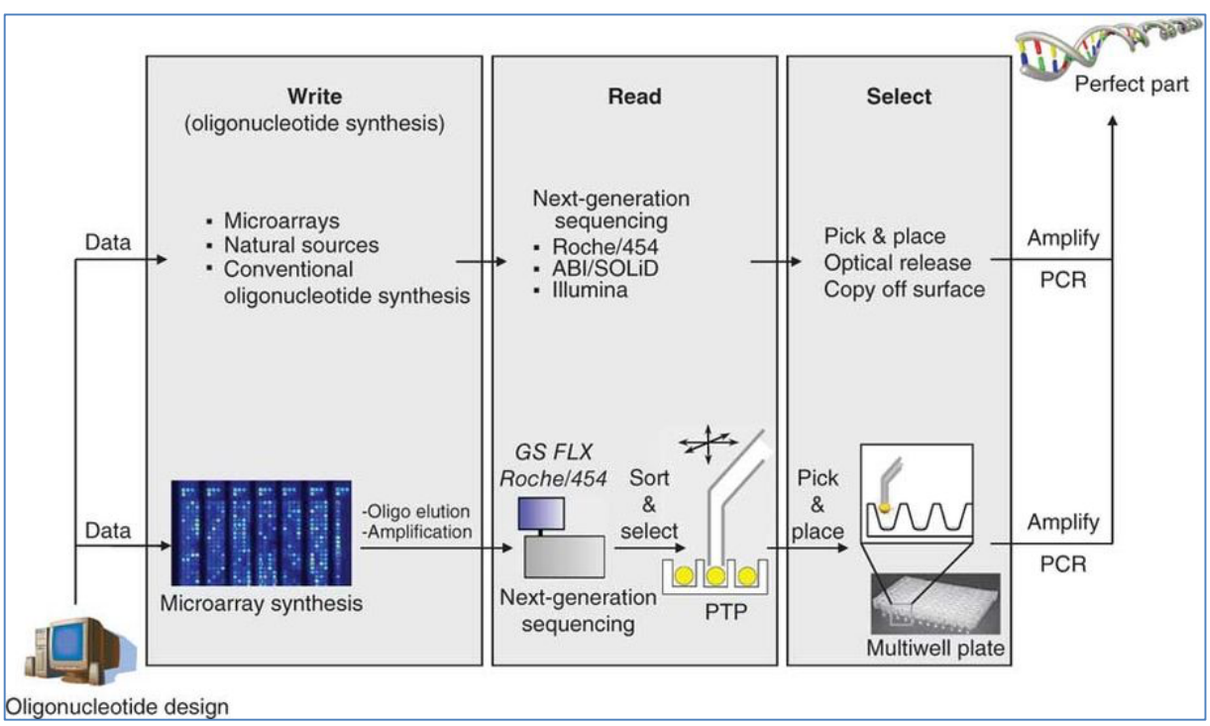


Abbildung 29: Grundkonzept des „Megakloners“.

Die Abbildung zeigt, wie wir die Geniom-Plattform verwendet haben um hoch präzise Oligonukleotide herzustellen. HTS wurde im Workflow eingesetzt, um die bereits zu größeren Stücken zusammengesetzten Fragmente Korrektur zu lesen. Die Abbildung entstammt aus Matzas et al.

Das in Kapitel 2.1. vorgestellte Geniom System besitzt genau diese Fähigkeit. In kurzer Zeit – innerhalb eines Arbeitstages – kann eine Anzahl von mehreren hunderttausend verschiedenen Oligonukleotiden der Länge von bis zu 50 Basen hergestellt werden. Diese können vom Glasträger gelöst und in der Synthetischen Biologie eingesetzt werden. Eine der Hauptherausforderungen ist dabei das möglichst fehlerfreie Herstellen von künstlicher DNA. Bereits das erste vollständige Bakteriengenom das künstlich erzeugt wurde hat aus mehr als einer Million Basen bestanden. Wenn man darüber nachdenkt höhere Organismen komplett oder zu sehr großen Teilen aus synthetischer DNA „herzustellen", müssen viele Millionen oder Milliarden Basen mit sehr geringer Fehlerrate erzeugt werden.

Die HTS Technologie bietet sich dabei an, um die erzeugten DNA Fragmente Korrektur zu lesen, bevor sie zu größeren DNA Stücken - wie zum Beispiel Genen - zusammengesetzt werden. Der entsprechende Ansatz wird Megacloning genannt [61] und ist in Abbildung 29 dargestellt.

Zunächst wird eine der oben genannten Quellen verwendet, um den benötigten grundlegenden Bausatz an DNA zu erhalten. Von jedem dieser grundlegenden Bausteine werden mehrere Instanzen erzeugt. Jede mögliche Variante der DNA-Bausteine kann Fehler enthalten. Daher werden sie mit einer sehr akkuraten Technologie, im vorliegenden Fall mit der 454 Sequenzier-Technologie der Firma Roche, sequenziert. Der Träger, der verwendet wurde um die Sequenzierung der DNA-Klone durchzuführen, wird im Anschluss an die Sequenzierung in den eigentlichen Megacloner gegeben. Ein Computerprogramm extrahiert die Positionen der korrekt gelesenen Reads auf dem Sequenzierträger. Ein Roboter steuert gezielt die Position mit dem korrekt gelesenen Fragment an und extrahiert den DNA-Klon vom Objektträger. Dadurch können gezielt die richtigen Fragmente ausgewählt werden. Diese werden anschließend Stück für Stück zu längeren Abschnitten zusammengesetzt.

Um die hohe Genauigkeit des Megacloners zu demonstrieren wurden 3.918 verschiedene Sequenzen auf einem Geniom-Mikroarray hergestellt. Daraus wurden 319 DNA-Klone die eine 100 %-ige Übereinstimmung zu den gewünschten Fragmenten zeigten mit dem Megacloner vollautomatisch ausgewählt. Während von den ursprünglich ausgewählten Sequenzen nur 3.1 % absolut korrekt waren, zeigten die vom Megacloner vorselektierten Fragmente eine 27,2-fach höhere Genauigkeit. Die Fehlerverteilung des nicht korrekturgelesenen Pools und des Pools vom Megacloner sind im Vergleich zueinander in Abbildung 30 dargestellt. Um zu demonstrieren, dass die entsprechenden Fragmente auch zusammengesetzt werden können, wurden jeweils neun und zehn DNA-Stücke nach dem Megacloning zu zwei Genen ligiert. Die Gene wurden dann durch Sanger-

Sequenzierung überprüft und es hat sich eine Erfolgsrate von 87,5 % gezeigt.

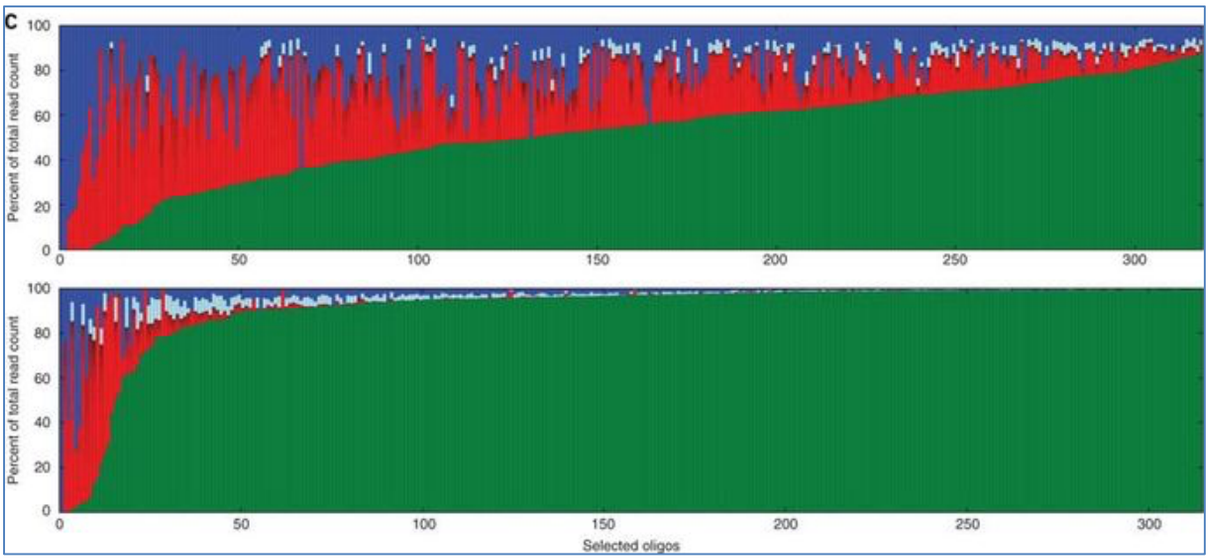


Abbildung 30: Performance des „Megakloners".

Der obere Teil der Abbildung zeigt für ausgewählte Reads den korrekten Anteil in grün und Fehler in rot. Im unteren Teil wird das Ergebnis nach dem Einsatz des „Megacloners" gezeigt. Die Abbildung entstammt aus Matzas et al.

Um die Leistungsfähigkeit des Megacloners noch besser abzuschätzen wurden längere Fragmente basierend auf fast 400 Basen langen Oligonukleotiden zusammengesetzt. Aus 29 korrekt vorselektierten DNA-Klonen konnte ein 7,195 Basen langes DNA-Fragment ohne Fehler erzeugt werden. Da kein Fehler in dem Fragment gefunden wurde, wurde ein statistisches Modell gebildet, das die verschiedenen Fehlerraten kombiniert. Dadurch wurde gezeigt, dass der Megacloner eine Genauigkeit von 5 Fehlern auf 100,000 Basen erreicht. Im Vergleich zu der ursprünglich nicht korrekturgelesenen Fraktion ist die Genauigkeit um einen Faktor von 500-mal verbessert worden. Durch den Megacloner können außerdem die Kosten der Gensynthese um einen Faktor von 10 verringert werden.

Gerade in immer komplexer werdenden Synthese Projekten bis hin zur synthetischen Herstellung von größeren Teilen höherer Organismen ist die Megacloner-Technologie ein wichtiger Baustein, um die benötigte hohe Qualität in der Synthetischen Biologie zu erlangen. Die Methode zum Herstellen von Oligonukleotiden und zum gezielten Korrekturlesen und Extrahieren der richtigen DNA-Fragmente wurde inzwischen vom Genomik-Pionier Graig Venter und seiner Firma Synthetic Genomics Incorporated gekauft und werden dort eingesetzt, um die DNA-Synthese substanziell zu verbessern.

Publikationen: Die Arbeiten, die in diesem Abschnitt beschrieben werden basieren hauptsächlich auf [61]. Im Bezug zum Megacloner wurden außerdem vier Patente angemeldet (US2017267999, US2010256012, EP2109499, DE102007018833).

Kapitel 5

Momentane Arbeit und Ausblick

Meine wissenschaftliche Tätigkeit habe ich als Ingenieur begonnen und Plattformen für den Einsatz in der Molekulardiagnostik entwickelt. Später habe ich mich mehr in biologische Aspekte eingearbeitet und Assays konzipiert und Biomarker erforscht. Dabei haben miRNAs eine essenzielle Rolle gespielt. Meine Arbeit ist dabei immer theoretischer und computerlastiger geworden und ich bin zu den Grundlagen meines Studdiums zurückgekehrt. Während ich am Anfang Auswertungen in Excel auf einem PC durchführen konnte, bedarf es heute spezieller Software wie ***R*** oder höheren und effizienteren Programmiersprachen wie C++ und großen Rechenclustern, um die Daten in alltäglichen Projekten zu verarbeiten. Neben dem klassischen maschinellen Lernen werden Deep Learning Aspekte und Künstliche Intelligenz quasi täglich wichtiger. Diese Entwicklung wird meine zukünftige Tätigkeit weiter mitbestimmen.

Die grundlegende Erforschung von miRNAs als Biomarker betrachte ich als wissenschaftlich weitestgehend abgeschlossen. Die nächsten Schritte bestehen hier im Messen größerer Kohorten und in der experimentellen Aufklärung der biologischen Wirkungsweise der miRNAs in Erkrankungen. Den ersten Teil der Arbeit, die klinische Validierung, gehen der Lehrstuhl für Klinische Bioinformatik und die Arbeitsgruppe für Humangenetik gemeinsam mit der Firma Hummingbird Diagnostic GmbH in Heidelberg an. Den zweiten Teil, die Erforschung biologischer Mechanismen, bearbeitet maßgeblich die Arbeitsgruppe Humangenetik.

Ich bin während meiner Doktorarbeit von meiner Position als CTO von Siemens Healthcare zu der Pharma Firma Merck KGaA in Darmstadt gewechselt. Dort leite ich das globale Medical Device & Service Geschäft. Dennoch spielt Forschung in meinem Alltag eine wichtige Rolle. Zwei natürliche Entwicklungen, die sich in meiner vorliegenden Ausarbeitung erkennen lassen, werden dabei weiter fortgeführt: Die Verlagerung der Medizin hin zum Patienten und die Digitalisierung im Gesundheitswesen. Der wichtigste Anwendungsfall ist dabei für mich nach wie vor die Multiple Sklerose.

Mit meinem Team bei Merck entwickele ich Software, die es Patienten erlaubt ihre medizinischen Daten – soweit möglich und sinnvoll – selbst zu verwalten und Ärzten gezielt mit geringem Aufwand Zugriff darauf zu geben. Dazu entwickeln wir bei Merck ein gesamtes IT Ökosystem, Software für Patienten und Ärzte, die sowohl auf mobilen Endgeräten wie auch auf klassischen PCs eingesetzt werden kann und die es erlaubt chronische Erkrankungen besser zu kontrollieren. Wir planen außerdem molekulare

Tests, wie die in dieser Arbeit beschriebenen miRNA-Signaturen für Multiple Sklerose, zu verwenden, um den Patienten gezielter die richtige Behandlung zum richtigen Zeitpunkt zukommen zu lassen. Dabei wird eine weitere Entwicklung in der Doktorarbeit fortgesetzt: Unsere Biomarker waren zunächst für die Anwendung in Speziallaboren gedacht, später für die Anwendung in Zentrallaboren von Krankenhäusern oder sogar Point-of-Care in den entsprechenden Fachabteilungen des Krankenhauses. Momentan etablieren wir diese Bluttests in einer Art und Weise, dass Patienten sich die Probe selbst zu Hause entnehmen können. Ein Stich mit einer kleinen Lanzette und ein Tropfen Blut aus dem Finger sind dazu ausreichend.

Durch unsere Lösungen zielen wir darauf ab, Ärzte in der Behandlung von MS Patienten besser zu unterstützen. Diese Forschung wollen wir gemeinsam mit dem Uniklinikum des Saarlandes und der Klinischen Bioinformatik an einem großen Patientenkollektiv im Saarland testen. Im Falle eines Erfolges kann die Lösung wegen ihrer Modularität und Flexibilität dann auch ohne Weiteres auf andere Anwendungsfelder, wie zum Beispiel Rheumatoide Arthritis oder neurodegenerative Erkrankungen, wie Alzheimer oder Parkinson, übertragen werden.

Anhang 1

Literaturverzeichnis

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
2. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR *et al*: **Accurate whole human genome sequencing using reversible terminator chemistry**. *Nature* 2008, **456**(7218):53-59.
3. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C: **Ten years of next-generation sequencing technology**. *Trends Genet* 2014, **30**(9):418-426.
4. Zhang J, Chiodini R, Badr A, Zhang G: **The impact of next-generation sequencing on genomics**. *J Genet Genomics* 2011, **38**(3):95-109.
5. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER: **The next-generation sequencing revolution and its impact on genomics**. *Cell* 2013, **155**(1):27-38.
6. Levy SE, Myers RM: **Advancements in Next-Generation Sequencing**. *Annu Rev Genomics Hum Genet* 2016, **17**:95-115.
7. Bumgarner R: **Overview of DNA microarrays: types, applications, and their future**. *Curr Protoc Mol Biol* 2013, **Chapter 22**:Unit 22 21.
8. Goyal R, Goyal D, Longo LD, Clyman RI: **Microarray gene expression analysis in ovine ductus arteriosus during fetal development and birth transition**. *Pediatr Res* 2016, **80**(4):610-618.
9. Belder N, Coskun O, Erdogan BD, Savas B, Ensari A, Ozdag H: **Optimization of gene expression microarray protocol for formalin-fixed paraffin-embedded tissues**. *Genom Data* 2016, **7**:303-306.
10. Di Salle P, Incerti G, Colantuono C, Chiusano ML: **Gene co-expression analyses: an overview from microarray collections in Arabidopsis thaliana**. *Brief Bioinform* 2017, **18**(2):215-225.
11. Kung JT, Colognori D, Lee JT: **Long noncoding RNAs: past, present, and future**. *Genetics* 2013, **193**(3):651-669.
12. Hombach S, Kretz M: **Non-coding RNAs: Classification, Biology and Functioning**. *Adv Exp Med Biol* 2016, **937**:3-17.
13. Gomes AQ, Nolasco S, Soares H: **Non-coding RNAs: multi-tasking molecules in the cell**. *Int J Mol Sci* 2013, **14**(8):16010-16039.
14. Sasaki H, Allen ND, Surani MA: **DNA methylation and genomic imprinting in mammals**. *EXS* 1993, **64**:469-486.
15. Li E, Zhang Y: **DNA methylation in mammals**. *Cold Spring Harb Perspect Biol* 2014, **6**(5):a019133.
16. Bannister AJ, Kouzarides T: **Regulation of chromatin by histone modifications**. *Cell Res* 2011, **21**(3):381-395.
17. Urban PL: **Quantitative mass spectrometry: an overview**. *Philos Trans A Math Phys Eng Sci* 2016, **374**(2079).
18. Nikolov M, Schmidt C, Urlaub H: **Quantitative mass spectrometry-based proteomics: an overview**. *Methods Mol Biol* 2012, **893**:85-100.

19. Hasin Y, Seldin M, Lusis A: **Multi-omics approaches to disease**. *Genome Biol* 2017, **18**(1):83.
20. Yan J, Risacher SL, Shen L, Saykin AJ: **Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data**. *Brief Bioinform* 2017.
21. Huang S, Chaudhary K, Garmire LX: **More Is Better: Recent Progress in Multi-Omics Data Integration Methods**. *Front Genet* 2017, **8**:84.
22. Bock C, Farlik M, Sheffield NC: **Multi-Omics of Single Cells: Strategies and Applications**. *Trends Biotechnol* 2016, **34**(8):605-608.
23. Palsson B, Zengler K: **The challenges of integrating multi-omic data sets**. *Nat Chem Biol* 2010, **6**(11):787-789.
24. Hu W, Yang Y, Li X, Huang M, Xu F, Ge W, Zhang S, Zheng S: **Multi-omics Approach Reveals Distinct Differences in Left- and Right-sided Colon Cancer**. *Mol Cancer Res* 2017.
25. Ma S, Ren J, Fenyo D: **Breast Cancer Prognostics Using Multi-Omics Data**. *AMIA Jt Summits Transl Sci Proc* 2016, **2016**:52-59.
26. Lin S, Yin YA, Jiang X, Sahni N, Yi S: **Multi-OMICs and Genome Editing Perspectives on Liver Cancer Signaling Networks**. *Biomed Res Int* 2016, **2016**:6186281.
27. Tenzer S, Leidinger P, Backes C, Huwer H, Hildebrandt A, Lenhof HP, Wesse T, Franke A, Meese E, Keller A: **Integrated quantitative proteomic and transcriptomic analysis of lung tumor and control tissue: a lung cancer showcase**. *Oncotarget* 2016, **7**(12):14857-14870.
28. Haas J, Mester S, Lai A, Frese KS, Sedaghat-Hamedani F, Kayvanpour E, Rausch T, Nietsch R, Boeckel JN, Carstensen A *et al*: **Genomic structural variations lead to dysregulation of important coding and non-coding RNA species in dilated cardiomyopathy**. *EMBO Mol Med* 2018, **10**(1):107-120.
29. Pimplikar SW: **Multi-omics and Alzheimer's disease: a slower but surer path to an efficacious therapy?** *Am J Physiol Cell Physiol* 2017, **313**(1):C1-C2.
30. Wood AJ, Lo TW, Zeitler B, Pickle CS, Ralston EJ, Lee AH, Amora R, Miller JC, Leung E, Meng X *et al*: **Targeted genome editing across species using ZFNs and TALENs**. *Science* 2011, **333**(6040):307.
31. Ran FA, Hsu PD, Lin CY, Gootenberg JS, Konermann S, Trevino AE, Scott DA, Inoue A, Matoba S, Zhang Y *et al*: **Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity**. *Cell* 2013, **154**(6):1380-1389.
32. Jung M, Schaefer A, Steiner I, Kempkensteffen C, Stephan C, Erbersdobler A, Jung K: **Robust microRNA stability in degraded RNA preparations from human tissue and cell samples**. *Clin Chem* 2010, **56**(6):998-1006.
33. Sun W, Julie Li YS, Huang HD, Shyy JY, Chien S: **microRNA: a master regulator of cellular processes for bioengineering systems**. *Annu Rev Biomed Eng* 2010, **12**:1-27.
34. Keller A, Leidinger P, Bauer A, Elsharawy A, Haas J, Backes C, Wendschlag A, Giese N, Tjaden C, Ott K *et al*: **Toward the blood-borne miRNome of human diseases**. *Nat Methods* 2011, **8**(10):841-843.
35. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14**. *Cell* 1993, **75**(5):843-854.
36. Ruvkun G: **Molecular biology. Glimpses of a tiny RNA world**. *Science* 2001, **294**(5543):797-799.
37. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T: **Identification of novel genes coding for small expressed RNAs**. *Science* 2001, **294**(5543):853-858.
38. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function**. *Cell* 2004, **116**(2):281-297.

39. Guimil R, Beier M, Scheffler M, Rebscher H, Funk J, Wixmerten A, Baum M, Hermann C, Tahedl H, Moschel E *et al*: **Geniom technology--the benchtop array facility**. *Nucleosides Nucleotides Nucleic Acids* 2003, **22**(5-8):1721-1723.
40. Summerer D, Hevroni D, Jain A, Oldenburger O, Parker J, Caruso A, Stahler CF, Stahler PF, Beier M: **A flexible and fully integrated system for amplification, detection and genotyping of genomic DNA targets based on microfluidic oligonucleotide arrays**. *N Biotechnol* 2010, **27**(2):149-155.
41. Baum M, Bielau S, Rittner N, Schmid K, Eggelbusch K, Dahms M, Schlauersbach A, Tahedl H, Beier M, Guimil R *et al*: **Validation of a novel, fully integrated and flexible microarray benchtop facility for gene expression profiling**. *Nucleic Acids Res* 2003, **31**(23):e151.
42. Vorwerk S, Ganter K, Cheng Y, Hoheisel J, Stahler PF, Beier M: **Microfluidic-based enzymatic on-chip labeling of miRNAs**. *N Biotechnol* 2008, **25**(2-3):142-149.
43. Kappel A, Backes C, Huang Y, Zafari S, Leidinger P, Meder B, Schwarz H, Gumbrecht W, Meese E, Staehler CF *et al*: **MicroRNA in vitro diagnostics using immunoassay analyzers**. *Clin Chem* 2015, **61**(4):600-607.
44. Hofmann S, Huang Y, Paulicka P, Kappel A, Katus HA, Keller A, Meder B, Stahler CF, Gumbrecht W: **Double-Stranded Ligation Assay for the Rapid Multiplex Quantification of MicroRNAs**. *Anal Chem* 2015, **87**(24):12104-12111.
45. Fehlmann T, Reinheimer S, Geng C, Su X, Drmanac S, Alexeev A, Zhang C, Backes C, Ludwig N, Hart M *et al*: **cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs**. *Clin Epigenetics* 2016, **8**:123.
46. Meder B, Haas J, Keller A, Heid C, Just S, Borries A, Boisguerin V, Scharfenberger-Schmeer M, Stahler P, Beier M *et al*: **Targeted next-generation sequencing for the molecular genetic diagnostics of cardiomyopathies**. *Circ Cardiovasc Genet* 2011, **4**(2):110-122.
47. Summerer D, Schracke N, Wu H, Cheng Y, Bau S, Stahler CF, Stahler PF, Beier M: **Targeted high throughput sequencing of a cancer-related exome subset by specific sequence capture with a fully automated microarray platform**. *Genomics* 2010, **95**(4):241-246.
48. Summerer D, Wu H, Haase B, Cheng Y, Schracke N, Stahler CF, Chee MS, Stahler PF, Beier M: **Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing**. *Genome Res* 2009, **19**(9):1616-1621.
49. Elsharawy A, Forster M, Schracke N, Keller A, Thomsen I, Petersen BS, Stade B, Stahler P, Schreiber S, Rosenstiel P *et al*: **Improving mapping and SNP-calling performance in multiplexed targeted next-generation sequencing**. *BMC Genomics* 2012, **13**:417.
50. Backes C, Leidinger P, Altmann G, Wuerstle M, Meder B, Galata V, Mueller SC, Sickert D, Stahler C, Meese E *et al*: **Influence of next-generation sequencing and storage conditions on miRNA patterns generated from PAXgene blood**. *Anal Chem* 2015, **87**(17):8910-8916.
51. Meder B, Backes C, Haas J, Leidinger P, Stahler C, Grossmann T, Vogel B, Frese K, Giannitsis E, Katus HA *et al*: **Influence of the confounding factors age and sex on microRNA profiles from peripheral blood**. *Clin Chem* 2014, **60**(9):1200-1208.
52. Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T, Pallasch C, Rheinheimer S, Meder B, Stahler C, Meese E *et al*: **Distribution of miRNA expression across human tissues**. *Nucleic Acids Res* 2016, **44**(8):3865-3877.
53. Keller A, Leidinger P, Gislefoss R, Haugen A, Langseth H, Staehler P, Lenhof HP, Meese E: **Stable serum miRNA profiles as potential tool for non-invasive lung cancer diagnosis**. *RNA Biol* 2011, **8**(3):506-516.
54. Keller A, Backes C, Leidinger P, Kefer N, Boisguerin V, Barbacioru C, Vogel B,

Matzas M, Huwer H, Katus HA *et al*: **Next-generation sequencing identifies novel microRNAs in peripheral blood of lung cancer patients**. *Mol Biosyst* 2011, **7**(12):3187-3199.
55. Leidinger P, Galata V, Backes C, Stahler C, Rheinheimer S, Huwer H, Meese E, Keller A: **Longitudinal study on circulating miRNAs in patients after lung cancer resection**. *Oncotarget* 2015, **6**(18):16674-16685.
56. Keller A, Backes C, Haas J, Leidinger P, Maetzler W, Deuschle C, Berg D, Ruschil C, Galata V, Ruprecht K *et al*: **Validating Alzheimer's disease micro RNAs using next-generation sequencing**. *Alzheimers Dement* 2016, **12**(5):565-576.
57. Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, Haas J, Ruprecht K, Paul F, Stahler C *et al*: **A blood based 12-miRNA signature of Alzheimer disease patients**. *Genome Biol* 2013, **14**(7):R78.
58. Keller A, Leidinger P, Steinmeyer F, Stahler C, Franke A, Hemmrich-Stanisak G, Kappel A, Wright I, Dorr J, Paul F *et al*: **Comprehensive analysis of microRNA profiles in multiple sclerosis including next-generation sequencing**. *Mult Scler* 2014, **20**(3):295-303.
59. Staehler CF, Keller A, Leidinger P, Backes C, Chandran A, Wischhusen J, Meder B, Meese E: **Whole miRNome-wide differential co-expression of microRNAs**. *Genomics Proteomics Bioinformatics* 2012, **10**(5):285-294.
60. Laczny C, Leidinger P, Haas J, Ludwig N, Backes C, Gerasch A, Kaufmann M, Vogel B, Katus HA, Meder B *et al*: **miRTrail--a comprehensive webserver for analyzing gene and miRNA patterns to enhance the understanding of regulatory mechanisms in diseases**. *BMC Bioinformatics* 2012, **13**:36.
61. Matzas M, Stahler PF, Kefer N, Siebelt N, Boisguerin V, Leonard JT, Keller A, Stahler CF, Haberle P, Gharizadeh B *et al*: **High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing**. *Nat Biotechnol* 2010, **28**(12):1291-1294.
62. Stahler P, Beier M, Gao X, Hoheisel JD: **Another side of genomics: synthetic biology as a means for the exploitation of whole-genome sequence information**. *J Biotechnol* 2006, **124**(1):206-212.
63. Ohno S: **So much "junk" DNA in our genome**. *Brookhaven Symp Biol* 1972, **23**:366-370.
64. Khajavinia A, Makalowski W: **What is "junk" DNA, and what is it worth?** *Sci Am* 2007, **296**(5):104.
65. Zuckerkandl E: **Revisiting junk DNA**. *J Mol Evol* 1992, **34**(3):259-271.
66. Flam F: **Hints of a language in junk DNA**. *Science* 1994, **266**(5189):1320.
67. Nowak R: **Mining treasures from 'junk DNA'**. *Science* 1994, **263**(5147):608-610.
68. Zuckerkandl E: **Junk DNA and sectorial gene repression**. *Gene* 1997, **205**(1-2):323-343.
69. Wong GK, Passey DA, Huang Y, Yang Z, Yu J: **Is "junk" DNA mostly intron DNA?** *Genome Res* 2000, **10**(11):1672-1678.
70. Meagher TR, Costich DE: **'Junk' DNA and long-term phenotypic evolution in Silene section Elisanthe (Caryophyllaceae)**. *Proc Biol Sci* 2004, **271 Suppl 6**:S493-497.
71. Slack FJ: **Regulatory RNAs and the demise of 'junk' DNA**. *Genome Biol* 2006, **7**(9):328.
72. Pennisi E: **Genomics. ENCODE project writes eulogy for junk DNA**. *Science* 2012, **337**(6099):1159, 1161.
73. Doolittle WF: **Is junk DNA bunk? A critique of ENCODE**. *Proc Natl Acad Sci U S A* 2013, **110**(14):5294-5300.
74. He L, Sedwick C: **Lin He: "Junk" DNA isn't**. *J Cell Biol* 2015, **211**(1):4-5.
75. Li SC, Chan WC, Hu LY, Lai CH, Hsu CN, Lin WC: **Identification of homologous**

microRNAs in 56 animal genomes. *Genomics* 2010, **96**(1):1-9.
76. Vejnar CE, Zdobnov EM: **MiRmap: comprehensive prediction of microRNA target repression strength**. *Nucleic Acids Res* 2012, **40**(22):11673-11683.
77. Griffiths-Jones S: **The microRNA Registry**. *Nucleic Acids Res* 2004, **32**(Database issue):D109-111.
78. Griffiths-Jones S: **miRBase: the microRNA sequence database**. *Methods Mol Biol* 2006, **342**:129-138.
79. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature**. *Nucleic Acids Res* 2006, **34**(Database issue):D140-144.
80. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics**. *Nucleic Acids Res* 2008, **36**(Database issue):D154-158.
81. Griffiths-Jones S: **miRBase: microRNA sequences and annotation**. *Curr Protoc Bioinformatics* 2010, **Chapter 12**:Unit 12 19 11-10.
82. Kozomara A, Griffiths-Jones S: **miRBase: integrating microRNA annotation and deep-sequencing data**. *Nucleic Acids Res* 2011, **39**(Database issue):D152-157.
83. Kozomara A, Griffiths-Jones S: **miRBase: annotating high confidence microRNAs using deep sequencing data**. *Nucleic Acids Res* 2014, **42**(Database issue):D68-73.
84. Londin E, Loher P, Telonis AG, Quann K, Clark P, Jing Y, Hatzimichael E, Kirino Y, Honda S, Lally M *et al*: **Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs**. *Proc Natl Acad Sci U S A* 2015, **112**(10):E1106-1115.
85. Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N: **Discovering microRNAs from deep sequencing data using miRDeep**. *Nat Biotechnol* 2008, **26**(4):407-415.
86. Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N: **miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades**. *Nucleic Acids Res* 2012, **40**(1):37-52.
87. Backes C, Meder B, Hart M, Ludwig N, Leidinger P, Vogel B, Galata V, Roth P, Menegatti J, Grasser F *et al*: **Prioritizing and selecting likely novel miRNAs from NGS data**. *Nucleic Acids Res* 2016, **44**(6):e53.
88. Ludwig N, Becker M, Schumann T, Speer T, Fehlmann T, Keller A, Meese E: **Bias in recent miRBase annotations potentially associated with RNA quality issues**. *Sci Rep* 2017, **7**(1):5162.
89. Fromm B, Billipp T, Peck LE, Johansen M, Tarver JE, King BL, Newcomb JM, Sempere LF, Flatmark K, Hovig E *et al*: **A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome**. *Annu Rev Genet* 2015, **49**:213-242.
90. Backes C, Fehlmann T, Kern F, Kehl T, Lenhof HP, Meese E, Keller A: **miRCarta: a central repository for collecting miRNA candidates**. *Nucleic Acids Res* 2017.
91. Haseeb A, Makki MS, Khan NM, Ahmad I, Haqqi TM: **Deep sequencing and analyses of miRNAs, isomiRs and miRNA induced silencing complex (miRISC)-associated miRNome in primary human chondrocytes**. *Sci Rep* 2017, **7**(1):15178.
92. Wallaert A, Van Loocke W, Hernandez L, Taghon T, Speleman F, Van Vlierberghe P: **Comprehensive miRNA expression profiling in human T-cell acute lymphoblastic leukemia by small RNA-sequencing**. *Sci Rep* 2017, **7**(1):7901.
93. Gong J, Wu Y, Zhang X, Liao Y, Sibanda VL, Liu W, Guo AY: **Comprehensive analysis of human small RNA sequencing data provides insights into expression profiles and miRNA editing**. *RNA Biol* 2014, **11**(11):1375-1385.
94. Ji H, Chen M, Greening DW, He W, Rai A, Zhang W, Simpson RJ: **Deep sequencing of RNA from three different extracellular vesicle (EV) subtypes released from the**

human LIM1863 colon cancer cell line uncovers distinct miRNA-enrichment signatures. *PLoS One* 2014, **9**(10):e110314.
95. Moreira FC, Assumpcao M, Hamoy IG, Darnet S, Burbano R, Khayat A, Goncalves AN, Alencar DO, Cruz A, Magalhaes L *et al*: **MiRNA expression profile for the human gastric antrum region using ultra-deep sequencing**. *PLoS One* 2014, **9**(3):e92300.
96. Cheng L, Sun X, Scicluna BJ, Coleman BM, Hill AF: **Characterization and deep sequencing analysis of exosomal and non-exosomal miRNA in human urine**. *Kidney Int* 2014, **86**(2):433-444.
97. Burgos KL, Javaherian A, Bomprezzi R, Ghaffari L, Rhodes S, Courtright A, Tembe W, Kim S, Metpally R, Van Keuren-Jensen K: **Identification of extracellular miRNA in human cerebrospinal fluid by next-generation sequencing**. *RNA* 2013, **19**(5):712-722.
98. Munch EM, Harris RA, Mohammad M, Benham AL, Pejerrey SM, Showalter L, Hu M, Shope CD, Maningat PD, Gunaratne PH *et al*: **Transcriptome profiling of microRNA by Next-Gen deep sequencing reveals known and novel miRNA species in the lipid fraction of human breast milk**. *PLoS One* 2013, **8**(2):e50564.
99. Moore LM, Kivinen V, Liu Y, Annala M, Cogdell D, Liu X, Liu CG, Sawaya R, Yli-Harja O, Shmulevich I *et al*: **Transcriptome and small RNA deep sequencing reveals deregulation of miRNA biogenesis in human glioma**. *J Pathol* 2013, **229**(3):449-459.
100. Sripada L, Tomar D, Prajapati P, Singh R, Singh AK, Singh R: **Systematic analysis of small RNAs associated with human mitochondria by deep sequencing: detailed analysis of mitochondrial associated miRNA**. *PLoS One* 2012, **7**(9):e44873.
101. Burroughs AM, Ando Y, de Hoon MJ, Tomaru Y, Suzuki H, Hayashizaki Y, Daub CO: **Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin**. *RNA Biol* 2011, **8**(1):158-177.
102. Fehlmann T, Backes C, Kahraman M, Haas J, Ludwig N, Posch AE, Wurstle ML, Hubenthal M, Franke A, Meder B *et al*: **Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs**. *Nucleic Acids Res* 2017, **45**(15):8731-8744.
103. Fehlmann T, Backes C, Alles J, Fischer U, Hart M, Kern F, Langseth H, Rounge T, Umu SU, Kahraman M *et al*: **A high-resolution map of the human small non-coding transcriptome**. *Bioinformatics* 2017.
104. Fehlmann T, Ludwig N, Backes C, Meese E, Keller A: **Distribution of microRNA biomarker candidates in solid tissues and body fluids**. *RNA Biol* 2016, **13**(11):1084-1088.
105. Chang TW: **Binding of cells to matrixes of distinct antibodies coated on solid surface**. *J Immunol Methods* 1983, **65**(1-2):217-223.
106. Barbulovic-Nad I, Lucente M, Sun Y, Zhang M, Wheeler AR, Bussmann M: **Bio-microarray fabrication techniques--a review**. *Crit Rev Biotechnol* 2006, **26**(4):237-259.
107. Wang J: **Computational biology of genome expression and regulation--a review of microarray bioinformatics**. *J Environ Pathol Toxicol Oncol* 2008, **27**(3):157-179.
108. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M: **Comparison of next-generation sequencing systems**. *J Biomed Biotechnol* 2012, **2012**:251364.
109. Shao J, Yu M, Jiang L, Wu F, Liu X: **Sequencing and bioinformatics analysis of the differentially expressed genes in herniated discs with or without calcification**. *Int J Mol Med* 2017, **39**(1):81-90.
110. Gruning BA, Fallmann J, Yusuf D, Will S, Erxleben A, Eggenhofer F, Houwaart T, Batut B, Videm P, Bagnacani A *et al*: **The RNA workbench: best practices for RNA**

and high-throughput sequencing bioinformatics in Galaxy. *Nucleic Acids Res* 2017, **45**(W1):W560-W566.
111. Djebali S, Wucher V, Foissac S, Hitte C, Corre E, Derrien T: **Bioinformatics Pipeline for Transcriptome Sequencing Analysis**. *Methods Mol Biol* 2017, **1468**:201-219.
112. Anslan S, Bahram M, Hiiesalu I, Tedersoo L: **PipeCraft: Flexible open-source toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data**. *Mol Ecol Resour* 2017, **17**(6):e234-e240.
113. Tsai EA, Shakbatyan R, Evans J, Rossetti P, Graham C, Sharma H, Lin CF, Lebo MS: **Bioinformatics Workflow for Clinical Whole Genome Sequencing at Partners HealthCare Personalized Medicine**. *J Pers Med* 2016, **6**(1).
114. Thiel WH: **Galaxy Workflows for Web-based Bioinformatics Analysis of Aptamer High-throughput Sequencing Data**. *Mol Ther Nucleic Acids* 2016, **5**:e345.
115. Bai B, Laiho M: **Deep Sequencing Analysis of Nucleolar Small RNAs: Bioinformatics**. *Methods Mol Biol* 2016, **1455**:243-248.
116. Oliver GR, Hart SN, Klee EW: **Bioinformatics for clinical next generation sequencing**. *Clin Chem* 2015, **61**(1):124-135.
117. Lin J, Cheng Z, Xu M, Huang Z, Yang Z, Huang X, Zheng J, Lin T: **Genome re-sequencing and bioinformatics analysis of a nutraceutical rice**. *Mol Genet Genomics* 2015, **290**(3):955-967.
118. Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, Leon A, Pullambhatla M, Temple-Smolkin RL, Voelkerding KV *et al*: **Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists**. *J Mol Diagn* 2018, **20**(1):4-27.
119. Roh SW, Abell GC, Kim KH, Nam YD, Bae JW: **Comparing microarrays and next-generation sequencing technologies for microbial ecology research**. *Trends Biotechnol* 2010, **28**(6):291-299.
120. Willenbrock H, Salomon J, Sokilde R, Barken KB, Hansen TN, Nielsen FC, Moller S, Litman T: **Quantitative miRNA expression analysis: comparing microarrays with next-generation sequencing**. *RNA* 2009, **15**(11):2028-2034.
121. Mestdagh P, Hartmann N, Baeriswyl L, Andreasen D, Bernard N, Chen C, Cheo D, D'Andrade P, DeMayo M, Dennis L *et al*: **Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study**. *Nat Methods* 2014, **11**(8):809-815.
122. Rietveld T, van Hout R: **The paired t test and beyond: Recommendations for testing the central tendencies of two paired samples in research on speech, language and hearing pathology**. *J Commun Disord* 2017, **69**:44-57.
123. Jankowski KRB, Flannelly KJ, Flannelly LT: **The t-test: An Influential Inferential Tool in Chaplaincy and Other Healthcare Research**. *J Health Care Chaplain* 2018, **24**(1):30-39.
124. Wang JA, Qin Y, Lv J, Tian YF, Dong YJ: **Clinical application of high-sensitivity cardiac troponin T test in acute myocardial infarction diagnosis**. *Genet Mol Res* 2015, **14**(4):17959-17965.
125. Kim TK: **T test as a parametric statistic**. *Korean J Anesthesiol* 2015, **68**(6):540-546.
126. Rietveld T, van Hout R: **The t test and beyond: Recommendations for testing the central tendencies of two independent samples in research on speech, language and hearing pathology**. *J Commun Disord* 2015, **58**:158-168.
127. Skaik Y: **The bread and butter of statistical analysis "t-test": Uses and misuses**. *Pak J Med Sci* 2015, **31**(6):1558-1559.
128. Shokirov B: **Test for normality of the gene expression data**. *Methods Mol Biol* 2013, **972**:193-208.

129. Dexter F: **Wilcoxon-Mann-Whitney test used for data that are not normally distributed**. *Anesth Analg* 2013, **117**(3):537-538.
130. Marx A, Backes C, Meese E, Lenhof HP, Keller A: **EDISON-WMW: Exact Dynamic Programing Solution of the Wilcoxon-Mann-Whitney Test**. *Genomics Proteomics Bioinformatics* 2016, **14**(1):55-61.
131. Darabi H, Czene K, Zhao W, Liu J, Hall P, Humphreys K: **Breast cancer risk prediction and individualised screening based on common genetic variation and breast density measurement**. *Breast Cancer Res* 2012, **14**(1):R25.
132. Barlow M, Schlabach D, Peiffer J, Cook C: **Differences in change scores and the predictive validity of three commonly used measures following concussion in the middle school and high school aged population**. *Int J Sports Phys Ther* 2011, **6**(3):150-157.
133. Gengsheng Q, Hotilovac L: **Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test**. *Stat Methods Med Res* 2008, **17**(2):207-221.
134. Cook NR: **Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve**. *Clin Chem* 2008, **54**(1):17-23.
135. Steinbach G, Bolke E, Schulte am Esch J, Peiper M, Zant R, Schwarz A, Spiess B, van Griensven M, Orth K: **Comparison of whole blood interleukin-8 and plasma interleukin-8 as a predictor for sepsis in postoperative patients**. *Clin Chim Acta* 2007, **378**(1-2):117-121.
136. Collinson PO, Gaze DC, Morris F, Morris B, Price A, Goodacre S: **Comparison of biomarker strategies for rapid rule out of myocardial infarction in the emergency department using ACC/ESC diagnostic criteria**. *Ann Clin Biochem* 2006, **43**(Pt 4):273-280.
137. Millimet CR, Greenberg RP: **Use of an analysis of variance technique for investigating the differential diagnosis of organic versus functional involvement of symptoms**. *J Consult Clin Psychol* 1973, **40**(2):188-195.
138. Virmani AK, Tsou JA, Siegmund KD, Shen LY, Long TI, Laird PW, Gazdar AF, Laird-Offringa IA: **Hierarchical clustering of lung cancer cell lines using DNA methylation markers**. *Cancer Epidemiol Biomarkers Prev* 2002, **11**(3):291-297.
139. Cordes D, Haughton V, Carew JD, Arfanakis K, Maravilla K: **Hierarchical clustering to measure connectivity in fMRI resting-state data**. *Magn Reson Imaging* 2002, **20**(4):305-317.
140. Vlad MO: **Hierarchical clustering-jump approach to analogs of renormalization-group transformations in fractal random processes**. *Phys Rev A* 1992, **45**(6):3600-3614.
141. Chatterjee SN, Datta RK: **Hierarchical clustering of 54 races and strains of the mulberry silkworm, Bombyx mori L: Significance of biochemical parameters**. *Theor Appl Genet* 1992, **85**(4):394-402.
142. Bernstein IH, Garbin CP: **Hierarchical clustering of pain patients' MMPI profiles: a replication note**. *J Pers Assess* 1983, **47**(2):171-172.
143. Harner EJ, Slater PB: **Identifying medical regions using hierarchical clustering**. *Soc Sci Med* 1980, **14D**(1):3-10.
144. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X: **Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells**. *PLoS One* 2014, **9**(1):e78644.
145. Su Z, Li Z, Chen T, Li QZ, Fang H, Ding D, Ge W, Ning B, Hong H, Perkins RG *et al*: **Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys**. *Chem Res Toxicol* 2011, **24**(9):1486-1493.

146. Pullat J, Kusnezow W, Jaakson K, Beier M, Hoheisel JD, Metspalu A: **Arrayed primer extension on in situ synthesized 5'-->3' oligonucleotides in microchannels**. *N Biotechnol* 2008, **25**(2-3):133-141.
147. Niyazi M, Zehentmayr F, Niemoller OM, Eigenbrod S, Kretzschmar H, Schulze-Osthoff K, Tonn JC, Atkinson M, Mortl S, Belka C: **MiRNA expression patterns predict survival in glioblastoma**. *Radiat Oncol* 2011, **6**:153.
148. Niemoeller OM, Niyazi M, Corradini S, Zehentmayr F, Li M, Lauber K, Belka C: **MicroRNA expression profiles in human cancer cells after ionizing radiation**. *Radiat Oncol* 2011, **6**:29.
149. Mogensen J, Nielsen HB, Hofmann G, Nielsen J: **Transcription analysis using high-density micro-arrays of Aspergillus nidulans wild-type and creA mutant during growth on glucose or ethanol**. *Fungal Genet Biol* 2006, **43**(8):593-603.
150. Duc L, Neuenschwander S, Rehrauer H, Wagner U, Sobek J, Schlapbach R, Zeyer J: **Development and experimental validation of a nifH oligonucleotide microarray to study diazotrophic communities in a glacier forefield**. *Environ Microbiol* 2009, **11**(8):2179-2189.
151. Coskun M, Bjerrum JT, Seidelin JB, Troelsen JT, Olsen J, Nielsen OH: **miR-20b, miR-98, miR-125b-1*, and let-7e* as new potential diagnostic biomarkers in ulcerative colitis**. *World J Gastroenterol* 2013, **19**(27):4289-4299.
152. Bosl W, Mandel J, Jonikas M, Ramoni RB, Kohane IS, Mandl KD: **Scalable decision support at the point of care: a substitutable electronic health record app for monitoring medication adherence**. *Interact J Med Res* 2013, **2**(2):e13.
153. Chenevier-Gobeaux C, Bonnefoy-Cudraz E, Charpentier S, Dehoux M, Lefevre G, Meune C, Ray P, Sfbc SFCSTw: **High-sensitivity cardiac troponin assays: answers to frequently asked questions**. *Arch Cardiovasc Dis* 2015, **108**(2):132-149.
154. St John A, Price CP: **Existing and Emerging Technologies for Point-of-Care Testing**. *Clin Biochem Rev* 2014, **35**(3):155-167.
155. Lequin RM: **Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA)**. *Clin Chem* 2005, **51**(12):2415-2418.
156. Alnasir J, Shanahan HP: **A Novel Method to Detect Bias in Short Read NGS Data**. *J Integr Bioinform* 2017, **14**(3).
157. Erhard F, Zimmer R: **Count ratio model reveals bias affecting NGS fold changes**. *Nucleic Acids Res* 2015, **43**(20):e136.
158. Backes C, Sedaghat-Hamedani F, Frese K, Hart M, Ludwig N, Meder B, Meese E, Keller A: **Bias in High-Throughput Analysis of miRNAs and Implications for Biomarker Studies**. *Anal Chem* 2016, **88**(4):2088-2095.
159. Pritchard CC, Kroh E, Wood B, Arroyo JD, Dougherty KJ, Miyaji MM, Tait JF, Tewari M: **Blood cell origin of circulating microRNAs: a cautionary note for cancer biomarker studies**. *Cancer Prev Res (Phila)* 2012, **5**(3):492-497.
160. Kapushesky M, Adamusiak T, Burdett T, Culhane A, Farne A, Filippov A, Holloway E, Klebanov A, Kryvych N, Kurbatova N *et al*: **Gene Expression Atlas update--a value-added database of microarray and sequencing-based functional genomics experiments**. *Nucleic Acids Res* 2012, **40**(Database issue):D1077-1081.
161. Kahraman M, Laufer T, Backes C, Schrors H, Fehlmann T, Ludwig N, Kohlhaas J, Meese E, Wehler T, Bals R *et al*: **Technical Stability and Biological Variability in MicroRNAs from Dried Blood Spots: A Lung Cancer Therapy-Monitoring Showcase**. *Clin Chem* 2017, **63**(9):1476-1488.
162. Backes C, Ludwig N, Leidinger P, Huwer H, Tenzer S, Fehlmann T, Franke A, Meese E, Lenhof HP, Keller A: **Paired proteomics, transcriptomics and miRNomics in non-small cell lung cancers: known and novel signaling cascades**. *Oncotarget* 2016, **7**(44):71514-71525.

163. Hart M, Rheinheimer S, Leidinger P, Backes C, Menegatti J, Fehlmann T, Grasser F, Keller A, Meese E: **Identification of miR-34a-target interactions by a combined network based and experimental approach**. *Oncotarget* 2016, **7**(23):34288-34299.
164. Leidinger P, Keller A, Meese E: **MicroRNAs - Important Molecules in Lung Cancer Research**. *Front Genet* 2011, **2**:104.
165. Leidinger P, Brefort T, Backes C, Krapp M, Galata V, Beier M, Kohlhaas J, Huwer H, Meese E, Keller A: **High-throughput qRT-PCR validation of blood microRNAs in non-small cell lung cancer**. *Oncotarget* 2016, **7**(4):4611-4623.
166. Keller A, Leidinger P, Borries A, Wendschlag A, Wucherpfennig F, Scheffler M, Huwer H, Lenhof HP, Meese E: **miRNAs in lung cancer - studying complex fingerprints in patient's blood cells by microarray experiments**. *BMC Cancer* 2009, **9**:353.
167. Leidinger P, Backes C, Meder B, Meese E, Keller A: **The human miRNA repertoire of different blood compounds**. *BMC Genomics* 2014, **15**:474.
168. Leidinger P, Backes C, Dahmke IN, Galata V, Huwer H, Stehle I, Bals R, Keller A, Meese E: **What makes a blood cell based miRNA expression pattern disease specific?--a miRNome analysis of blood cell subsets in lung cancer patients and healthy controls**. *Oncotarget* 2014, **5**(19):9484-9497.
169. Juzenas S, Venkatesh G, Hubenthal M, Hoeppner MP, Du ZG, Paulsen M, Rosenstiel P, Senger P, Hofmann-Apitius M, Keller A *et al*: **A comprehensive, cell specific microRNA catalogue of human peripheral blood**. *Nucleic Acids Res* 2017, **45**(16):9290-9301.
170. Schwarz EC, Backes C, Knorck A, Ludwig N, Leidinger P, Hoxha C, Schwar G, Grossmann T, Muller SC, Hart M *et al*: **Deep characterization of blood cell miRNomes by NGS**. *Cell Mol Life Sci* 2016, **73**(16):3169-3181.
171. Backes C, Keller A: **Reanalysis of 3,707 novel human microRNA candidates**. *Proc Natl Acad Sci U S A* 2015, **112**(22):E2849-2850.
172. Zhang L, Lin J, Ye Y, Oba T, Gentile E, Lian J, Wang J, Zhao Y, Gu J, Wistuba, II *et al*: **Serum MicroRNA-150 Predicts Prognosis for Early-Stage Non-Small Cell Lung Cancer and Promotes Tumor Cell Proliferation by Targeting Tumor Suppressor Gene SRCIN1**. *Clin Pharmacol Ther* 2017.
173. Su K, Zhang T, Wang Y, Hao G: **Diagnostic and prognostic value of plasma microRNA-195 in patients with non-small cell lung cancer**. *World J Surg Oncol* 2016, **14**(1):224.
174. Guo J, Meng R, Yin Z, Li P, Zhou R, Zhang S, Dong X, Liu L, Wu G: **A serum microRNA signature as a prognostic factor for patients with advanced NSCLC and its association with tissue microRNA expression profiles**. *Mol Med Rep* 2016, **13**(6):4643-4653.
175. Chu G, Zhang J, Chen X: **Serum level of microRNA-147 as diagnostic biomarker in human non-small cell lung cancer**. *J Drug Target* 2016, **24**(7):613-617.
176. Mo D, Gu B, Gong X, Wu L, Wang H, Jiang Y, Zhang B, Zhang M, Zhang Y, Xu J *et al*: **miR-1290 is a potential prognostic biomarker in non-small cell lung cancer**. *J Thorac Dis* 2015, **7**(9):1570-1579.
177. Wu C, Cao Y, He Z, He J, Hu C, Duan H, Jiang J: **Serum levels of miR-19b and miR-146a as prognostic biomarkers for non-small cell lung cancer**. *Tohoku J Exp Med* 2014, **232**(2):85-95.
178. Wang ZX, Bian HB, Wang JR, Cheng ZX, Wang KM, De W: **Prognostic significance of serum miRNA-21 expression in human non-small cell lung cancer**. *J Surg Oncol* 2011, **104**(7):847-851.
179. Zhou R, Zhou X, Yin Z, Guo J, Hu T, Jiang S, Liu L, Dong X, Zhang S, Wu G: **Tumor invasion and metastasis regulated by microRNA-184 and microRNA-574-5p in small-cell lung cancer**. *Oncotarget* 2015, **6**(42):44609-44622.

180. Tiberio P, Callari M, Angeloni V, Daidone MG, Appierto V: **Challenges in using circulating miRNAs as cancer biomarkers**. *Biomed Res Int* 2015, **2015**:731479.
181. Moldovan L, Batte KE, Trgovcich J, Wisler J, Marsh CB, Piper M: **Methodological challenges in utilizing miRNAs as circulating biomarkers**. *J Cell Mol Med* 2014, **18**(3):371-390.
182. Polman CH, Reingold SC, Banwell B, Clanet M, Cohen JA, Filippi M, Fujihara K, Havrdova E, Hutchinson M, Kappos L *et al*: **Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria**. *Ann Neurol* 2011, **69**(2):292-302.
183. Polman CH, Reingold SC, Edan G, Filippi M, Hartung HP, Kappos L, Lublin FD, Metz LM, McFarland HF, O'Connor PW *et al*: **Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria"**. *Ann Neurol* 2005, **58**(6):840-846.
184. Miller DH, Weinshenker BG, Filippi M, Banwell BL, Cohen JA, Freedman MS, Galetta SL, Hutchinson M, Johnson RT, Kappos L *et al*: **Differential diagnosis of suspected multiple sclerosis: a consensus approach**. *Mult Scler* 2008, **14**(9):1157-1174.
185. Cummings J, Aisen PS, DuBois B, Frolich L, Jack CR, Jr., Jones RW, Morris JC, Raskin J, Dowsett SA, Scheltens P: **Drug development in Alzheimer's disease: the path to 2025**. *Alzheimers Res Ther* 2016, **8**:39.
186. Hegde ML, Bharathi P, Suram A, Venugopal C, Jagannathan R, Poddar P, Srinivas P, Sambamurti K, Rao KJ, Scancar J *et al*: **Challenges associated with metal chelation therapy in Alzheimer's disease**. *J Alzheimers Dis* 2009, **17**(3):457-468.
187. Hendrix JA, Bateman RJ, Brashear HR, Duggan C, Carrillo MC, Bain LJ, DeMattos R, Katz RG, Ostrowitzki S, Siemers E *et al*: **Challenges, solutions, and recommendations for Alzheimer's disease combination therapy**. *Alzheimers Dement* 2016, **12**(5):623-630.
188. Sabbagh MN, Richardson S, Relkin N: **Disease-modifying approaches to Alzheimer's disease: challenges and opportunities-Lessons from donepezil therapy**. *Alzheimers Dement* 2008, **4**(1 Suppl 1):S109-118.
189. Zafari S, Backes C, Meese E, Keller A: **Circulating Biomarker Panels in Alzheimer's Disease**. *Gerontology* 2015, **61**(6):497-503.
190. Keller A, Leidinger P, Vogel B, Backes C, ElSharawy A, Galata V, Mueller SC, Marquart S, Schrauder MG, Strick R *et al*: **miRNAs can be generally associated with human pathologies as exemplified for miR-144**. *BMC Med* 2014, **12**:224.
191. Chou CH, Shrestha S, Yang CD, Chang NW, Lin YL, Liao KW, Huang WC, Sun TH, Tu SJ, Lee WH *et al*: **miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions**. *Nucleic Acids Res* 2018, **46**(D1):D296-D302.
192. Chou CH, Chang NW, Shrestha S, Hsu SD, Lin YL, Lee WH, Yang CD, Hong HC, Wei TY, Tu SJ *et al*: **miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database**. *Nucleic Acids Res* 2016, **44**(D1):D239-247.
193. Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH, Chu CF, Huang HY, Lin CM, Ho SY *et al*: **miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions**. *Nucleic Acids Res* 2014, **42**(Database issue):D78-85.
194. Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, Tsai WT, Chen GZ, Lee CJ, Chiu CM *et al*: **miRTarBase: a database curates experimentally validated microRNA-target interactions**. *Nucleic Acids Res* 2011, **39**(Database issue):D163-169.
195. Li JH, Liu S, Zhou H, Qu LH, Yang JH: **starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data**. *Nucleic Acids Res* 2014, **42**(Database issue):D92-97.
196. Yang JH, Li JH, Shao P, Zhou H, Chen YQ, Qu LH: **starBase: a database for**

exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Res* 2011, **39**(Database issue):D202-209.
197. Selkoe DJ: **The cell biology of beta-amyloid precursor protein and presenilin in Alzheimer's disease**. *Trends Cell Biol* 1998, **8**(11):447-453.
198. O'Brien RJ, Wong PC: **Amyloid precursor protein processing and Alzheimer's disease**. *Annu Rev Neurosci* 2011, **34**:185-204.
199. Backes C, Kehl T, Stockel D, Fehlmann T, Schneider L, Meese E, Lenhof HP, Keller A: **miRPathDB: a new dictionary on microRNAs and target pathways**. *Nucleic Acids Res* 2017, **45**(D1):D90-D96.
200. Backes C, Meese E, Lenhof HP, Keller A: **A dictionary on microRNAs and their putative target pathways**. *Nucleic Acids Res* 2010, **38**(13):4476-4486.
201. Kehl T, Backes C, Kern F, Fehlmann T, Ludwig N, Meese E, Lenhof HP, Keller A: **About miRNAs, miRNA seeds, target genes and target pathways**. *Oncotarget* 2017, **8**(63):107167-107175.
202. Endy D: **Foundations for engineering biology**. *Nature* 2005, **438**(7067):449-453.
203. Serrano L: **Synthetic biology: promises and challenges**. *Mol Syst Biol* 2007, **3**:158.
204. Smith M: **Applications of synthetic oligodeoxyribonucleotides to problems in molecular biology**. *Nucleic Acids Symp Ser* 1980(7):387-395.
205. Ellis RW: **The applications of synthetic oligonucleotides to molecular biology**. *Pharm Res* 1986, **3**(4):195-207.
206. Smith M: **Nobel lecture. Synthetic DNA and biology**. *Biosci Rep* 1994, **14**(2):51-66.
207. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM *et al*: **Creation of a bacterial cell controlled by a chemically synthesized genome**. *Science* 2010, **329**(5987):52-56.
208. Seghal Kiran G, Ramasamy P, Sekar S, Hassan S, Ninawe AS, Selvin J: **Synthetic biology approaches: Towards sustainable exploitation of marine bioactive molecules**. *Int J Biol Macromol* 2018.
209. Cao J, de la Fuente-Nunez C, Ou RW, Torres MD, Pande SG, Sinskey AJ, Lu TK: **Yeast-based synthetic biology platform for antimicrobial peptide production**. *ACS Synth Biol* 2018.
210. Brenner MJ, Cho JH, Wong NML, Wong WW: **Synthetic Biology: Immunotherapy by Design**. *Annu Rev Biomed Eng* 2018.
211. Anderson LA, Islam MA, Prather KLJ: **Synthetic biology strategies for improving microbial synthesis of "green" biopolymers**. *J Biol Chem* 2018.
212. Dyo YM, Purton S: **The algal chloroplast as a synthetic biology platform for production of therapeutic proteins**. *Microbiology* 2018, **164**(2):113-121.
213. Heidari Feidt R, Ienca M, Elger BS, Folcher M: **Synthetic Biology and the Translational Imperative**. *Sci Eng Ethics* 2017.
214. Gerstmans H, Criel B, Briers Y: **Synthetic biology of modular endolysins**. *Biotechnol Adv* 2017.
215. Averesch NJH, Martinez VS, Nielsen LK, Kromer JO: **Toward Synthetic Biology Strategies for Adipic Acid Production: An in Silico Tool for Combined Thermodynamics and Stoichiometric Analysis of Metabolic Networks**. *ACS Synth Biol* 2017.
216. Kowarschik K, Hoehenwarter W, Marillonnet S, Trujillo M: **UbiGate: a synthetic biology toolbox to analyse ubiquitination**. *New Phytol* 2017.
217. Kim HJ, Jeong H, Lee SJ: **Synthetic biology for microbial heavy metal biosensors**. *Anal Bioanal Chem* 2018, **410**(4):1191-1203.

Anhang 2

Patentverzeichnis

Title	Publikations-nummer	Datum	Erfinder
Immunoassay for detection of miRNAs	SI2639312 (T1)	29.12.17	KAPPEL ANDREAS [DE] KELLER ANDREAS [DE] STAHLER CORD FRIEDRICH [DE] WRIGHT IAN [US] ANDERSON-MAUSER LINDA MARIE [US] BEDZYK WILLIAM [US] SCHWARZ HERBERT [DE] WICKE MICHAELA [DE]
Genetic testing for predicting resistance of Shigella species against antimicrobial agents	AU2016273220 (A1)	21.12.17	KELLER ANDREAS SCHMOLKE SUSANNE STÄHLER CORD FRIEDRICH BACKES CHRISTINA
GENETIC TESTING FOR PREDICTING RESISTANCE OF KLEBSIELLA SPECIES AGAINST ANTIMICROBIAL AGENTS	US2017283862 (A1)	05.10.17	KELLER ANDREAS [DE] SCHMOLKE SUSANNE [DE] STAHLER CORD FRIEDRICH [DE] BACKES CHRISTINA [DE]
Diagnosis of neuromyelitis optica vs. multiple sclerosis using mirna biomarkers	CN106661623 (A)	10.05.17	KELLER ANDREAS KIRSTEN JAN STAHLER CORD FRIEDRICH BACKES CHRISTINA LEIDINGER PETRA MEESE ECKART
MIRNA FINGERPRINT IN THE DIAGNOSIS OF MULTIPLE SCLEROSIS	EP3184651 (A1)	28.06.17	KELLER ANDREAS [DE] MEESE ECKART [DE] BORRIES ANNE [DE] STÄHLER PEER [DE] BEIER MARKUS [DE]
GENETIC RESISTANCE TESTING	US2017009277 (A1)	12.01.17	BACKES CHRISTINA [DE] KELLER ANDREAS [DE] KIRSTEN JAN [DE] RENSEN GABRIEL [US] SCHMOLKE SUSANNE [DE] STÄHLER CORD FRIEDRICH [DE]
SPECIFIC SIGNATURES IN ALZHEIMER'S DISEASE BY MEANS OF MULTICENTER MIRNA PROFILES	WO2017108535 (A1)	29.06.17	KELLER ANDREAS [DE] STÄHLER CORD FRIEDRICH [DE] SICKERT DANIEL [DE] BACKES CHRISTINA [DE]
DIAGNOSTIC MIRNA SIGNATURES IN MS AND CIS PATIENTS	WO2017108491 (A1)	29.06.17	KELLER ANDREAS [DE] STÄHLER CORD FRIEDRICH [DE] GEISEN STEFANIE [DE] SICKERT DANIEL [DE]
DIAGNOSTIC MIRNA MARKERS FOR ALZHEIMER	US2016273040 (A1)	22.09.16	KELLER ANDREAS [DE] STÄHLER CORD FRIEDRICH [DE] MEESE ECKART [DE] KAPPEL ANDREAS [DE] BACKES CHRISTINA [DE] LEIDINGER PETRA [DE]
COMPLEX MIRNA SETS AS NOVEL BIOMARKERS FOR AN ACUTE CORONARY SYNDROME	EP3124623 (A1)	01.02.17	KELLER ANDREAS [DE] STÄHLER PEER [DE] BEIER MARKUS [DE] MEDER BENJAMIN [DE] KATUS HUGO [DE] ROTTBAUER WOLFGANG [DE]
GENETIC RESISTANCE PREDICTION AGAINST ANTIMICROBIAL DRUGS IN MICROORGANISM USING STRUCTURAL CHANGES IN THE GENOME	WO2017021529 (A1)	09.02.17	KELLER ANDREAS [DE] SCHMOLKE SUSANNE [DE] STÄHLER CORD FRIEDRICH [DE] BACKES CHRISTINA [DE] GALATA VALENTINA [DE]
GENETIC TESTING FOR PREDICTING RESISTANCE OF STENOTROPHOMONAS SPECIES AGAINST ANTIMICROBIAL AGENTS	WO2017017045 (A1)	02.02.17	KELLER ANDREAS [DE] SCHMOLKE SUSANNE [DE] STÄHLER CORD FRIEDRICH [DE] BACKES CHRISTINA [DE] GALATA VALENTINA [DE]
GENETIC TESTING FOR PREDICTING RESISTANCE OF ENTEROBACTER SPECIES AGAINST ANTIMICROBIAL AGENTS	WO2017017044 (A1)	02.02.17	KELLER ANDREAS [DE] SCHMOLKE SUSANNE [DE] STÄHLER CORD FRIEDRICH [DE] BACKES CHRISTINA [DE] GALATA VALENTINA [DE]

GENETIC TESTING FOR PREDICTING RESISTANCE OF MORGANELLA SPECIES AGAINST ANTIMICROBIAL AGENTS	WO2017013223 (A1)	26.01.17	KELLER ANDREAS [DE] SCHMOLKE SUSANNE [DE] STÄHLER CORD FRIEDRICH [DE] BACKES CHRISTINA [DE] GALATA VALENTINA [DE]
GENETIC TESTING FOR PREDICTING RESISTANCE OF SERRATIA SPECIES AGAINST ANTIMICROBIAL AGENTS	WO2017013220 (A2); WO2017013220 (A3)	26.01.17	KELLER ANDREAS [DE] SCHMOLKE SUSANNE [DE] STÄHLER CORD FRIEDRICH [DE] BACKES CHRISTINA [DE] GALATA VALENTINA [DE]
GENETIC TESTING FOR PREDICTING RESISTANCE OF GRAM-NEGATIVE PROTEUS AGAINST ANTIMICROBIAL AGENTS	WO2017013219 (A2); WO2017013219 (A3)	26.01.17	SCHMOLKE SUSANNE [DE] STÄHLER CORD FRIEDRICH [DE] KELLER ANDREAS [DE] BACKES CHRISTINA [DE]
GENETIC TESTING FOR PREDICTING RESISTANCE OF SALMONELLA SPECIES AGAINST ANTIMICROBIAL AGENTS	WO2017013217 (A2); WO2017013217 (A3)	26.01.17	KELLER ANDREAS [DE] SCHMOLKE SUSANNE [DE] STÄHLER CORD FRIEDRICH [DE] BACKES CHRISTINA [DE] GALATA VALENTINA [DE]
GENETIC TESTING FOR PREDICTING RESISTANCE OF PSEUDOMONAS SPECIES AGAINST ANTIMICROBIAL AGENTS	WO2017013204 (A1)	26.01.17	KELLER ANDREAS [DE] SCHMOLKE SUSANNE [DE] STÄHLER CORD FRIEDRICH [DE] BACKES CHRISTINA [DE] GALATA VALENTINA [DE]
GENETIC TESTING FOR PREDICTING RESISTANCE OF ACINETOBACTER SPECIES AGAINST ANTIMICROBIAL AGENTS	WO2017009374 (A1)	19.01.17	KELLER ANDREAS [DE] SCHMOLKE SUSANNE [DE] STÄHLER CORD FRIEDRICH [DE] BACKES CHRISTINA [DE] GALATA VALENTINA [DE]
METHOD AND SYSTEM FOR DETERMINING A BACTERIAL RESISTANCE TO AN ANTIBIOTIC DRUG	US2016162635 (A1)	09.06.16	KELLER ANDREAS [DE] STÄHLER CORD FRIEDRICH [DE] RENSEN GABRIEL [US]
Verfahren zum Erkennen von Mikroorganismen	DE102015206444 (B3)	19.05.16	KELLER ANDREAS [DE] STÄHLER CORD FRIEDRICH [DE] KOCH NIKOLA MARGA DIANA [DE] BACKES CHRISTINA [DE] LEIDINGER PETRA [DE] MEESE ECKART [DE]
COMBINATION OF STRUCTURAL VARIATIONS AND SINGLE NUCLEOTIDE CHANGES IN ONE STATISTICAL MODEL FOR IMPROVED THERAPY SELECTION	EP3216873 (A1)	13.09.17	BACKES CHRISTINA [DE] GALATA VALENTINA [DE] KELLER ANDREAS [DE] SCHMOLKE SUSANNE [DE] STÄHLER FRIEDRICH [DE]
NEW DIAGNOSTIC MIRNA MARKERS FOR PARKINSON DISEASE	US2015315641 (A1)	05.11.15	KELLER ANDREAS [DE] STÄHLER CORD FRIEDRICH [DE] KIRSTEN JAN [DE] MEESE ECKART [DE] BACKES CHRISTINA [DE] LEIDINGER PETRA [DE]
NOVEL MIRNAS AS DIAGNOSTIC MARKERS	US2015292013 (A1)	15.10.15	KELLER ANDREAS [DE] STÄHLER CORD FRIEDRICH [DE] BACKES CHRISTINA [DE] MEESE ECKART [DE] LEIDINGER PETRA [DE] KAPPEL ANDREAS [DE]
GENETIC TESTING FOR ALIGNMENT-FREE PREDICTING RESISTANCE OF MICROORGANISMS AGAINST ANTIMICROBIAL AGENTS	WO2017020967 (A1)	09.02.17	KELLER ANDREAS [DE] SCHMOLKE SUSANNE [DE] STÄHLER CORD FRIEDRICH [DE] BACKES CHRISTINA [DE] GALATA VALENTINA [DE]
GENETIC TESTING FOR PREDICTING RESISTANCE OF GRAM-NEGATIVE PROTEUS AGAINST ANTIMICROBIAL AGENTS	CA2991090 (A1)	26.01.17	SCHMOLKE SUSANNE [DE] STAEHLER CORD FRIEDRICH [DE] KELLER ANDREAS [DE] BACKES CHRISTINA [DE]
DIAGNOSIS OF NEUROMYELITIS OPTICA VS. MULTIPLE SCLEROSIS USING MIRNA BIOMARKERS	US2017130269 (A1)	11.05.17	KELLER ANDREAS [DE] KIRSTEN JAN [DE] STAEHLER CORD FRIEDRICH [DE] BACKES CHRISTINA [DE] LEIDINGER PETRA [DE] MEESE ECKART [DE]
miRNA FINGERPRINT IN THE DIAGNOSIS OF DISEASES	US2016076103 (A1); US9702008 (B2)	17.03.16	KELLER ANDREAS [DE] MEESE ECKART [DE] BORRIES ANNE [DE] STAEHLER PEER FRIEDRICH [DE] BEIER MARKUS [DE]
System and methods for integrated and predictive analysis of molecular, imaging, and clinical data for patient-specific management of diseases	CN104737172 (A)	24.06.15	MANSI TOMMASO LIM WEI KEAT KING VANESSA KREMER ANDREAS GEORGESCU BOGDAN ZHENG XUDONG KAMEN ALI KELLER ANDREAS STAEHLER CORD FRIEDRICH WIRSZ EMIL COMANICIU DORIN

Diagnostic miRNA profiles in multiple sclerosis	CN104428426 (A)	18.03.15	KELLER ANDREAS MEESE ECKART STAEHLER CORD FRIEDRICH KAPPEL ANDREAS LEIDINGER PETRA BACKES CHRISTINA
IMMUNOASSAY FOR DETECTION OF SPECIFIC NUCLEIC ACID SEQUENCES SUCH AS MIRNAS	KR20140137360 (A)	02.12.14	KAPPEL ANDREAS [DE] KELLER ANDREAS [DE] STAEHLER CORD FRIEDRICH [DE] WRIGHT IAN [GB] ANDERSON MAUSER LINDA MARIE [US] BEDZYK WILLIAM [US] SCHWARZ HERBERT [DE] WICKE MICHAELA [DE]
COMPUTER NETWORK FOR QUALITY TESTING CLINICAL TRIAL DATA	US2014058752 (A1)	27.02.14	BECHTOLD MARIO [DE] KELLER ANDREAS [DE] KUTH RAINER [DE] SCHMIDT GERD [DE] STAEHLER CORD FRIEDRICH [DE]
METHOD FOR IMPROVED QUANTIFICATION OF MIRNAS	WO2014001400 (A1)	03.01.14	KELLER ANDREAS [DE] KAPPEL ANDREAS [DE] STAEHLER CORD FRIEDRICH [DE]
Complex miRNA Sets as Novel Biomarkers for an Acute Coronary Syndrome	US2013157883 (A1); US9611511 (B2)	20.06.13	KELLER ANDREAS [DE] STAEHLER PEER F [DE] BEIER MARKUS [DE] MEDER BENJAMIN [DE] KATUS HUGO A [DE] ROTTBAUER WOLFGANG [DE]
METHOD FOR ANALYSIS OF NUCLEIC ACID POPULATIONS	US2012045771 (A1)	23.02.12	BEIER MARKUS [DE] STAEHLER PEER F [DE] STAEHLER CORD F [DE] SUMMERER DANIEL [DE] LEONARD JACK T [US] BAU STEPHAN [DE] CARUSO ANTHONY [US] SCHRACKE NADINE [DE] KELLER ANDREAS [DE] HANENBERG HELMUT [DE] ECKERMANN OLAF [DE]
SYNTHESIS OF SEQUENCE-VERIFIED NUCLEIC ACIDS	US2017267999 (A1)	21.09.17	STAEHLER PEER F [DE] CARAPITO RAPHAEL [FR] STAEHLER CORD F [DE] MATZAS MARK [DE] LEONARD JACK T [US] JAEGER JOACHIM [DE] BEIER MARKUS [DE]
METHOD FOR PRODUCING POLYMERS	US2017147748 (A1)	25.05.17	STAEHLER PEER F [DE] STAEHLER CORD F [DE] MUELLER MANFRED [DE]
Fremgangsmåde til fremstilling af polymerer	DK2175021 (T3)	23.09.13	STAEHLER PEER F [DE] STAEHLER CORD F [DE] MUELLER MANFRED [DE]
INDEXING OF NUCLEIC ACID POPULATIONS	US2012071327 (A1)	22.03.12	STAEHLER PEER F [DE] STAEHLER CORD F [DE] BEIER MARKUS [DE] CHEE MARK S [US] SCHRACKE NADINE [DE]
MICROFLUIDIC EXTRACTION METHOD	AT545019 (T)	15.02.12	MUELLER MANFRED [DE] STAEHLER CORD F [DE] STAEHLER PEER F [DE]
Biochemical analysis instrument uses mixture of sample and smart reagent beads viewed by color CCD camera to perform chemical analysis of medical samples.	AT535814 (T)	15.12.11	STAEHLER CORD F [DE] STAEHLER PEER F [DE] MUELLER MANFRED [DE] STAEHLER FRITZ [DE] LINDNER HANS [DE]
MULTI-USE OF BIOCHIPS	WO2010106109 (A1)	23.09.10	STAEHLER PEER F [DE] STAEHLER CORD F [DE] BEIER MARKUS [DE] PETERMANN RABEA [DE] GUEIMIL GARCIA RAMON [DE]
INTEGRATED AMPLIFICATION, PROCESSING AND ANALYSIS OF BIOMOLECULES IN A MICROFLUIDIC REACTION MEDIUM	WO2010043418 (A2); WO2010043418 (A3)	22.04.10	STAEHLER PEER [DE] STAEHLER CORD F [DE] BEIER MARKUS [DE] SUMMERER DANIEL [DE]
INCREASING THE SENSITIVITY AND SPECIFICITY OF NUCLEIC ACID CHIP HYBRIDIZATION TESTS	AT444376 (T)	15.10.09	STAEHLER CORD [DE] STAEHLER PEER [DE] BEIER MARKUS [DE]
IMPROVED MOLECULAR BIOLOGICAL PROCESSING SYSTEM	EP2109499 (A2)	21.10.09	STAEHLER PEER [DE] BEIER MARKUS [DE] STAEHLER CORD [DE] SUMMERER DANIEL [DE] MATZAS MARK [DE] VORWERK SONJA [DE]
		06.06.02	STAEHLER CORD F [DE] STROBELT TILO [DE]

Method for producing a fluid device, fluid device and analysis apparatus	US2002068021 (A1); US7226862 (B2)		FRECH JOHANNES [DE] NOMMENSEN PETER [DE] MUELLER MARTIN [DE]
Highly parallel template-based dna synthesizer	US2007087349 (A1)	19.04.07	STAEHLER PEER [DE] STAHLER CORD [DE] BEIER MARKUS [DE]
Method and device for the integrated synthesis and analysis of analytes on a support	US2004043509 (A1); US7470540 (B2)	04.03.04	STAHLER CORD F [DE] GUIMIL RAMON [DE] SCHEFFLER MATTHIAS [DE] STAHLER PEER F [DE] HEIDBREDE ANKE [DE]
Microfluid reaction carrier having three flow levels and a transparent protective layer	US7361314 (B1)	22.04.08	STAEHLER CORD FREDRICH [DE] MUELLER MANFRED [DE] STAEHLER PEER FRIEDRICH [DE] MAURITZ RALF [DE]
Dynamic determination of analytes using arrays on internal surfaces	EP1650314 (A1)	26.04.06	STAEHLER PEER F [DE] STAEHLER CORD F [DE] BEIER MARKUS [DE] SCHLAUERSBACH ANDREA [DE] BAUM MICHAEL [DE] MUELLER MANFRED [DE]
METHOD AND DEVICE FOR THE INTEGRATED SYNTHESIS AND ANALYSIS OF ANALYTES ON A SUPPORT	EP1330307 (A1); EP1330307 (B1); EP1330307 (B8)	30.07.03	STAEHLER CORD F [DE] GUEIMIL RAMON [DE] SCHEFFLER MATTHIAS [DE] STAEHLER PEER F [DE] HEIDBREDE ANKE [DE]
HYBRID METHOD FOR THE PRODUCTION OF CARRIERS FOR ANALYTE DETERMINATION	WO02089971 (A2); WO02089971 (A3)	14.11.02	STAEHLER CORD F [DE] STAEHLER PEER F [DE] BEIER MARKUS [DE] WIXMERTEN ANKE [DE] MAURITZ RALF [DE] SCHLAUERSBACH ANDREA [DE]
SYNTHESIS OF SEQUENCE-VERIFIED NUCLEIC ACIDS	DK2398915 (T3)	12.12.16	STÄHLER PEER F [DE] LEONARD JACK T [US] JÄGER JOACHIM [DE] BEIER MARKUS [DE] STÄHLER CORD F [DE] MATZAS MARK [DE] CARAPITO RAPHAEL [FR]
METHOD FOR IMPROVED QUANTIFICATION OF MIRNAS	US2015184223 (A1)	02.07.15	KELLER ANDREAS [DE] KAPPEL ANDREAS [DE] STÄHLER CORD FRIEDRICH [DE]
DIAGNOSTIC MIRNA PROFILES IN MULTIPLE SCLEROSIS	US2015111772 (A1)	23.04.15	KELLER ANDREAS [DE] MEESE ECKART [DE] STÄHLER CORD FRIEDRICH [DE] KAPPEL ANDREAS [DE] LEIDINGER PETRA [DE] BACKES CHRISTINA [DE]
SYSTEM AND METHODS FOR INTEGRATED AND PREDICTIVE ANALYSIS OF MOLECULAR, IMAGING, AND CLINICAL DATA FOR PATIENT-SPECIFIC MANAGEMENT OF DISEASES	DE112013003363 (T5)	02.04.15	COMANICIU DORIN [US] STÄHLER CORD FRIEDRICH [DE] WIRSZ EMIL [DE] KELLER ANDREAS [DE] KAMEN ALI [US] GEORGESCU BOGDAN [US] MANSI TOMMASO [US] ZHENG XUDONG [US] KREMER ANDREAS [AT] KING VANESSA [US] KEAT WEI [US]
MIRNAS AS ADVANCED DIAGNOSTIC TOOL IN PATIENTS WITH CARDIOVASCULAR DISEASE, IN PARTICULAR ACUTE MYOCARDIAL INFARCTION (AMI)	WO2015079060 (A2); WO2015079060 (A3)	04.06.15	KELLER ANDREAS [DE] MEDER BENJAMIN [DE] HAAS JAN [DE] VOGEL BRITTA [DE] KATUS HUGO A [DE] KIRSTEN JAN [DE] STÄHLER CORD FRIEDRICH [DE]
Prädiktion der Wirksamkeit eines Arzneimittels mittels 3D Modeling in der personalisierten Medizin	DE102013209424 (A1); DE102013209424 (B4)	27.11.14	KIRSTEN JAN [DE] STÄHLER CORD FRIEDRICH [DE] KELLER ANDREAS [DE] HENGERER ARNE [DE]
Method for obtaining gene signature scores	EP2733634 (A1)	21.05.14	KELLER ANDREAS [DE] STÄHLER CORD FRIEDRICH [DE]
IMPROVED MOLECULAR-BIOLOGICAL PROCESSING EQUIPMENT	US2011092380 (A1)	21.04.11	STAHLER PEER [DE] BEIER MARKUS [DE] STAHLER CORD [DE] SUMMERER DANIEL [DE] MATZAS MARK [DE] VORWERK SONJA [DE]
Hybrid method for the production of carriers for analyte determination	AU2002312899 (A1)	18.11.02	SCHLAUERSBACH ANDREA WIXMERTEN ANKE STAHLER PEER F STAHLER CORD F BEIER MARKUS MAURITZ RALF
METHOD FOR PRODUCING POLYMERS		02.07.09	STAHLER PEER F [DE] STAHLER CORD F [DE]

	US2009170802 (A1); US9568839 (B2)		MULLER MANFRED [DE]
METHOD AND DEVICE FOR INTEGRATED SYNTHESIS AND ANALYSIS OF ANALYTES ON A SUPPORT	US2009156423 (A1)	18.06.09	STAHLER CORD F [DE] GUIMIL RAMON [DE] SCHEFFLER MATTHIAS [DE] STAHLER PEER F [DE] HEIDBREDE ANKE [DE]
MICROFLUIDIC EXTRACTION METHOD	AT426158 (T)	15.04.09	ST HLER CORD [DE] ST HLER PEER [DE] M LLER MANFRED [DE]
METHOD AND DEVICE FOR PREPARING AND/OR ANALYZING BIOCHEMICAL REACTION CARRIERS	US2008214412 (A1)	04.09.08	STAHLER CORD F [DE] STAHLER PEER F [DE] MULLER MANFRED [DE] STAHLER FRITZ [DE] LINDNER HANS [DE]
Methods and apparatuses for electronic determination of analytes	US2005037407 (A1)	17.02.05	BEIER MARKUS [DE] STAHLER CORD F [DE]
Increasing the sensitivity and specificity of nucleic acid chip hybridization tests	US2005164407 (A1)	28.07.05	STAHLER CORD F [DE] STAHLER PEER F [DE] BEIER MARKUS [DE]
Dynamic sequencing by hybridization	US2003138790 (A1)	24.07.03	SCHLAUERSBACH ANDREA [DE] STAHLER CORD F [DE] STAHLER PEER F [DE] BAUM MICHAEL [DE] MULLER MANFRED [DE]
Dynamic determination of analytes	US2003138789 (A1)	24.07.03	STAHLER PEER [DE] STAHLER CORD F [DE] SCHLAMERSBACH ANDREA [DE] MULLER MANFRED [DE] BAUM MICHAEL [DE] BEIER MARKUS [DE]
Dynamic determination of analytes	US2005164293 (A1)	28.07.05	STAHLER PEER [DE] STAHLER CORD F [DE] SCHLAUERSBACH ANDREA [DE] MULLER MANFRED [DE] BAUM MICHAEL [DE] BEIER MARKUS [DE]
Microfluidic reaction support having three flow levels and a transparent cover layer	US2008132430 (A1)	05.06.08	STAHLER CORD FREDRICH [DE] MULLER MANFRED [DE] STAHLER PEER FRIEDRICH [DE] MAURITZ RALF PETER [DE]
Method for producing a fluid component, fluid component and an analysis device	AU2796902 (A)	18.06.02	STROBELT TILO FRECH JOHANNES NOMMENSEN PETER MULLER MARTIN STAHLER CORD-F

Anhang 3

Abbildungsverzeichnis

Anhang 4

Abkürzungsverzeichnis

Abkürzung	Bedeutung
AD	Alzheimer's Disease
ANOVA	Analysis of Variance
APP	Amyloid Precursor Protein
ATP	Adenosintriphosphat
AUC	Area under the Curve
BD	Bipolare Disorder
BH	Benjamin-Hochberg
CCD	Charge-coupled device
CIS	Clinical Isolated Syndrome
CNS	Central Nervous System
COPD	Chronisc Obstructive Pulmonary Disease
cPAS	Combinatorial probe-anchor synthesis
CRISPR Cas	Clustered Regularly Interspaced Short Palindromic Repeats)
CTP	Cytidintriphosphat
dATP	Deoxyadenosine triphosphate
dep	Depression
DNA	Deoxyribonucleic acid
DNB	DNA Nanoball
ENCODE	Encyclopedia of DNA Elements
ELISA	Enzyme-linked Immunosorbent Assay
FDR	False Discovery Rate
GTP	Guanosintriphosphat
HGP	Human Genome Project
hsa	Homo Sapiens
HTS	High-Throughput-Sequencing
lncRNA	Long non-coding RNA
MCI	Mild Cognitive Impairment
ML	Maschinelles Lernen
MMSE	Mini Mental State Exam
MPEA	Microfluidic primer extension assay
MRT	Magnet Resonanz Tomographie
MS	Multiple Sklerose
ncRNA	Non-Coding RNA

NGS	Next Generation Sequencing (siehe auch HTS)
NSCLC	Non Small-Cell Lung Carcinoma
PD	Parkinson's Disease
PoC	Point-of-Care
PCR	Polymerase Kettenreaktion
RNA	ribonucleic acid
ROC	Receiver-Operating-Characteristic
RRMSE	Relapsing-Remitting Multiple Sclerosis
RT-qPCR	Reverse Transcription quantitative polymerase chain reaction
SOP	Standard Operating Procedure
SVM	Support Vector Machines
TALEN	Transcription Activator-like Effector Nuclease
tNGS	Targeted Next-Generation Sequencing
TP	Time Point
tRNA	Transfer RNA
UTP	Uridintriphosphat
UTR	Untranslated Region
WMW	Wilcoxon-Mann-Whitney
ZNF	Zinc finger nucleases
ZNS	Zentrales Nervensystem

Anhang 5

Eigene Manuskripte

Als Anhang sind hier die vollständigen Originalarbeiten, die im Rahmen meiner Tätigkeit mit meinem Beitrag angefertigt wurden angefügt.

Nucleic Acids Research, 2003, Vol. 31, No. 23 ***e151***
DOI: 10.1093/nar/gng151

Validation of a novel, fully integrated and flexible microarray benchtop facility for gene expression profiling

Michael Baum*, Simone Bielau, Nicole Rittner, Kathrin Schmid, Kathrin Eggelbusch, Michael Dahms, Andrea Schlauersbach, Harald Tahedl, Markus Beier, Ramon Güimil, Matthias Scheffler, Carsten Hermann, Jörg-Michael Funk[1], Anke Wixmerten, Hans Rebscher, Matthias Hönig, Claas Andreae, Daniel Büchner, Erich Moschel, Andreas Glathe, Evelyn Jäger, Marc Thom, Andreas Greil, Felix Bestvater, Frank Obermeier, Josef Burgmaier, Klaus Thome, Sigrid Weichert, Silke Hein, Tim Binnewies, Volker Foitzik, Manfred Müller, Cord Friedrich Stähler and Peer Friedrich Stähler

febit ag, Käfertaler Strasse 190, 68167 Mannheim, Germany and [1]Carl Zeiss Jena GmbH, Carl Zeiss Group, Business Group: Microscopy, Division: Advanced Imaging Microscopy, Carl-Zeiss-Promenade 10, 07745 Jena, Germany

Received July 6, 2003; Revised September 20, 2003; Accepted October 9, 2003

ABSTRACT

Here we describe a novel microarray platform that integrates all functions needed to perform any array-based experiment in a compact instrument on the researcher's laboratory benchtop. Oligonucleotide probes are synthesized *in situ* via a light-activated process within the channels of a three-dimensional microfluidic reaction carrier. Arrays can be designed and produced within hours according to the user's requirements. They are processed in a fully automatic workflow. We have characterized this new platform with regard to dynamic range, discrimination power, reproducibility and accuracy of biological results. The instrument detects sample RNAs present at a frequency of 1:100 000. Detection is quantitative over more than two orders of magnitude. Experiments on four identical arrays with 6398 features each revealed a mean coefficient of variation (CV) value of 0.09 for the 6398 unprocessed raw intensities indicating high reproducibility. In a more elaborate experiment targeting 1125 yeast genes from an unbiased selection, a mean CV of 0.11 on the fold change level was found. Analyzing the transcriptional response of yeast to osmotic shock, we found that biological data acquired on our platform are in good agreement with data from Affymetrix GeneChips, quantitative real-time PCR and—albeit somewhat less clearly—to data from spotted cDNA arrays obtained from the literature.

INTRODUCTION

Microarrays have become a standard tool in molecular biology that has revolutionized genomics research. Microarrays are used extensively for gene expression profiling (1,2) in many applications including the discovery of gene function (3,4), drug evaluation (4–6), pathway dissection (7), classification of clinical samples (8–10), exon mapping (11) and investigation of splicing events (12). Arrays may be produced either by deposition of presynthesized material (1,13–15) or by *in situ* oligonucleotide synthesis (16,17). DNA arrays manufactured by physical deposition of presynthesized material require labor-intensive preparation and record-keeping of DNA probes. In contrast, oligonucleotide arrays synthesized *in situ* using a photolithographic method (18) only require DNA sequence data. However, cost and time spent in generating the photolithographic masks render this approach as slow and inflexible as the deposition methods. Recently, more flexible microarray technologies have been developed. These employ either ink-jet printing (19) or micromirror devices (20,21) for *in situ* synthesis of customized oligonucleotide arrays. Although these techniques provide full flexibility with respect to the array design, the actual generation of the array and in some cases even the hybridization and detection steps are restricted to centralized manufacturer facilities. Again, the investigator's flexibility remains limited. In addition, array synthesis and subsequent processing steps are not physically linked and require error-prone manual handling. The geniom platform described here is the first system to overcome these restrictions. The investigator gains full control of the complete workflow of any microarray experiment. The technology integrates microarray production, hybridization and detection in a compact benchtop unit. Automation of these processes

*To whom correspondence should be addressed. Tel: +49 621 3804 257; Fax: +49 621 3804 400; Email: michael.baum@febit.de

and a powerful software interface allow the scientist to design and perform microarray-based experiments using sequence information derived from public databases. Microarrays are generated by *in situ* oligonucleotide synthesis via a light-activated process employing a digital micromirror device and highly efficient photochemistry (22,23). Instead of a conventional microscope slide, a truly micro-machined three-dimensional microstructure bearing four individual channel-like chambers (arrays) is used as a reaction carrier. This approach allows one to run several array experiments on a single carrier since up to four individual microarrays are generated and may be hybridized sequentially or in parallel. In contrast to the recently described maskless array synthesizer, which also uses a micromirror device for *in situ* oligonucleotide synthesis (20), geniom is highly automated and integrates all functions required to perform an array-based experiment within a single device on the investigator's laboratory benchtop. A more detailed description of this technology is presented by Stähler *et al*. (24) and can also be found in Supplementary Material Figure 1.

In the study presented here, we characterized the geniom technology on a technical level with regard to dynamic range, discrimination power and reproducibility. In addition, we validated complex biological results acquired on the geniom platform by comparison with existing technologies and conventional standards. Analyzing the transcriptional response of *Saccharomyces cerevisiae* to osmotic shock, we found a good agreement of data obtained on geniom arrays, Affymetrix GeneChip data, and expression results obtained by quantitative real-time PCR. Our study also revealed a high concordance of geniom results and cDNA data from the literature (25). While the actual fold-change values are less consistent in this latter comparison, the vast majority of genes included in our study showed the same trend of regulation in both assay systems.

MATERIALS AND METHODS

Oligonucleotide arrays

Light-activated *in situ* oligonucleotide synthesis was performed essentially as described by Singh-Gasson *et al.* (20) using a digital micromirror device (Texas Instruments). The synthesis was performed within the geniom device on an activated three-dimensional reaction carrier consisting of a glass-silicon-glass sandwich (DNA processor; see Supplementary Material Fig. 1). Four individually accessible microchannels (referred to as arrays) etched into the silicon layer of the DNA processor are connected to the microfluidic system of the geniom device. Using standard DNA synthesis reagents (Proligo) and 3′-phosphoramidites carrying a 5′-photolabile protective group (22,23), oligonucleotides were synthesized in parallel in all four translucent arrays of one reaction carrier. Prior to synthesis, the glass surface was activated by coating with a spacer. The synthesized probe sets may be the same or different for all four arrays. Actually, the time needed for synthesis of standard arrays used in this study is independent of the number of different probe sets, the probe sequences and the number of probes synthesized within one probe set (current limit: 14 000 features per array; corresponding to $4 \times 14\,000 = 56\,000$ features per reaction carrier). However, the probe length substantially influences synthesis time. According to the conservative protocol used in this study, the synthesis of four typical 25mer arrays (with 12 880 features each) takes ~15.5 h (including 1.5 h for the final deprotection step). The yeast probe set (ten 25mer probes per transcript) was calculated based on the full genome sequence (retrieved online from http://genome-www.stanford.edu/Saccharomyces/) using a combination of sequence uniqueness criteria and rules for selection of oligonucleotides likely to hybridize with high specificity and sensitivity. The selection criteria were essentially as described in Lockhart *et al.* (2) with modifications for the longer probes used here (25mers instead of 20mers).

Yeast strain and growth conditions

Saccharomyces cerevisiae, wild-type strain W303-1A, MATa, ura3-52, trp1Δ2, leu2-3_112, his3-11, ade2-1, can1-100 (accession no. 20000A; EUROSCARF, Frankfurt a.M., Germany) was grown in 240 ml batch cultures at 30°C in YPD (1% yeast extract, 2% peptone, 2% glucose) to an A_{600} of 1.0. At this point, cells were collected for determination of expression profiles under baseline conditions. Osmotic stress was applied by adding prewarmed (30°C) 5 M NaCl to a final concentration of 0.7 M NaCl. Cells were collected 45 min after the addition of NaCl. Ten milliliters of suspension culture were chilled on ice, cells were pelleted, washed once with ice-cold water, frozen in a dry ice/ethanol bath and stored at –20°C until use.

RNA extraction and preparation for hybridization

Total RNA was extracted from frozen cell pellets using a hot phenol method (26). Amplification and labeling was achieved using a modification of the procedure first described by Van Gelder *et al*. (27) and Eberwine *et al.* (28). In brief, 5 µg of total RNA were used as a starting material and converted into double-stranded cDNA using an oligo(dT) primer with a 5′ T7 RNA polymerase promoter sequence and the Superscript II system for cDNA synthesis (Invitrogen). Double-stranded cDNA was purified by phenol–chloroform extraction followed by ethanol precipitation. Using the purified double-stranded cDNA as a template, *in vitro* transcription was performed using T7 RNA Polymerase (T7 Megascript Kit, Ambion) in the presence of a mixture of unlabeled ATP, CTP, GTP and UTP and biotin-labeled CTP and UTP [biotin-11-CTP (PerkinElmer); biotin-16-UTP (Roche)]. Biotinylated cRNA was purified on an affinity resin (RNeasy, Qiagen). The cRNA yield was determined by measuring the light absorbance at 260 nm (1 OD at 260 nm corresponds to 40 µg/ml RNA). Prior to hybridization, 15 µg of cRNA were fragmented randomly to an average length of ~100 nt by incubating at 94°C for 35 min in a 5 µl volume of 40 mM Tris-acetate pH 8.1, 100 mM potassium acetate and 30 mM magnesium acetate. A detailed description of the labeling protocol will be provided upon request. Transcripts of the ampicillin^r (amp^r), kanamycin^r (kan^r) and chloramphenicol^r (cm^r) resistance genes used for the determination of the dynamic range were prepared as follows. Each gene was PCR amplified from a plasmid vector (amp^r from pBR322; cm^r from pDNR-LIB; kan^r from pLP-GBKT7) and the PCR product was cloned into pBluescript II SK (+) (downstream of the T3 polymerase promotor sequence; between the BamHI and the EcoRI restriction sites). In

addition, an $A_{(50)}$ sequence was inserted between the EcoRI and HindIII sites of the same vector (immediately downstream of the resistance gene). Run-off transcripts [with a 3′ $A_{(50)}$ tail] were generated using the T3 Megascript Kit (Ambion) and 1 µg of the HindIII-digested construct as a template. One microgram of the *in vitro* transcript was used as a template for cRNA synthesis as described above. Different amounts of the biotinylated cRNA were then spiked into the yeast cRNA samples (prior to fragmentation).

Array hybridization, detection and data analysis

Microarrays were hybridized with 15 µg of fragmented cRNA in a final volume of 20 µl. Hybridization solutions contained 100 mM MES (pH 6.6), 0.9 M NaCl, 20 mM EDTA and 0.01% (v/v) Tween-20 (referred to as MES-hyb). In addition, the solutions contained 0.1 mg/ml sonicated herring sperm DNA (Promega) and 0.5 mg/ml BSA (Sigma). RNA samples were heated in the hybridization solution to 95°C for 3 min followed by 45°C for 3 min before being placed in an array which had been prehybridized for 15 min with 1% (w/v) BSA in MES-hyb at RT. Hybridizations were carried out at 45°C for 16 h without agitation. After removing the hybridization solutions, arrays were first washed with non-stringent buffer [0.005% (v/v) Triton X-100 in 6× SSPE] for 20 min at 25°C and subsequently with stringent buffer [0.005% (v/v) Triton X-100 in 0.5× SSPE] for 20 min at 45°C. After washing, the hybridized RNA was fluorescence-stained by incubating with 10 µg/ml streptavidin–phycoerythrin (Molecular Probes) and 2 µg/µl BSA in 6× SSPE at 25°C for 15 min. Unbound streptavidin–phycoerythrin was removed by washing with non-stringent buffer for 20 min at 25°C. Detection and feature readout were performed using the CCD-based detection system of the geniom device (Cy3 filter set). Processing of raw data including background correction, array to array normalization and determination of gene expression levels as well as calculation of fold-change values were essentially as described by Zhou and Abagyan (29). All steps were carried out using the PROP algorithm of the geniom application software which is based on the MOID algorithm described by Zhou and Abagyan (29).

Affymetrix GeneChip reference data

Aliquots of the same biotinylated cRNA samples analyzed on the geniom platform were sent to a service provider. The samples were hybridized to Affymetrix yeast GeneChips (YG-S98) according to the protocol in the Affymetrix GeneChip Expression Manual. Starting from the raw data files (.cel files), analysis was performed using both the Affymetrix MAS4 algorithm (at the service provider) and the PROP algorithm (at febit).

Quantitative PCR

In vitro transcripts [with a 3′ $A_{(50)}$ tail] of amp^r (250 pg), kan^r (25 pg) and cm^r (2.5 pg) were spiked into 5 µg of total RNA from yeast (control and treated). cRNA was prepared as described above but omitting the biotin labeling. The cRNA was then converted into cDNA using random hexamer primers and the Superscript II Kit. Quantitative PCR was performed using the iCycler iQ™ (Bio-Rad). Reactions contained ~250 pg non-purified cDNA, 300 nM forward and reverse primers (designed using the DNAMAN software; sequences will be provided upon request) and 25 µl of 2× QuantiTect SYBR Green PCR Master Mix (Qiagen) in a final volume of 50 µl. Samples were incubated for 13.5 min at 95°C followed by 50 cycles of denaturation (30 s at 95°C), annealing (30 s at 62°C) and extension (45 s at 72°C). The data obtained were normalized using all three spike-in controls. Fold-change values were calculated taking the PCR efficiencies into account (30,31).

RESULTS

Dynamic range and discrimination power

Spiking experiments were performed to determine the dynamic range of oligonucleotide arrays processed on the geniom platform. Biotinylated cRNAs from three prokaryotic genes (antibiotic resistance genes: amp^r, kan^r, cm^r) were mixed and spiked into 0.75 µg/µl biotinylated cRNA background from yeast total RNA at molar ratios of 1:100–1:100 000. In addition, kan^r and cm^r cRNAs were spiked at a molar ratio of 1:10. Using an estimate of 15 000 copies of mRNA per yeast cell (32–34) a frequency of 1:100 000 corresponds to that of an mRNA present at a density of one copy per six to seven cells. In 15 µg of cRNA background and a hybridization volume of 20 µl, a frequency of 1:100 000 corresponds to a concentration of ~22.7 pM and an absolute amount of 0.45 fmol (approximately 2.7×10^8 molecules or ~0.15 ng) of specific RNA. Each combination of dilution and background was hybridized six times with the exception of the 1:10 ratios which do not reflect situations encountered in normal cells and therefore were hybridized only once. In order to ensure optimal comparability of the data generated with the geniom instrument to those from other *in situ* synthesized short oligonucleotide arrays that mostly include mismatch (MM) controls, all samples were hybridized to arrays containing 16 perfect match (PM)/MM probe pairs (25mers) for each of 100 randomly chosen yeast genes, and 20 PM/MM probe pairs (25mers) for each of the three prokaryotic genes, although the geniom application software does not necessarily require MM probes for gene expression analysis. The arrays had been pretested for cross-hybridization. Yeast probes cross-hybridizing to the spiked-in transcripts as well as probes designed for these transcripts cross-hybridizing to the yeast background had been removed.

As indicated in Figure 1, the hybridization intensity is linearly correlated to the cRNA target concentration in the range of 1:100 000–1:1000. In the range of 1:1000–1:100, the signal increases by a factor of approximately six rather than 10 because the probes immobilized on the array are beginning to saturate. Between 1:100 and 1:10, saturation proceeds and the hybridization signal only increases by a factor of 1.5. At a molar ratio of approximately 1:100 000, the critical level for the discrimination power of the system is reached. While the presence of the prokaryotic transcripts was detected above the background in 14 out of 18 experiments at this level (six replicate hybridizations for each of the three genes), the remaining four experiments (three times kan^r and once amp^r) indicate that a ratio of 1:100 000 is the threshold level for at least some probe sets. In experiments lacking the complex cRNA background, the transcripts could be detected at concentrations corresponding to a ratio of 1:1 000 000 (data

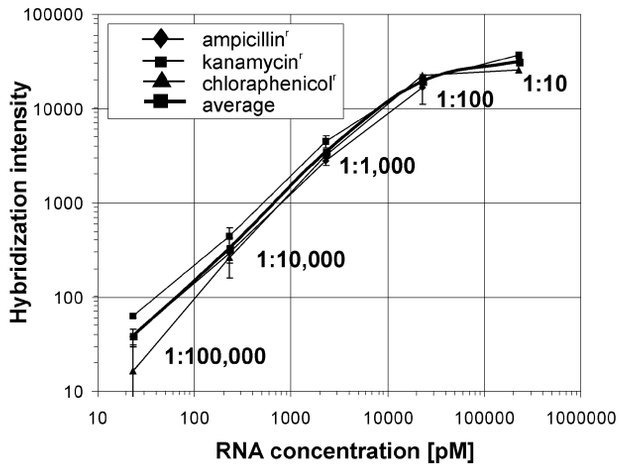


Figure 1. Dynamic range of oligonucleotide arrays from the geniom platform. Log–log plot of the normalized hybridization intensity (average of the 20 PM–MM intensity differences for each gene) versus concentration for three different prokaryotic cRNA targets. The three cRNA targets (amp^r, kan^r, cm^r) were spiked into labeled yeast cRNA at molar ratios of 1:100 000–1:100 and each dilution was measured six times. kan^r and cm^r cRNA was additionally spiked once at a molar ratio of 1:10. The error bars indicate the standard deviation calculated across the replicates after elimination of outliers.

not shown). The dynamic range of two to three orders of magnitude and the discrimination power of 1:100 000 measured here for arrays of the geniom platform compare very well to data obtained with other commercially available *in situ* synthesized (35) or spotted (13) oligonucleotide arrays. For the Affymetrix GeneChips a dynamic range of three to four orders of magnitude was initially reported (2). However, these data were obtained using a customized array containing probe sets with more than 500 PM/MM probe pairs per transcript. In a more recent study on commercial GeneChips with 20 PM/MM probe pairs per gene, a linear relationship between transcript abundance and signal intensity was observed at ratios of 1:150 000–1:15 000. Linearity ceased above the 1:15 000 ratio and saturation emerged around the 1:150 level (36).

Reproducibility of raw data

Replicate experiments were performed to determine the reproducibility of array synthesis, hybridization and technical readout. Aliquots of the same cRNA sample were hybridized to four identical arrays and the coefficient of variation (CV) for each individual feature was calculated based on the raw fluorescence intensities across the four replicates without applying any data preprocessing steps like background correction, array-to-array normalization, removal of outliers or removal of low-intensity spots. Since we expected the CV to be higher for features with a low intensity and lower for features with a high intensity we again designed the arrays with PM/MM probe pairs to obtain a balanced ratio between high intensity (PM probes) and low intensity features (MM probes). The four arrays each contained 6398 25mer probes (corresponding to 3199 PM/MM probe pairs). The probe sequences were derived from the Affymetrix HuGeneFL and the Test2 GeneChips. In addition, each array included 154 negative control features where a single 'T' mononucleotide was synthesized instead of a 25mer probe. The arrays were hybridized to aliquots of a cRNA sample from a pool of total RNAs (*Homo sapiens*, *Arabidopsis thaliana*, *Drosophila melanogaster*). This sample may be inappropriate for meaningful biological experiments focusing on the expression of specific genes but is very well suited for experiments with a technical scope. Due to its high complexity, this sample is likely to undergo specific hybridization, unspecific cross-hybridization (including cross-species hybridization) as well as extensive target–target interactions and thus will serve as a good indicator for the reproducibility of the array synthesis and the hybridization process in particular. For the analysis, we first performed a pairwise comparison of the four arrays (Fig. 2A–F). The average Pearson correlation coefficient calculated on the raw intensities for all possible combinations of two arrays was 0.986. To further investigate the reproducibility of the system on the raw data level, the CV for each of the 6398 features was calculated across the four replicates and CVs were plotted as a frequency distribution (Fig. 2G): 95% of all 6398 values were in the range of 0.03 (2.5th percentile) to 0.19 (97.5th percentile), the median CV being 0.09. A slightly higher median CV of 0.10 was found when the analysis was restricted to the 10% of features with the lowest intensities. These features do not represent the lowest features within a group consisting of only high-intensity features but indeed have very low intensities close to non-specific background. This is evident from the comparison of the average intensity of these features to the local background and to the negative control spots, where a single 'T' was synthesized instead of a 25mer probe. The average intensity of the 10% lowest features within the total of 6398 features (value: 911), the average of the local background of all spots on the array (value: 1198) and the average intensity for the negative controls (value: 1040) were all in the same range. Actually, the average of the 10% lowest features is even slightly lower than the average of the negative controls and the average of the local background. The latter phenomenon is due to the fact that the local background—at least for high intensity features—is increased by a 'neighborhood' effect caused by blooming of the hybridization signal. This is in agreement with a recent study published by Machl *et al.* that describes a similar 'neighborhood' effect for cDNA arrays spotted on nylon membranes and hybridized with radioactive labeled samples (37). Why the average intensity of the negative control features somewhat exceeds the average intensity found for the 10% lowest features is less obvious. A possible explanation could result from the higher negative charge of a 25mer probe as compared with a single 'T' nucleotide. In this case, the higher density of negative charges would lead to an increased repulsion of the equally negatively charged non-cognate targets that might reduce unspecific binding of non-cognate targets at the 25mer features. Another possible explanation is that steric hindrance for non-specific binding of the streptavidin–phycoerythrin complex to the glass surface might be higher for a feature with 25mers than for a feature carrying 'T' mononucleotides. This could result in a slightly higher blocking effect of 25mers as compared with 'T' mononucleotides. In summary, our technical experiments indicate a high reproducibility of geniom arrays on the raw data level and suggest that the good reproducibility is

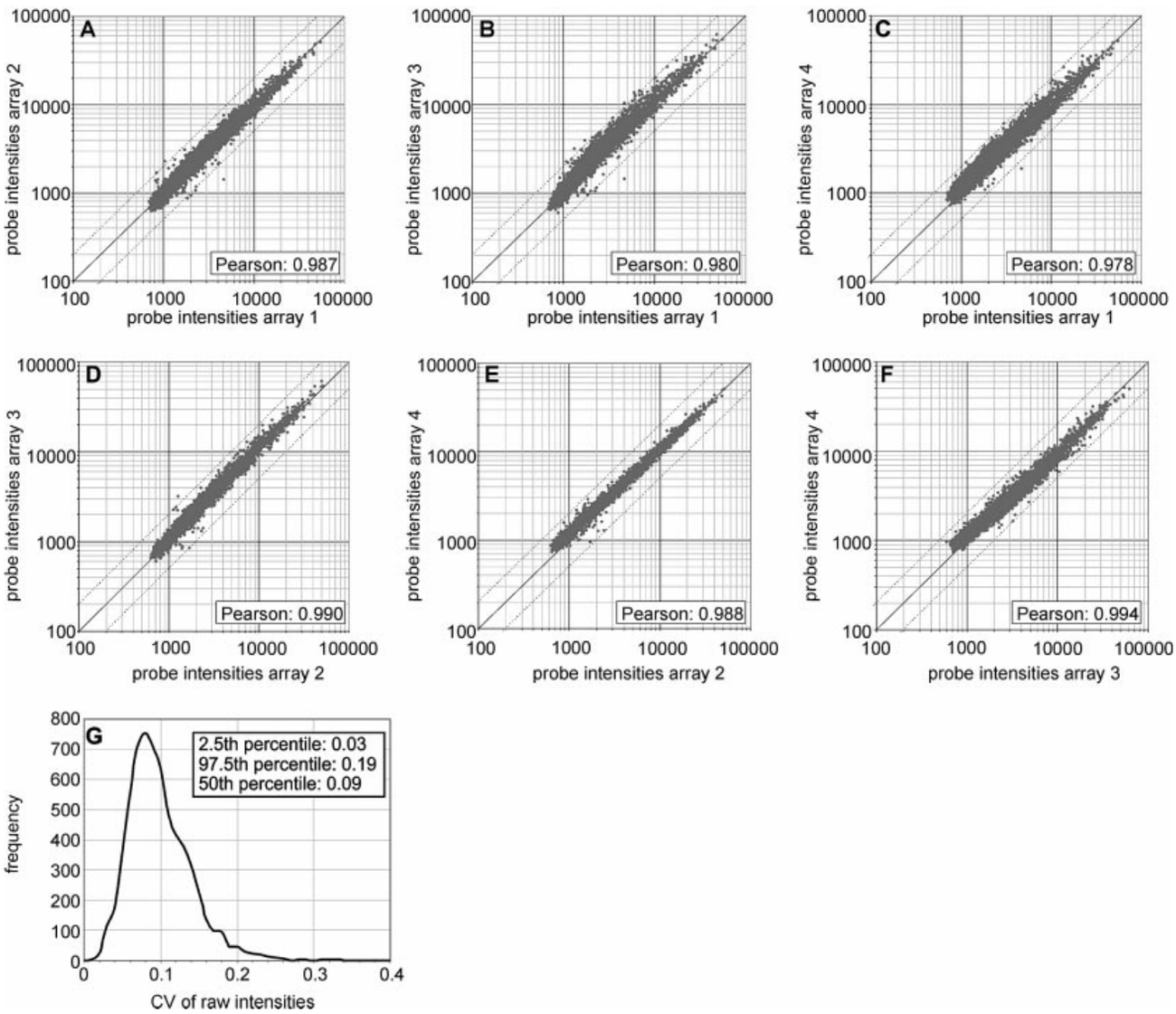


Figure 2. Reproducibility of the geniom platform on the raw data level. Aliquots of a single cRNA sample from a pool of total RNA (*H.sapiens*, *D.melanogaster*, *A.thaliana*) have been hybridized to four different arrays with 6398 features each. Raw intensity values (Supplementary Material Table 1) represent the median of approximately 30 CCD pixels for each feature. No data preprocessing (such as background correction, normalization, elimination of outliers or removal of low intensity features) was performed. (**A**–**F**) Pairwise comparison of raw intensities from the four arrays as scatter plots. (**G**) Frequency distribution of CVs. The CV for each of the 6398 features (probes) was calculated across the four replicates and CVs were plotted as a frequency distribution.

retained when applying geniom arrays to complex biological expression profiling experiments with the majority of features being in the low intensity range. However, in this case the average CV value might be slightly higher compared to our analysis with an unbiased distribution of raw intensity data across the entire intensity range.

Reproducibility of fold-change and expression level values

Having demonstrated a high reproducibility for the raw intensity data, we evaluated the variability of fold-change values, the ultimate result of standard gene expression profiling experiments. We therefore measured the transcriptional response of 1125 randomly chosen yeast genes to osmotic shock in four identical experiments on eight arrays. In contrast to the technical experiments described in the previous sections, this experiment was designed as a real-world gene expression profiling. As a consequence, the array design, which included MM controls beforehand, was adapted to our standard for expression arrays and the MM controls were omitted. This approach was supported by the geniom application software which operates on a algorithm similar to the MOID principle (29) for gene expression profiling experiments and thus does not require MM controls for calculating expression levels and fold-change values. The eight arrays used in this study each contained 12 880 features (including all controls) with ten 25mer PM probes per transcript. Following hybridization with aliquots of either a control sample or a treated sample, we first calculated the CV of the 12 880 unprocessed raw intensities across the four arrays hybridized

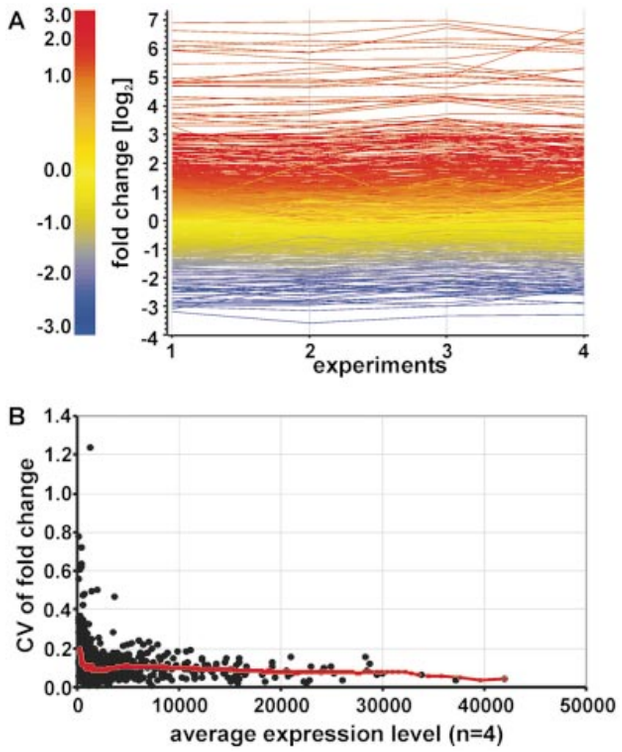


Figure 3. Reproducibility of the geniom platform on the fold-change level. The transcriptional response of 1125 yeast genes to osmotic shock was analyzed in four identical experiments and the fold-change values (Supplementary Material Table 2A) were compared. (**A**) Diagram of $\log_2$ fold-change values. Each transcript is represented as a line colored according to the $\log_2$ fold-change value. The color code is given on the left. (**B**) CVs of fold-change values. The CV for each of the 1125 genes was calculated across the four experiments and graphed as a function of the gene's expression level (E_k value). The gene's expression level represents the average of the E_k values from the four control arrays. A trend line representing the moving average of 100 genes is shown.

with the same sample. Mean CV values of 0.12 and 0.10 were found for the arrays hybridized with the control sample and the treated sample, respectively. A pairwise comparison of raw intensity data from the four control arrays in each possible combination revealed a mean Pearson correlation coefficient of 0.984 (min: 0.979; max: 0.993). In an identical analysis performed on the four arrays, hybridized with the cRNA sample from osmotically shocked yeast cells, we found a mean Pearson correlation coefficient of 0.986 (min: 0.977; max: 0.995). In conclusion, these values confirm the high reproducibility of the raw intensity data demonstrated in the last section and also suggest that the CV of the raw data is almost the same for arrays designed with PM/MM probe pairs (last section) and arrays with PM probes only (this section).

We next focused on the reproducibility of fold-change values obtained from geniom arrays. Fold-change values were calculated based on background-corrected and normalized intensities of one control and one treated array. They were subsequently compared between the four experiments (Fig. 3A). For this purpose, the CV of the fold-change value for each of the 1125 genes was calculated across the four replicates and graphed as a function of the gene's expression

Table 1. CV of fold-change values as a function of the expression level (E_k value)[a]

Classes of E_k values	Number of genes	Average CV	95% of CVs between
Up to 400	104	0.20	0.05–0.63
400–1000	391	0.12	0.03–0.29
1000–2000	299	0.10	0.02–0.20
2000–5000	185	0.10	0.03–0.21
>5000	146	0.09	0.03–0.22
All genes	1125	0.11	0.03–0.29

[a]The CVs were calculated across four replicates (Fig. 3). Genes were grouped into five different classes according to the average E_k value on the four control arrays. The range that includes 95% of the CV values of a certain class was determined by calculating the 2.5th and 97.5th percentiles on all CVs within this class.

level (Fig. 3B). As expected, the CV was highest for genes expressed at low levels (low E_k values) and decreased with rising expression levels (high E_k values). Table 1 shows the average CV for each of five different classes of 1125 genes classified according to their expression level. With the exception of genes expressed at very low levels ($E_k < 400$), the average CV value remains below 0.2 throughout all classes and even drops below 0.1 for highly expressed genes (Table 1). The probe sets for the three prokaryotic spike-in controls (amp^r, kan^r, cm^r; see Dynamic range and discrimination power) produce E_k values of ~350 in the absence of these transcripts. E_k values below 400 therefore indicate genes expressed at very low levels or not at all. As shown in Figure 3B and in Table 1, the distribution of CV values within a class is considerably wider for classes with genes expressed at low levels and narrower for classes including highly expressed genes. For genes with E_k values below 400, for instance, 95% of CVs fall into the range between 0.05 (2.5th percentile) and 0.63 (97.5th percentile), whereas for genes with an E_k level above 5000, the 95% range of the CVs is 0.03 and 0.22. The wider distribution together with the higher average CV render fold-change values for genes expressed at low levels less reliable than those of genes expressed at high levels. This limitation is shared by most if not all array platforms and is also documented for *in situ* synthesized 24mer arrays (38) and the Affymetrix GeneChip arrays (39). The average CV calculated for all 1125 genes irrespective of the expression levels is 0.11. It is worth noting, however, that this value is strongly influenced by the selection of genes. Adding more highly expressed genes would lower this value. On the contrary, a biased selection of genes expressed at low level would lead to a considerably higher CV. The selection of genes included in our study was unbiased and spans the entire expression range (Table 1). Thus, the average value of 0.11 presented in this study is likely to reflect the level of reproducibility encountered in typical gene expression profiling experiments on geniom arrays. In summary, our study revealed CV values that suggest a high reproducibility of fold-change values and compare favorably to data from spotted 35mer arrays where an average CV for the fold-change values of ~0.3 was found (13). In addition, the CV values found on the geniom platform are significantly lower than those obtained with 24mer arrays synthesized on microscopic slides using a maskless photolithographic instrument. For these arrays, average CVs of the fold-change data typically are in

Table 2. CVs of expression levels as a function of the expression level[a]

Classes of E_k values	Control			Treated		
	Number of genes	Average CV	95% of CVs between	Number of genes	Average CV	95% of CVs between
<400	104	0.17	0.04–0.38	128	0.15	0.05–0.34
400–1000	391	0.13	0.04–0.30	383	0.12	0.03–0.25
1000–2000	299	0.13	0.05–0.25	274	0.11	0.04–0.21
2000–5000	185	0.14	0.06–0.24	167	0.12	0.06–0.20
>5000	146	0.12	0.05–0.20	173	0.10	0.04–0.18
All genes	1125	0.14	0.05–0.28	1125	0.12	0.04–0.26

[a]The CVs were calculated across four arrays hybridized with aliquots of a control sample (control) and another four arrays hybridized with aliquots of a sample from yeast cells which were harvested after an osmotic shock (treated). The genes were grouped into five different classes according to their mean expression level on the four arrays. The range that includes 95% of the CV values of a certain class was determined by calculating the 2.5th and 97.5th percentiles on all CVs within this class.

the range between 0.45 (average for low expressed genes) and 0.29 (average for highly expressed genes) (38). Besides the fold-change data, the gene expression level is the most important result of a gene expression profiling experiment. This is particularly true for experiments which determine relative mRNA levels within a single sample rather than comparing two or more samples. In the experiment described above, four arrays were hybridized with aliquots of a yeast control sample and another four arrays were hybridized with aliquots of a sample from yeast cells treated with an osmotic shock. In order to investigate the reproducibility of expression levels (E_k values) obtained with our platform, we calculated CVs of the E_k values across the four replicates hybridized with the same sample for each of the 1125 genes included in the experiment. In agreement with results of a recent study performed on Affymetrix GeneChips (40), the CV of the expression levels was higher for genes expressed at low levels (low E_k values) and lower for genes expressed at high levels (high E_k values). As described above, we grouped the genes into five different classes according to their expression level. The average CV values calculated for these classes were in the 0.17–0.10 range. As shown in Table 2, a trend towards higher CVs for genes expressed at low levels and towards lower CVs for highly expressed genes is evident in the arrays hybridized with the control sample as well as in the arrays hybridized with the treated sample. This is a remarkable finding because the same gene may have different E_k levels on the 'control' and the 'treated' array: the genes that make up a certain expression class are not necessarily the same for the control and the treated sample. We therefore conclude that the high variability found for genes expressed at low levels is indeed due to technical parameters and is only slightly influenced by the individual genes analyzed.

Accuracy of biological results

In an attempt to validate the accuracy of results from the geniom platform we have analyzed the transcriptional response of yeast to osmotic shock. The data acquired with the geniom platform were compared with data from cDNA arrays published by Rep *et al.* (25) and to reference data from Affymetrix GeneChips which were generated as described in the experimental protocol. Our study comprised 4857 genes which were all analyzed twice on standard gene expression arrays containing 10 PM probes per gene (25mers; without MM controls). Using the same type of arrays we also measured an additional group of 203 genes in 10 replicates.

These 203 genes were found to be involved in the cellular response of yeast to osmotic shock in the experiments on spotted cDNA arrays published by Rep *et al.* (25). This selection of genes thus is biased with respect to the expected fold-change values and is also likely to be biased with respect to the expected expression level. However, since we were interested in the accuracy of biological results obtained from geniom arrays and the regulation of these 203 genes is known to be the major response of yeast cells to osmotic shock, we first focused the data analysis on these particular genes before extending it to the total of 4857 genes. Figure 4 shows fold-change values for these genes compared pair-wise between geniom arrays, Affymetrix GeneChips and the cDNA arrays used by Rep *et al*. (25). As indicated by a Pearson correlation coefficient of 0.914 and a Spearman rank correlation coefficient of 0.889, a high conformity was found between the geniom data and the GeneChip data despite comparing two completely independent array platforms (Fig. 4A). Note that the only parameter kept constant on both platforms was the biological sample. When reducing the complexity by applying the same analysis algorithm to both the raw intensity values from the geniom arrays and the raw data from the GeneChips (as found in the .cel file) an even higher similarity was found and the Pearson correlation coefficient increased to 0.959 (Fig. 4B). For further analysis, we again focused on the comparison of independent platforms (Fig. 4A, C and D) grouping the genes into three different categories. Genes with fold-change values $\geqslant$1.5 ($\log_2$ value: 0.58) were considered to be upregulated. Genes with fold-change values $\leqslant$–1.5 ($\log_2$ value: –0.58) were considered to be downregulated and genes with fold-change values between –1.5 and 1.5 ($\log_2$ value: –0.58 to 0.58) were considered to be unaffected. Based on this categorization, 184 out of 203 genes showed the same tendency on Affymetrix and geniom arrays (142 upregulated, 30 downregulated, 12 unchanged). From the remaining genes, nine were found to be regulated on the Affymetrix GeneChip but unaffected on the geniom arrays and nine genes behaved vice versa. Only one gene switched between the upregulated and the downregulated categories. As indicated by the correlation coefficients, the geniom data closely match the GeneChip data. In addition, they are very similar to the data obtained with cDNA arrays. A total of 174 out of 203 genes showed the same tendency in the geniom and the cDNA data set. A minority of 21 genes switched between unchanged on the geniom arrays and regulated on the cDNA arrays, one gene vice versa, and seven genes were found to be regulated in the

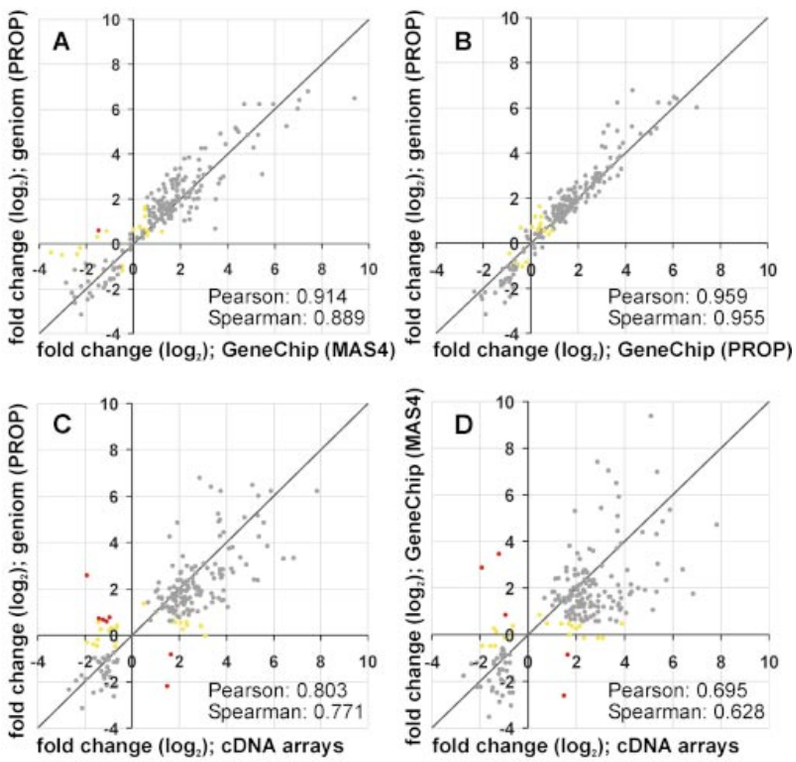


Figure 4. Log–log plots comparing fold-change data from three different array formats. The transcriptional response of 203 yeast genes to osmotic shock was analyzed on the geniom platform in 10 replicates. The average fold-change values were compared with data from Affymetrix GeneChips and to cDNA array data from the literature (25) (Supplementary Material Table 2B). Genes that fall into the same category of regulation on the respective platforms are shown in gray (cut-offs for categorization: fold-change $\leqslant$–0.58, downregulated; fold-change $>$–0.58 but $<$0.58, unchanged; fold-change $\geqslant$0.58, upregulated). Genes that were found to be up- or downregulated on one platform but unchanged on the other are shown in yellow. Genes that behave the opposite way are shown in red. (**A**) Comparison of geniom data and MAS4 calculated fold-change values from the Affymetrix GeneChips. (**B**) Comparison of geniom data and PROP-calculated fold-change values from the Affymetrix GeneChips. (**C**) Comparison of geniom data and the cDNA array data from the literature (25). (**D**) Comparison of Affymetrix GeneChip data and cDNA array data from the literature (25).

opposite sense on both platforms. The conformity between geniom data and GeneChip data, however, is greater than the similarity found between the cDNA data and either of the oligonucleotide arrays (Fig. 4). In general, most genes showed the same tendency on the spotted cDNA arrays and on both oligonucleotide array formats. Thus, the major findings described by Rep *et al*. (25) could be reproduced on geniom arrays (Supplementary Material Table 2B). The actual fold-change values, however, differ significantly between the cDNA arrays and the oligonucleotide arrays. This is in good agreement with studies that revealed substantial differences in the overall performance of cDNA arrays and oligonucleotide arrays. Generally, spotted cDNA arrays show a higher sensitivity than short oligonucleotide arrays (19,41). Conversely, spotted cDNA arrays are known to exhibit lower specificity than short oligonucleotide arrays, primarily because of cross-hybridization of highly homologous transcripts and non-cognate cDNA probes and due to varying hybridization efficiencies of long cDNA probes (42–45). An additional factor that might contribute to the variance in the fold-change values observed in our study is the biological sample itself. The cDNA data were taken from the literature. Therefore, the total RNA source used for the experiments on the cDNA arrays was not identical to that used for the geniom and the Affymetrix oligonucleotide arrays. A recently published, extensive study designed as an interlaboratory comparison revealed that variations introduced by *in vitro* handling steps and variations between replicate cultures in particular can significantly influence the result of a gene expression experiment (46). In addition, the labeling procedures differ significantly: the oligonucleotide arrays used in this study were hybridized to an amplified biotinylated cRNA sample (synthesized starting from the total RNA, as described in Materials and Methods) while the cDNA arrays used by Rep *et al.* (25) were hybridized with a non amplified, [^{33}P]CTP-labeled cDNA sample (synthesized from the total RNA via reverse transcription). Taken together, the first part of our study focusing on the 203 genes known to be regulated in the cellular response of yeast to osmotic shock suggests a high conformity of biological data obtained on geniom arrays and data aquired on Affymetrix GeneChips. We also found that the great majority of the 203 genes (86% when applying the categorization criteria described above) showed the same tendency on geniom arrays and spotted cDNA arrays. The significant variation of the actual fold-change values found in the latter comparison is likely to be caused by differences in

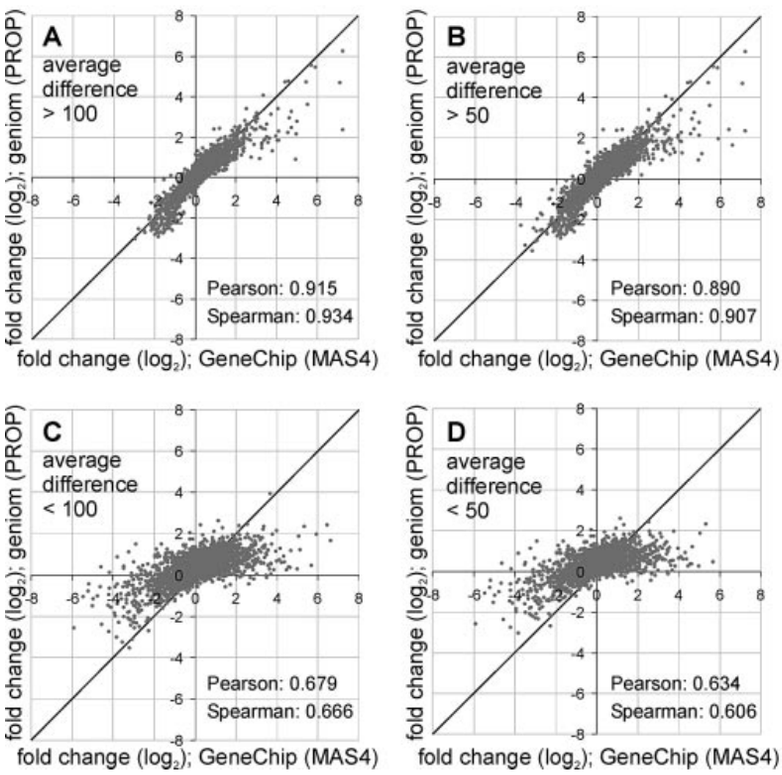


Figure 5. Log–log plots comparing fold-change data from geniom arrays and Affymetrix GeneChips. The transcriptional response of 4857 randomly chosen yeast genes to osmotic shock was analyzed on geniom arrays in two replicates and the average fold-change values were compared with data from Affymetrix GeneChips (Supplementary Material Table 2C). For this comparison, the genes were grouped into 'expressed at low level' and 'expressed at a higher level' according to their expression level (average difference) on the GeneChip base array. Cut-offs at either 100 or 50 were used for the categorization. (**A**) Log–log plot of the 1688 genes with an average difference above a cut-off at 100. (**B**) Log–log plot of the 2596 genes with an average difference above a cut-off at 50. (**C**) Log–log plot of the 3169 genes with an average difference below 100. (**D**) Log–log plot of the 2261 genes with an average difference below 50.

the general performance of the two array formats and by differences in the samples used for the experiments on the respective platforms.

So far, we have restricted our analysis to the 203 genes known to be involved in the cellular response of yeast to osmotic shock. Most of these genes are highly regulated and tend to be expressed at higher levels. They are therefore much more likely to show the same trends on different platforms than randomly selected genes. In order to investigate if the high concordance of geniom and GeneChip data is confirmed in experiments with a completely unbiased selection of genes, we extended the analysis to all 4857 genes included in our study. We compared the average fold-change values calculated on the two replicate experiments performed on the geniom instrument to the fold-change values obtained from the Affymetrix GeneChips. 3276 (68%) out of 4857 genes fell into the same category; 1076 genes (22%) were unchanged on the febit arrays but downregulated or upregulated on the GeneChips; 436 genes (9%) were unchanged on the Affymetrix GeneChips but regulated on the geniom arrays and 69 genes (1%) were found to be regulated in the opposite sense on both platforms (Supplementary Material Table 2C). Overall, a mean Pearson correlation coefficient of 0.742 and an average Spearman rank correlation coefficient of 0.759 were calculated on the fold-change level. Taken together, these data indicate a considerably lower agreement of the fold-change values for the 4857 randomly selected genes than for the 203 genes from a biased selection. To address the question of whether the poor conformity applies to all 4857 genes analyzed or if it is restricted to a certain subgroup of genes, we refined our analysis taking the expression levels into account. Mills and Gordon (39) investigated false-positive rates using Affymetrix Mu11KsubA and Mu11KsubB GeneChips. All genes recognized as increased or decreased in same-to-same comparisons were defined as noise. Most of these genes were clustered at expression levels below 250 (measured by the average difference between PM and MM of all PM/MM probe pairs for one transcript). Grundschober *et al.* (40) used GeneChip U34 and estimated CVs of triplicate hybridizations to determine significant fold-changes thresholds. They found the fold-change value to be reliable above a cut-off expression level of 100. We applied this 100 cut-off as well as a less stringent cut-off at 50 to our analysis. We classified the 4857 genes according to their average difference (expression level) on the GeneChip array (base array) into 'expressed at low level' (below the respective cutoff) and 'expressed at a higher level' (above the respective cut-off). Then, we analyzed the agreement of the fold-change values obtained with the GeneChips and the fold-change values acquired from geniom arrays within these groups. As shown in Figure 5A and B, we found a substantial correlation between GeneChip data and geniom data for genes expressed at an

elevated level: for genes with an expression level above 100, a Pearson correlation coefficient of 0.915 was calculated and, applying the same categorization criteria as used for the 203 genes, 83% of all genes (1401 out of 1688) showed the same tendency on both platforms (Fig. 5A). A slightly lower but still significant conformity was found for genes with an expression level above the less stringent cut-off at 50: for these genes, the Pearson correlation coefficient was 0.890, and 80% of all genes (2067 out of 2596) showed the same tendency of regulation (Fig. 5B). In contrast, we found only poor correlations of fold-change values for the genes with an expression level lower than the respective cut-offs: for genes with an expression level lower than 100 (3169 genes) or lower than 50 (2261 genes) we observed a Pearson correlation coefficient of 0.679 and 0.634, respectively (Fig. 5C and D).

From these data we conclude that—at least for yeast—fold-change values obtained from geniom arrays are in good concordance with fold-change values acquired with Affymetrix GeneChips (with the exception of genes expressed at very low levels). This is a remarkable finding if the context of the experimental design is considered. The only parameter kept constant between the two platforms was the biological sample. All other parameters, including the probe design and the algorithm used for data analysis, were different for both platforms. Despite this high correlation found for genes expressed at elevated levels, our comparison also revealed substantial differences in the fold-change values obtained with both platforms with regard to genes expressed at low levels. This finding was not unexpected and is likely to be caused by a higher variation of fold-change values calculated on low signal intensities. The fact that calculations based on such low signal intensities are prone to increased variation is known for most if not all array formats, including spotted 35mer arrays (13), *in situ* synthesized 24mer arrays (38) and GeneChips (39,40)—and was also found for the geniom platform in this study.

We further demonstrated that geniom data not only match data acquired with other array formats but also reflect the true gene expression pattern of the biological system analyzed. We used a non-array reference system and compared the gene expression data from the geniom platform with those obtained by quantitative RT–PCR (SYBR Green assay). For this experiment, a subset of 56 genes from the 203 genes shown in Figure 4 was selected. The choice was based on the fold-change distribution in the array experiments, such that the validated data set spans the entire range of fold-change values observed. The selection was otherwise unbiased and random. The quantitative RT–PCR analysis was performed with the same RNA samples used for the array experiments. Seven out of the 56 genes were excluded from the analysis due to PCR efficiencies below 1.70. Table 3 compares the fold-change values of the evaluable genes to the average fold-change values from the 10 replicate experiments on the geniom platform described above (Fig. 4). As indicated by the Pearson correlation coefficient of 0.966 and the Spearman rank correlation coefficient of 0.972, a very high conformity was found between the two data sets. Due to the lower dynamic range of oligonucleotide arrays as compared with quantitative RT–PCR, the fold-change values for highly regulated genes are compressed on the geniom platform. This phenomenon has been described before for other spotted (13) or *in situ* synthesized (38) oligonucleotide arrays. Despite those differences in the fold-change values of highly regulated genes, our study provided evidence that geniom arrays generate accurate and reliable results and thus enable scientists to address complex biological questions.

DISCUSSION

This study was designed to validate the geniom technology, a novel and fully integrated oligonucleotide array platform for gene expression profiling applications. We first focused on the technical aspects and evaluated the discrimination power, the dynamic range, and the reproducibility of the system. The system is able to detect RNAs present at a frequency of 1:100 000. In good agreement with data published for other oligonucleotide array platforms (13,35,36), detection is quantitative over more than two orders of magnitude. The geniom technology integrates array synthesis, hybridization and detection in a single benchtop device located in the investigator's laboratory. As quality assurance is a more demanding issue for benchtop instruments compared with centralized facilities, special attention was paid to data reproducibility. Primary experiments on four identical arrays with 6398 features each revealed a mean CV value of 0.09 for the non-processed raw intensities with an unbiased distribution across the entire intensity range. In a more elaborate experiment targeting 1125 randomly chosen yeast genes, we found the CV for the fold-change values to be substantially influenced by the expression level. The average CV values range between 0.20 for genes expressed at very low levels and 0.09 for genes expressed at high levels. The CVs for the expression levels range between 0.19 (average for genes expressed at very low levels) and 0.10 (average for genes expressed at high levels). Taken together, the CV values indicate a good reproducibility of raw data, fold-change values and expression levels but also revealed that expression results for genes expressed at low level are considerably less consistent than those of genes expressed at higher levels. This phenomenon is common to most if not all array platforms and is known for the widely used GeneChip arrays (39,40), *in situ* synthesized 24mer arrays (38) and spotted 35mer arrays (13). By extending our study from inter-array to inter-instrument comparisons we demonstrated that different individual geniom instruments perform equally well. For all four instruments included in our study, the mean CVs for the fold-change values (mean value across the entire expression range) were in the range of 0.11–0.18 (data not shown). As a next step, the accuracy of biological data was demonstrated by comparing the geniom data from a real-world experiment to reference data obtained from Affymetrix GeneChips, data from quantitative RT–PCR and cDNA array data from the literature (25). In this experiment, we were able to reproduce the major findings of Rep *et al.*, who investigated the transcriptional response of yeast to osmotic shock in great detail on cDNA arrays and generated a list of 203 genes which they identified as the main responders to the osmotic shock treatment (25). Despite substantial differences in the actual fold-change values, the great majority of the 203 genes showed the same tendency of regulation on the geniom oligonucleotide arrays. By comparing the geniom data for these genes to reference data acquired on Affymetrix GeneChips we found a high conformity of

Table 3. Comparison of fold-change data from geniom arrays and quantitative RT–PCR[a]

Gene	Geniom arrays		Quantitative RT–PCR	
	Average fold-change	Average fold-change (log_2)	Average fold-change	Average fold-change (log_2)
YMR175W	111.05	6.80	164.81	7.36
YBR117C	85.37	6.42	2112.88	11.04
YER150W	34.41	5.10	76.91	6.27
YDL223C	29.00	4.86	42.23	5.40
YAL061W	21.70	4.44	23.62	4.56
YDL204W	19.18	4.26	33.15	5.05
YGR248W	19.12	4.26	2.04	4.46
YKL151C	13.54	3.76	10.67	3.42
YHR087W	9.88	3.30	33.58	5.07
YGR066C	8.57	3.10	16.75	4.07
YML054C	7.25	2.86	4.39	2.14
YHR094C	6.44	2.69	8.31	3.05
YML100W	6.35	2.67	7.08	2.82
YLR267W	5.12	2.36	3.06	1.61
YER103W	4.91	2.30	4.32	2.11
YKL150W	4.21	2.07	5.48	2.45
YHR022C	4.02	2.01	5.63	2.49
YLR031W	3.98	1.99	4.36	2.12
YEL039C	3.71	1.89	2.96	1.56
YMR031C	3.06	1.61	1.50	0.59
YCL040W	2.91	1.54	4.25	2.09
YER054C	2.62	1.39	3.18	1.67
YDR533C	2.31	1.21	2.06	1.04
YGR170W	2.27	1.18	1.69	0.76
YJL149W	2.22	1.15	3.30	1.72
YDR100W	2.09	1.06	1.84	0.88
YDR463W	2.08	1.06	1.50	0.58
YLR042C	2.06	1.04	2.13	1.09
YER041W	1.72	0.78	–1.05	–0.07
YGR146C	1.54	0.63	1.46	0.55
YMR030W	1.45	0.54	2.11	1.08
YHR086W	1.30	0.38	–1.09	–0.12
YDL135C	1.15	0.21	–1.93	–0.95
YKL160W	–1.17	–0.22	–2.56	–1.35
YBL002W	–1.18	–0.24	–1.53	–0.61
YGR138C	–1.24	–0.31	–2.64	–1.40
YDR324C	–1.39	–0.48	–10.36	–3.37
YKL109W	–1.84	–0.88	–4.99	–2.32
YER165W	–2.19	–1.13	–7.59	–2.92
YGR155W	–3.04	–1.61	–7.15	–2.84
YGL055W	–3.29	–1.72	–10.69	–3.42
YDL198C	–3.32	–1.73	–37.42	–5.23
YHR128W	–3.68	–1.88	–12.13	–3.60
YJL217W	–3.92	–1.97	–7.19	–2.85
YDL014W	–4.47	–2.16	–22.89	–4.52
YGR060W	–4.55	–2.19	–16.32	–4.03
YKR013W	–4.60	–2.20	–1.96	–0.97
YER052C	–6.24	–2.64	–21.92	–4.45
YGR234W	–8.91	–3.16	–33.48	–5.07

[a]Values in the geniom columns represent averages from 10 identical experiments (Fig. 4). All fold-change values can be found in Supplementary Material Table 2B.

fold-change data. A larger experiment comprising 4857 yeast genes from a random selection, confirmed the high correlation of geniom data and Affymetrix data. Despite a high correlation of fold-change data for highly expressed genes, however, substantial differences in the fold-change values were evident for genes in the low expression level. This was not an unexpected finding and is in good agreement with a higher variation of fold-change data found for genes expressed at low levels on both the Affymetrix GeneChips (39) and the geniom arrays. In an attempt to demonstrate that geniom data not only match data obtained from other array formats but also reflect the gene expression pattern of the biological system analyzed, we used quantitative real-time PCR to measure the fold-change of 56 yeast genes that span the entire expression range. Due to the lower dynamic range of geniom arrays as compared with real-time PCR we observed some differences in the fold-change values of highly regulated genes, reflecting the compression of geniom data in the high-intensity range. Nevertheless, a Pearson correlation coefficient of 0.966 clearly indicated a high concordance between the geniom data and the data obtained by quantitative real-time PCR. In conclusion, our data suggest that the geniom technology

produces reproducible and reliable results and complements other well established array platforms. Due to its design, however, it provides a number of new opportunities. The sequence of each individual probe may be varied on each array and all that is required to generate a new array is sequence information. Sequence updates or results from a previously performed array experiment can be incorporated into new array designs. The automation ensures convenient handling of the machine and thus may contribute to a more widespread use of the complex array technologies.

In this study, we have validated the geniom platform for gene expression profiling experiments. Supported by the small reaction volumes and the design of the arrays as three-dimensional microchannels, however, the system is also well suited for other applications involving enzymatic reactions such as primer extension, ligation or on-chip PCR.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

Part of the work presented here was supported by a grant from the German BMBF (Bundesministerium für Bildung und Forschung, Deutschland) to febit ag.

REFERENCES

1. Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
2. Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. and Brown,E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
3. Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. [Erratum: *Science*, **282**, 5393] *Science*, **282**, 699–705.
4. Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D., Kidd,M.J., King,A.M., Meyer,M.R., Slade,D., Lum,P.Y., Stepaniants,S.B., Shoemaker,D.D., Gachotte,D., Chakraburtty,K., Simon,J., Bard,M. and Friend,S.H. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
5. Gray,N.S., Wodicka,L., Thunnissen,A.M., Norman,T.C., Kwon,S., Espinoza,F.H., Morgan,D.O., Barnes,G., LeClerc,S., Meijer,L., Kim,S.H., Lockhart,D.J. and Schultz,P.G. (1998) Exploiting chemical libraries, structure and genomics in the search for kinase inhibitors. *Science*, **218**, 533–538.
6. Marton,M.J., DeRisi,J.L., Bennett,H.A., Iyer,V.R., Meyer,M.R., Roberts,C.J., Stoughton,R., Burchard,J., Slade,D., Dai,H., Bassett,D.E.,Jr, Hartwell,L.H., Brown,P.O. and Friend,S.H. (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.*, **4**, 1293–1301.
7. Roberts,C.J., Nelson,B., Marton,M.J., Stoughton,R., Meyer,M.R., Bennett,H.A., He,Y.D., Dai,H., Walker,W.L., Hughes,T.R., Tyers,M., Boone,C. and Friend,S.H. (2000) Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science*, **287**, 873–880.
8. Khan,J., Simon,R., Bittner,M., Chen,Y., Leighton,S.B., Pohida,T., Smith,P.D., Jiang,Y., Gooden,G.C., Trent,J.M. and Meltzer,P.S. (1998) Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.*, **58**, 5009–5013.
9. Perou,C.M., Jeffrey,S.S., van de Rijn,M., Rees,C.A., Eisen,M.B., Ross,D.T., Pergamenschikov,A., Williams,C.F., Zhu,S.X., Lee,J.C., Lashkari,D., Shalon,D., Brown,P.O. and Botstein,D. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA*, **96**, 9212–9217.
10. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
11. Shoemaker,D.D., Schadt,E.E., Armour,C.D., He,Y.D., Garrett-Engele,P., McDonagh,P.D., Loerch,P.M., Leonardson,A., Lum,P.Y., Cavet,G., Wu,L.F., Altschuler,S.J., Edwards,S., King,J., Tsang,J.S., Schimmack,G., Schelter,J.M., Koch,J., Ziman,M., Marton,M.J., Li,B., Cundiff,P., Ward,T., Castle,J., Krolewski,M., Meyer,M.R., Mao,M., Burchard,J., Kidd,M.J., Dai,H., Phillips,J.W., Linsley,P.S., Stoughton,R., Scherer,S. and Boguski,M.S. (2001) Experimental annotation of the human genome using microarray technology. *Nature*, **409**, 922–927.
12. Hu,G.K., Madore,S.J., Moldover,B., Jatkoe,T., Balaban,D., Thomas,J. and Wang,Y. (2001) Predicting splice variants from DNA chip expression data. *Genome Res.*, **11**, 1237–1245.
13. Ramakrishnan,R., Dorris,D., Lublinsky,A., Nguyen,A., Domanus,M., Prokhorova,A., Gieser,L., Touma,E., Lockner,R., Tata,M., Zhu,X., Patterson,M., Shippy,R., Sendera,T.J. and Mazumder,A. (2002) An assessment of Motorola CodeLink™ microarray performance for gene expression profiling applications. *Nucleic Acids Res.*, **30**, e30.
14. Yue,H., Eastman,P.S., Wang,B.B., Minor,J., Doctolero,M.H., Nuttall,R.L., Stack,R., Becker,J.W., Montgomery,J.R., Vainer,M. and Johnston,R. (2001) An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Res.*, **29**, e41.
15. Guckenberger,M., Kurz,S., Aepinus,C., Theiss,S., Haller,S., Leimbach,T., Panzner,U., Weber,J., Paul,H., Unkmeir,A., Frosch,M. and Dietrich,G. (2002) Analysis of the heat shock response of *Neisseria meningitidis* with cDNA- and oligonucleotide-based DNA microarrays. *J. Bacteriol.*, **184**, 2546–2551.
16. Southern,E.M., Maskos,U. and Elder,J.K. (1992) Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. *Genomics*, **13**, 1008–1017.
17. Maskos,U. and Southern,E.M. (1992) Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesized *in situ*. *Nucleic Acids Res.*, **20**, 1679–1684.
18. Fodor,S.P., Read,J.L., Pirrung,M.C., Stryer,L., Lu,A.T. and Solas,D. (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science*, **251**, 767–773.
19. Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R., Kobayashi,S., Davis,C., Dai,H., He,Y.D., Stephaniants,S.B., Cavet,G., Walker,W.L., West,A., Coffey,E., Shoemaker,D.D., Stoughton,R., Blanchard,A.P., Friend,S.H. and Linsley,P.S. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.*, **19**, 342–347.
20. Singh-Gasson,S., Green,R.D., Yue,Y., Nelson,C., Blattner,F., Sussman,M.R. and Cerrina,F. (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat. Biotechnol.*, **17**, 974–978.
21. Pellois,J.P., Zhou,X., Srivannavit,O., Zhou,T., Gulari,E. and Gao,X. (2002) Individually addressable parallel peptide synthesis on microchips. *Nat. Biotechnol.*, **20**, 922–926.
22. Beier,M. and Hoheisel,J.D. (2000) Production by quantitative photolithographic synthesis of individually quality-checked DNA microarrays. *Nucleic Acids Res.*, **28**, e11.
23. Hasan,A. Stengele,K.-P., Giegrich,H., Cornwell,P., Isham,K.R., Sachleben,R.A., Pfleiderer,W. and Foote,S. (1997) Photolabile protecting groups for nucleotides: synthesis and photodeprotection rates. *Tetrahedron*, **53**, 4247–4264.
24. Stähler,C.F., Stähler,P.F., Müller,M., Stähler,F. and Lindner,H. (1999) Patent DE-19940750.9-52; PCT/WO/EP/99/0617; AU-749884B2.
25. Rep,M., Krantz,M., Thevelein,J.M. and Hohmann,S. (2000) The transcriptional response of *Saccharomyces cerevisiae* to osmotic shock. *J. Biol. Chem.*, **275**, 8290–8300.
26. Schmitt,M.E., Brown,T.A. and Trumpower,B.L. (1990) A rapid and simple method for preparation of RNA from *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **18**, 3091–3092.

27. Van Gelder,R.N., von Zastrow,M.E., Yool,A., Dement,W.C., Barchas,J.D., Eberwine,J.H. (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Science*, **87**, 1663–1667.
28. Eberwine,J., Yeh,H., Miyashiro,K., Cao,Y., Nair,S., Finnell,R., Zettel,M. and Coleman,P. (1992) Analysis of gene expression in single live neurons. *Proc. Natl Acad. Sci. USA*, **89**, 3010–3014.
29. Zhou,Y. and Abagyan,R. (2002) Match-only integral distribution (MOID) algorithm for high-density oligonucleotide array analysis. *BMC Bioinformatics*, **3**, 3.
30. Rasmussen,R. (2001) Quantification on the lightCycler. In Meuer,S., Wittwer,C. and Nakagawara,K. (eds), *Rapid Cycle Real-time PCR, Methods and Applications*. Springer Press, Heidelberg, pp. 21–34.
31. Muller,P.Y., Janovjak,H., Miserez,A.R. and Dobbie,Z. (2002) Processing of gene expression data generated by quantitative real-time RT-PCR. [Erratum: *Biotechniques*, **33**, 514] *Biotechniques*, **32**, 1372–1374, 1376, 1378–1379.
32. Iyer,V. and Struhl,K. (1996) Absolute mRNA levels and transcriptional initiation rates in *Saccharomyces cerevisiae. Proc. Natl Acad. Sci. USA*, **93**, 5208–5212.
33. Hereford,L.M. and Rosbash,M. (1997) Number and distribution of polyadenylated RNA sequences in yeast. *Cell*, **10**, 453–462.
34. Lewin,B. (1980) *Gene Expression.* Wiley-Interscience, New York, NY, Vol. 2.
35. Albert,T.J., Norton,J., Ott,M., Richmond,T., Nuwaysir,K., Nuwaysir,E.F., Stengele,K.P. and Green,R.D. (2003) Light-directed 5′→3′ synthesis of complex oligonucleotide microarrays. *Nucleic Acids Res.*, **31**, e35.
36. Chudin,E., Walker,R., Kosaka,A., Wu,S.X., Rabert,D., Chang,T.K. and Kreder,D.E. (2002) Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biol.*, **3**, research0005.1–research0005.10.
37. Machl,A.W., Schaab,C. and Ivanov,I. (2002) Improving DNA array data quality by minimising 'neighbourhood' effects. *Nucleic Acids Res.*, **30**, e127.
38. Nuwaysir,E.F., Huang,W., Albert,T.J., Singh,J., Nuwaysir,K., Pitas,A., Richmond,T., Gorski,T., Berg,J.P., Ballin,J., McCormick,M., Norton,J., Pollock,T., Sumwalt,T., Butcher,L., Porter,D., Molla,M., Hall,C., Blattner,F., Sussman,M.R., Wallace,R.L., Cerrina,F. and Green,R.D. (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res.*, **12**, 1749–1755.
39. Mills,J.C. and Gordon,J.I. (2001) A new approach for filtering noise from high-density oligonucleotide microarray datasets. *Nucleic Acids Res.*, **29**, e72.
40. Grundschober,C., Malosio,M.L., Astolfi,L., Giordano,T., Nef,P. and Meldolesi,J. (2002) Neurosecretion competence. A comprehensive gene expression program identified in PC12 cells. *J. Biol. Chem.*, **277**, 36715–36724.
41. Schulze,A. and Downward,J. (2001) Navigation gene expression using microarrays—a technology review. *Nature Cell Biol.*, **3**, E190–E195.
42. Li,J., Pankratz,M. and Johnson,J.A. (2002) Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays. *Toxicol. Sci.*, **69**, 383–390.
43. Bartosiewicz,M., Trounstine,M., Barker,D., Johnston,R. and Buckpitt,A. (2000) Development of a toxicological gene array and quantitative assessment of this technology. *Arch. Biochem. Biophys.*, **376**, 66–73.
44. Heller,R.A., Schena,M., Chai,A., Shalon,D., Bedilion,T., Gilmore,J., Woolley,D.E. and Davis,R.W. (1997) Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl Acad. Sci. USA*, **94**, 2150–2155.
45. Richmond,C.S., Glasner,J.D., Mau,R., Jin,H. and Blattner,F.R. (1999) Genome-wide expression profiling in *Escherichia coli* K-12. *Nucleic Acids Res.*, **27**, 3821–3835.
46. Piper,M.D.W., Daran-Lapujade,P., Bro,C., Regenberg,B., Knudsen,S., Nielsen,J. and Pronk,J.T. (2002) Reproducibility of oligonucleotide microarray transcriptome analyses. An interlaboratory comparison using chemostat cultures of *Saccharomyces cerevisiae. J. Biol. Chem.*, **277**, 37001–37008.

Methods

Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing

Daniel Summerer,[1,4] Haiguo Wu,[2] Bettina Haase,[1] Yang Cheng,[1] Nadine Schracke,[1] Cord F. Stähler,[1] Mark S. Chee,[3] Peer F. Stähler,[1] and Markus Beier[1]

[1]*febit biomed gmbh, 69120 Heidelberg, Germany;* [2]*febit Inc., Lexington, Massachusetts 02421, USA;* [3]*Prognosys Biosciences Inc., La Jolla, California 92037, USA*

The lack of efficient high-throughput methods for enrichment of specific sequences from genomic DNA represents a key bottleneck in exploiting the enormous potential of next-generation sequencers. Such methods would allow for a systematic and targeted analysis of relevant genomic regions. Recent studies reported sequence enrichment using a hybridization step to specific DNA capture probes as a possible solution to the problem. However, so far no method has provided sufficient depths of coverage for reliable base calling over the entire target regions. We report a strategy to multiply the enrichment performance and consequently improve depth and breadth of coverage for desired target sequences by applying two iterative cycles of hybridization with microfluidic Geniom biochips. Using this strategy, we enriched and then sequenced the cancer-related genes *BRCA1* and *TP53* and a set of 1000 individual dbSNP regions of 500 bp using Illumina technology. We achieved overall enrichment factors of up to 1062-fold and average coverage depths of 470-fold. Combined with high coverage uniformity, this resulted in nearly complete consensus coverages with >86% of target region covered at 20-fold or higher. Analysis of SNP calling accuracies after enrichment revealed excellent concordance, with the reference sequence closely mirroring the previously reported performance of Illumina sequencing conducted without sequence enrichment.

[Supplemental material is available online at http://www.genome.org. The sequence data from this study have been submitted to the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession no. SRA009002.]

Next-generation sequencing (NGS) platforms have transformed genetic variation studies by a massive reduction of cost and sequencing effort (Shendure et al. 2004, 2005; Margulies et al. 2005; Bentley 2006; Johnson et al. 2007; Harris et al. 2008). However, this technology advance has not yet been matched by an equal improvement at the front end: the isolation of target DNA sequences for analysis (Garber 2008). Although untargeted sequencing of even whole human genomes has been shown to be feasible, such large projects exceed the current capacity of NGS instruments and are cost prohibitive for the majority of research laboratories (Bentley et al. 2008; Wang et al. 2008). Many future applications would greatly benefit from focusing on specific genomic subsets. This can be the targeted sequencing of components of a single genome such as the whole exome but also fractions of more complex samples, for example, when applied to microbial communities, host–pathogen mixtures, or somatic variants.

Technologies are thus urgently required to selectively isolate genomic sequences at a scale and specificity that cannot easily be met by traditional enrichment approaches like PCR. An ideal enrichment technology for NGS would allow highly multiplexed access to any desired genomic loci. Enrichment thereby has to be uniform and efficient to enable maximal consensus coverage of the target region with sufficient depth for accurate base calling and with minimal sequencing effort. Furthermore, the method should not interfere with accuracy of base calling by causing allelic bias or dropout.

Several recent studies have started to address this bottleneck by using solution- or microarray-based sequence capture relying on hybridization. Two studies using solution-phase sequence capture with padlock or molecular inversion probes have been published that targeted large numbers of small genomic regions in a single reaction. Although the multiplexing level of one of these methods was high, low uniformity of coverage was reported as a serious drawback of both of these approaches (Dahl et al. 2007; Porreca et al. 2007). Still another approach made use of long, biotinylated RNA probes for solution-phase hybridization. However, the overall workflow depended on multistep enzymatic processing of DNA capture probes including PCR and in vitro transcription, possibly introducing bias and errors into the probe library. Moreover, very long hybridization times of several days were applied (Gnirke et al. 2009), which is rather time-consuming even compared with approaches relying on solid-phase hybridization.

Recently, sequence enrichment using solid-phase hybridization to DNA microarrays with flexible content has been described (Albert et al. 2007; Hodges et al. 2007; Okou et al. 2007; Bau et al. 2009). For several projects targeting different regions, enrichment factors of several hundred- to a 1000-fold have been reported, resulting in good depth of coverage for at least a fraction of the target region. However, covering the full target region with the depth sufficient for reliable base calling has emerged as a key challenge (Garber 2008).

In fact, no method has so far been able to reach an enrichment performance that allows for full consensus coverage of a target with satisfactory depth, and before now, it was not clear whether optimization of the most obvious experimental variables such as hybridization stringency, probe design, or blocking conditions would overcome this problem. Given that reported target sizes are typically in the range of kilobases to megabases, the fraction of target sequence in a human DNA sample relative to background is only 3.1×10^{-5}% to 3.1×10^{-2}% for 1 kb and 1 Mb, respectively. This range of concentration presents a serious

[4]Corresponding author.
E-mail daniel.summerer@febit.de; fax +49-6221-6510-390.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.091942.109. Freely available online through the *Genome Research* Open Access option.

19:1616–1621 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

purification challenge, e.g., similar to the most demanding protein purifications. Although the specificity of protein–protein interactions employed in protein purifications (e.g., antibody–antigen interactions or affinity tag binding) can be much higher than the specificity of Watson-Crick base pairing, the application of multiple rounds of chromatography is a standard procedure to obtain target protein of sufficient purity (Coligan et al. 2008).

We transferred this purification strategy to DNA sequence isolation by performing two instead of one cycles of enrichment using microfluidic Geniom biochips before Illumina NGS. We show that for different target sequences enrichment performance dramatically increases from the first to the second cycle, indicating a multiplicative effect. This effect on enrichment performance is accompanied by a significant increase of the percentage of target region being covered. This results in higher enrichment factors than previously reported for sequence capture methods prior to Illumina NGS (Hodges et al. 2007; Gnirke et al. 2009). A comprehensive analysis of SNP calling performance after enrichment shows that the method does not interfere with base-calling accuracy.

Using a microfluidic array platform with integrated hardware thereby results in several advantages. The hybridization steps employed are four times shorter than in other methods, which results in shorter overall process times. Furthermore, the process can be highly automated, which supports improved handling effort, reduces contamination risk, and increases reproducibility.

Results and Discussion

The sequence enrichment technology reported here, called HybSelect, is conducted in three main steps: hybridization, washing, and elution. First, a genomic DNA library is hybridized to a Geniom biochip containing target-specific DNA capture probes. After washing and elution, the sample is subjected to a second cycle of enrichment and analyzed by an NGS platform. Though the process should be applicable to any NGS platform, experiments for this study were analyzed using the Illumina Genome Analyzer II (GAII).

Capture of cancer-related genes

We chose the human genes *BRCA1* and *TP53* as our first targets for enrichment, because of their well-known role in the development of certain cancers.

We designed an array of 50mer DNA oligonucleotide probes with a tiling density of 8 bp. A Geniom biochip is composed of eight individual microfluidic channels, each having a capacity for >15,000 capture probes; we used part of one channel for synthesis of the tiling array. To prevent the enrichment of repetitive elements, we excluded low-complexity probes from the array design, which reduces the region of interest (ROI) of 100 kb to a core region of 54 kb actually covered by capture probes (hybselected region [HR]). This corresponds to a capacity of >1.8 Mb ROI or >1 Mb HR per biochip. Next, we subjected a human Illumina paired-end library to a first round of hybridization on the biochip for 16 h with active mixing of the sample.

Two independent experiments, A and B, were conducted in parallel to test the reproducibility of the process. After four consecutive washing steps, we eluted the samples and amplified them using the Illumina paired-end primers, which afforded sufficient amounts for a second hybridization step. Processing of the enriched samples on an Illumina GAII instrument yielded 8,217,673 and 7,624,181 paired-end reads of 2 × 36 bp for the individual samples. The reads were used for further analysis after homopolymeric and ambiguous sequences were filtered out.

After this first cycle of enrichment, mapping of the reads to the ROI revealed that 61.8% to 88.8% of the HR was covered at least once, exhibiting a similar range to what was previously reported for one cycle of microarray-based sequence enrichment and Illumina sequencing (Table 1). In this study, between 12% and 91% of target sequence were reported to be covered at least once, depending on sequence context and library fragment size (Hodges et al. 2007). The average depth of coverage was between 2.9- and 5.0-fold for all target regions for both experiments (Table 1). Overall, the data suggest similar or better reproducibility than previously reported for microarray-based sequence capture (Albert et al. 2007; Hodges et al. 2007; Okou et al. 2007; Bau et al. 2009). Importantly, analysis of the uniqueness of obtained read pairs revealed that more than 98% for both runs, were unique, which is higher than previously reported for standard Illumina GAII sequencing without any enrichment method (Quail et al. 2008). This clearly shows that no detectable library representation bias has been introduced during the HybSelect process that would compromise the information value of obtained reads.

Impact of a second enrichment cycle on capture performance

We next subjected the enriched sample from experiment A to a hybridization process under the same conditions applied in the first enrichment cycle. Sequencing yielded 7,433,555 paired end reads of 2 × 36 bp that were filtered as described above.

Figure 1 shows a graphic view of the ROI with HR regions and coverage depth distribution of mapped reads from the first and

Table 1. Mapping data of reads obtained from one or two cycles of array-based sequence enrichment of human genomic DNA samples for different target regions and Illumina GAII paired-end sequencing

Experiment[a]	Target	ROI	HR	Reads on HR	Average depth of coverage (fold/base)	Enrichment (fold)	1× consensus (%)	5× consensus (%)	10× consensus (%)	20× consensus (%)
A (cycle 1)	*BRCA1*	81,155	45,498	5265	3.8	22.9	77.3	22.8	5.2	1.5
	TP53	19,179	8178	1131	5.0	27.3	88.8	47.9	8.7	0.9
B (cycle 1)	*BRCA1*	81,155	45,498	4426	2.9	20.5	61.8	8.2	2.2	1.1
	TP53	19,179	8178	737	3.3	19.0	83.3	19.8	2.6	0.8
A (cycle 2)	*BRCA1*	81,155	45,498	74,269	58.1	356.4	96.5	87.3	79.5	68.8
	TP53	19,179	8178	23,109	101.3	616.9	98.5	92.9	89.6	86.2
NA18558	1000 loci	1,498,000	498,000	4,300,087	315.6	713.3	96.9	92.1	87.5	80.4
NA18561	1000 loci	1,498,000	498,000	6,281,911	469.1	1061.9	97.5	93.7	90.5	85.5

[a]First cycle of enrichment for *BRCA1* and *TP53* was conducted in duplicate (Experiments A and B).
(ROI) Region of interest; (HR) hybselected region (see text).

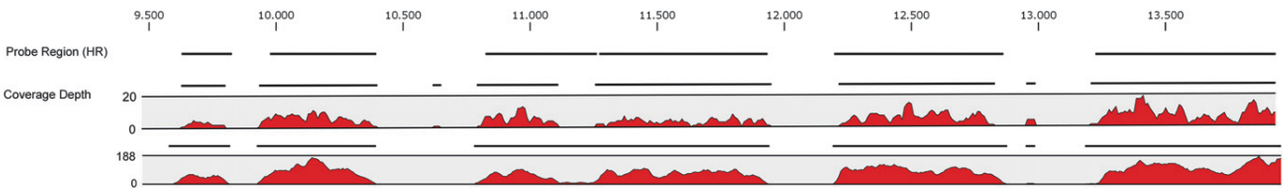


Figure 1. Graphic overview of mapping analysis of an Illumina paired-end sequencing run with a human genomic DNA sample enriched for the genes *BRCA1* and *TP53*. Shown is the capture probe region used for array-based enrichment (black line at *top*), coverage depth distribution obtained from the first enrichment cycle (*middle*), and coverage depth distribution from the second enrichment cycle (*bottom*) to a representative part of the *TP53* gene (nucleotides ~9500–14,000). The obtained consensus sequences are shown as black lines. *X*-axis, the nucleotide position of the gene; *y*-axis, the fold coverage depth. Note that the scale of the *y*-axis varies between the two mappings.

second cycle for a representative region of *TP53*. Reads were obtained almost exclusively in the HR that is covered by capture probes with some overlap to adjacent regions. Moreover, the second cycle experiment strongly increased depth of coverage and apparently also uniformity over the whole region compared with the first cycle of enrichment. Overall, 96.5% and 98.5% of *BRCA1* and *TP53* were covered at least once after this second enrichment cycle (Table 1). The individual enrichment factors (representation of HR sequence in the obtained sequence reads divided by their representation in the human genome) for the two genes obtained from the second cycle were 15.6- and 22.6-fold, respectively, similar to the enrichment factors for the first cycle (22.9- and 27.3-fold), which indicates a multiplicative enrichment effect. This resulted in final enrichment factors for the overall process of 356.4- and 616.9-fold. Interestingly, quantitative analyses suggest that biochips that are reused for the second enrichment cycle result in comparable enrichment factors as observed for the standard process (Supplemental Fig. 1).

Further analysis revealed that the average depth of coverage was also higher for both regions after the second enrichment cycle, being 58.1- and 101.3-fold for *BRCA1* and *TP53*, respectively.

However, the most striking effect was observed for consensus coverages of the HR (percent of HR covered with reads) at increased minimum coverage depths. These numbers are especially important, since a certain minimal depth of coverage is generally required for base calling. This makes a consensus coverage with the minimal depth for reliable base calling the most relevant parameter of an experiment in terms of analytical value for the targeted region. Recent whole human genome sequencing projects using Illumina technology revealed that >95% of both homo- and heterozygous single nucleotide polymorphisms (SNPs) can be accurately called at a coverage depth of 20-fold or higher when paired-end reads are used (Bentley et al. 2008; Wang et al. 2008). The consensus coverage of the HR (i.e., target region) at more than 20-fold depth of coverage can therefore be considered a key parameter for targeted NGS using Illumina instruments.

Strikingly, the consensus coverage with at least 20-fold coverage depth increased between 46- and 96-fold for the two genes from the first to the second cycle of enrichment (Table 1). In total, 68.8%–86.2% of the target regions were covered at ≥20-fold, exceeding previously reported data for targeted sequencing using microarray-based enrichment and Illumina NGS (Hodges et al. 2007).

Capture of 1000 SNP loci

A crucial performance criterion of an enrichment method is its accuracy of base calling. In principle, several steps of the overall process could lead to allelic bias or dropout, which would prevent the practical use of the method for resequencing studies.

To evaluate our method in this direction, we aimed at the enrichment of 1000 nonoverlapping loci of 500-bp size throughout the human genome, each harboring a central dbSNP position. Capture probes with a tiling density of 8 bp were synthesized on four channels of a Geniom biochip, and genomic DNA of two CHB individuals (Chinese individuals from Beijing, HapMap IDs NA18558 and NA18561) was subjected to the two-cycle HybSelect process as described above.

A total of 19,762,440 and 19,405,469 paired end reads of 2 × 36 bp were obtained that were mapped to the ROI after filtering. For the two samples, enrichment factors of 713.3- and 1061.9-fold were obtained. This resulted in average depths of coverage of 315.6- and 469.1-fold over the whole HR (Table 1). Importantly, 80.4% or 85.5% of the HR for all 1000 regions was covered with a depth of at least 20-fold, corresponding well to the obtained consensus coverages for *BRCA1* and *TP53*. This should allow for reliable analysis of most nucleotide positions within the targeted sequence regions.

We performed detailed analysis of consensus coverages and read distributions on the level of the individual loci (a list containing the locus-wise analysis of obtained reads, consensus coverages at one-, five-, 10-, and 20-fold depth of coverage, enrichment factors, and average coverage depths can be found in Supplemental Table 1). Figure 2 shows a histogram of the average depths of coverage for all loci. Remarkably, most regions were covered at a depth of between 250- to 500-fold, with decreasing numbers for higher and lower coverage depths. On average, 90% and 94% of the regions were covered at ≥20-fold, respectively.

Next, we analyzed the uniformity of coverage depth for the whole set of loci. For the most cost-effective sequence capture, uniformity should be maximal since this avoids redundant reads in overcaptured regions. We found that across all regions a fraction of 27%–30% exhibited the average depth of coverage or more. Fifty-one percent to 53% had a normalized coverage depth of 0.5-fold, the average depth of coverage (Supplemental Fig. 2). These data match a uniformity recently reported for a solution-phase capture experiment combined with Illumina NGS technology for a comparable, discontiguous exon target (Gnirke et al. 2009). The availability of long-read platforms like the Roche/454 instrument and the continuing increase of read lengths of the Illumina Genome Analyzer and the ABI SOLiD system raise the question how this might impact the coverage characteristics of the method when applied to these systems. We anticipate that longer read lengths might further improve uniformity and consensus coverages, since regions with lower coverage could be rescued by reads from fragments captured at more distant sites.

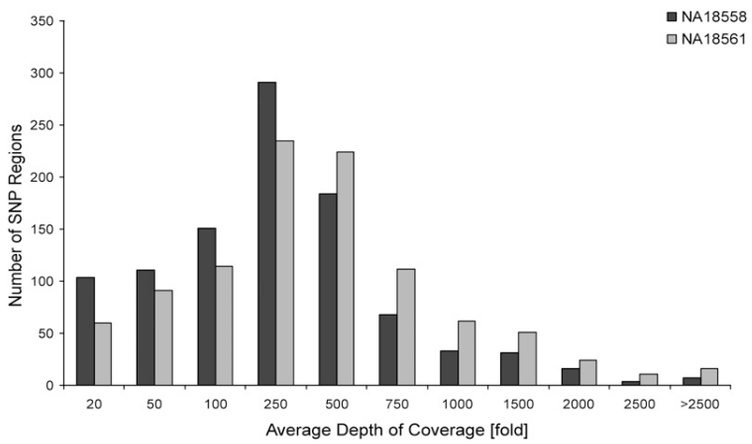


Figure 2. Statistical analysis of average coverage depths and consensus coverages of 1000 human 500-bp loci obtained from mapping analysis after sequence enrichment from the two HapMap reference samples NA18558 and NA18561 and Illumina GAII sequencing. Shown is a histogram of average coverage depths for the HR of individual 500-bp loci for both samples as depicted in the figure.

We next questioned how individual sequence contexts impact the capture performance for the specific regions. Analysis of the correlation between average depth of coverage and GC-content of the 1000 regions for NA18561 revealed that 99.9% of all regions with a moderate GC-content of 40%–60% were covered >20-fold and 98.8% even >50-fold (Supplemental Fig. 3). This suggests considerable potential to even improve the observed capture performance by simple alterations in probe design.

Regional coverage distribution

The design of the dbSNP loci capture experiment with non-overlapping regions of identical size and targeted with identical numbers of capture probes allows a facile statistical analysis of the average spatial distribution of coverage depth over all 1000 ROIs.

It is important to evaluate which fraction of coverage falls into the HR. Since library molecules can extend into the adjacent region within range of the fragment size distribution of the library, sequencing reads can be generated for this noninformative part of the ROI. This effectively decreases the achievable fraction of desired data in the NGS instruments sequence output. Previous microarray studies indicate that the fraction of reads falling into a probe region follows a binomial pattern and depends on the sizes of these regions and the length of the library fragments. The larger the probe region and the shorter the fragment size are, the lower the overlap and the lower the content of noninformative sequence tend to be (Hodges et al. 2007).

In a recent publication, there is further supporting evidence for the notion that longer capture probes could also increase the fraction of noninformative reads. In this study (Gnirke et al. 2009), 170mer probes were used, exceeding the 120-bp median length of human exons. Since library fragments preferentially hybridize with a maximal part of the probe sequence, this leads to considerable overlap into surrounding regions and only a small fraction of 47% in the informative regions. This diminishes the practical use of this enrichment approach for Illumina end sequencing with standard read length.

Analysis of spatial coverage depth distribution for our experiment (NA18561) revealed a binomial pattern with maximal coverage depths in the middle of the HR and relatively low representation of reads falling into noninformative regions (Fig. 3). Coverage depth was thereby highly uniform with only approximately twofold higher depth for the center compared with the edges of the probe regions. Overall, 81% of total coverage was obtained for the targeted HR.

SNP calling accuracy

To assess the applicability of the approach for SNP detection, we analyzed the nucleotide representations of the 1000 captured dbSNP positions. Six hundred of these SNPs were chosen from chromosome 1 and have previously been genotyped in the HapMap project; 400 additional HapMap SNPs were chosen from ENCODE regions on several different chromosomes (dbSNP IDs can be found in Supplemental Table 1). SNPs were thereby selected to have an increased content of 50% heterozygous genotypes within the HapMap CHB population. This allows a balanced analysis of homo- and heterozygous positions and imposes a higher challenge to the process owing to higher coverage requirements and potential bias in nucleotide representation for heterozygous positions. We first filtered the regions for SNP coverage depths of 20-fold or higher as a stringent and pre-established criterion for reliable base calling (Bentley et al. 2008; Wang et al. 2008). Of 1000 SNPs, 913 SNPs fullfilled this criterion, with 449 being homozygous and 464 being heterozygous in the reference data (sample NA18561,

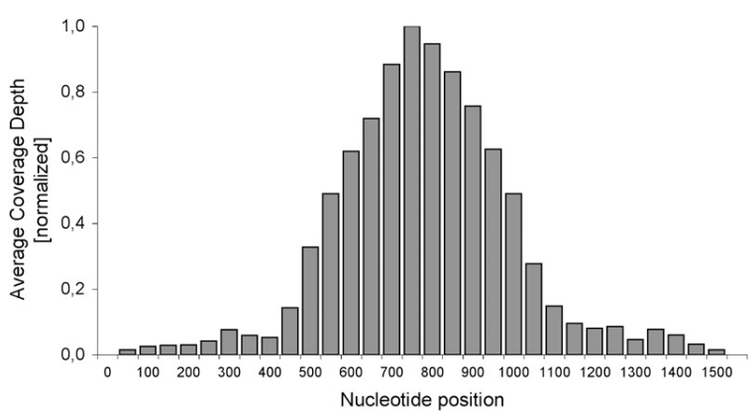


Figure 3. Average spatial distribution of coverage depths for ROI of 1000 human 500-bp dbSNP loci obtained from mapping analysis after sequence enrichment from a human genomic DNA sample and Illumina GAII sequencing. The *x*-axis shows the nucleotide positions of the ROI, consisting of the core region covered by capture probes for array-based sequence enrichment (HR, nucleotide positions 501–1000) with flanking regions of ±500 nucleotides. The *y*-axis shows the coverage depth for all 1000 loci of sample NA18561 averaged for each 50-bp segment and normalized to the maximal depth of coverage.

Supplemental Table 2). Nucleotide analysis and comparison with HapMap reference data (data from HapMap project phases 1 and 2) revealed an overall concordance of 98.6% for all SNPs. Notably, concordance was significantly higher for homozygous positions (99.1%) than for heterozygous positions (98.1%), which suggests that combined call rates for both allele types would be higher for regions that are not enriched for heterozygous occurrences. Analysis of all 464 heterozygous SNP positions revealed an allelic ratio of 0.49, indicating a well-balanced enrichment of both alleles.

Interestingly, very similar concordance (98.8–99.1%, depending on mapping algorithm) was previously reported for nontargeted whole-genome sequencing using Illumina technology and comparison to HapMap reference data of the same project phases (Bentley et al. 2008; Wang et al. 2008). This indicates that the HybSelect process does not interfere with the accuracy of SNP calling and provides a useful tool for resequencing studies.

Conclusion

Sequence enrichment performance

Although several approaches for enrichment of genomic sequences have been reported, no method so far has shown an enrichment performance allowing for reliable SNP calling over the full target region. This has previously been highlighted as the main challenge for hybridization-based sequence enrichment and severely impairs the actual power of NGS technologies (Garber 2008).

Our data show that enrichment factors, consensus coverage, and average depth of coverage for target regions can be multiplied by applying two instead of one enrichment cycle. Compared with two recent studies reporting targeted enrichment using Illumina NGS technology, this resulted in superior enrichment performance and excellent consensus coverages for all targeted regions. Importantly, our calculation of enrichment factors does not include a prefiltering of raw reads for reads uniquely mapping to the human genome. This can reduce the fraction of usable raw reads by a factor of ~0.4–0.5 (Gnirke et al. 2009), whereas the number of unique reads mapping to the target should not be altered. Since this affects the ratio of on-target reads vs. total reads and thus the calculation of enrichment factors and the fraction of on-target reads, we believe that our actual process performance is even better in terms of these parameters than reported here.

Furthermore, this performance was achieved with standard short-read end-sequencing and should further improve with increasing read lengths. Average coverage depths in our experiments exceed those in other studies using this sequencing mode by up to more than one order of magnitude. Uniformity of coverage thereby matches comparable experiments as reported previously.

Uniqueness of NGS reads received after sequence enrichment has not been analyzed in previous studies and consequently the actual value of published coverage depths remains unclear. In contrast, our data show that no significant representation bias is observed in libraries after the HybSelect process, which indicates that no PCR duplicates account for the observed performance. We further showed that the process does not interfere with SNP calling and allows for efficient resequencing of large fractions of the targeted regions with accuracies typically observed for Illumina NGS technology with nonenriched samples.

Advantages of microfluidic biochip architecture

Previous approaches for sequence enrichment employed hybridization steps of >60 h and multiple manual washing and elution steps resulting in long processing times (Albert et al. 2007; Hodges et al. 2007; Okou et al. 2007; Gnirke et al. 2009).

Microfluidic array architecture with associated short hybridization times and a high level of automation throughout the HybSelect procedure enables fast processing and easy handling, despite the use of two enrichment cycles. The total process time starting with a sequencing library and resulting in an enriched, purified, and quantified library ready for Illumina sequencing is less than 60 h, shorter than the hybridization step of any previously reported approach alone.

The used biochips are scalable between one and eight samples and/or 230 kb and >1.8 Mb ROI (125 kb–1 Mb HR) with only 1.5 μg of Illumina library needed per array. This scalability facilitates adjustment of an experiment to different target sizes and can significantly reduce per sample cost for small targets. Further quantitative analyses suggest that biochips can be reused within the two-cycle protocol with typical enrichment performances, which would reduce cost of the approach.

We believe that further improvements in probe design and process optimization will allow us to reach depths of coverage that will enable efficient multiplexing of pooled samples. The general strategy to apply iterative cycles of sequence enrichment might thereby not only facilitate efficient targeted NGS for human genomic subsets. It might also enable analysis of much more complex samples that demand enrichment factors far beyond the possible limit of a single-cycle experiment, e.g., for environmental samples, low abundance cancer cells, or pathogens in a human background. We are therefore convinced that the HybSelect enrichment method will find wide application for large-scale, targeted genomics studies.

Methods

Microarray design and synthesis

Light-activated in situ oligonucleotide synthesis on Geniom biochips (febit biomed gmbh, Heidelberg, Germany) was performed as described previously (Baum et al. 2003). One biochip contains eight individual, microfluidic channels each containing an array of >15,000 individual DNA probe features.

For the enrichment of the two human genes *BRCA1* and *TP53*, 50mer probes were tiled across the target regions with a density of 8 bp, corresponding to a total ROI of 100 kb or a capacity of >1.8 Mb per biochip. Probes were allowed to have a maximal content of 25 low-complexity bases in a row and a maximal total content of low-complexity bases of 80% according to the Hg18 annotation. This resulted in 6700 probes and a reduction of the ROI to the actual probe region (Hybselected region [HR]) of 54 kb, corresponding to a total capacity per biochip of >1 Mb HR.

For enrichment of the 500-bp dbSNP loci, 1000 nonoverlapping regions from high-complexity sequence context throughout the human genome were chosen containing a central dbSNP position. A total of 57,000 50mer probes were designed with a tiling density of 8 bp and synthesized on four array channels again resulting in a capacity of >1 Mb HR per biochip. For all experiments, array designs for the two enrichment cycles were identical.

DNA sample preparation

Human genomic DNA samples NA18558 and NA18561 were obtained from Coriell Repositories. DNA samples for enrichment of *BRCA1* and *TP53* were purchased from Promega. Five micrograms of human genomic DNA were dissolved in 190 μL of water and fragmented for 30 min by sonication at high intensity (Bioruptor, Diagenode). Preparation of the paired-end adaptor-ligated gDNA

library ready for sequencing on an Illumina Genome Analyzer II (Illumina) was performed according to the manufacturer's standard protocol including excision of the size fraction of 300–400 bp from an agarose gel. The sample was analyzed by a Bioanalyzer experiment (Agilent), quantified by UV measurement (Nanodrop 1000, Thermo Scientific), and stored in water at −20°C until use.

Hybridization and elution

For each array, 1.5 μg of an adaptor-ligated gDNA library were dissolved in febit Hybmix-4 or -5, heated to 95°C for 5 min, and placed on ice. The sample mixture was injected into the microfluidic arrays of the biochip and hybridization was performed for 16 h at 45°C or 50°C with active movement of the sample using a febit active mixing device. After hybridization, each array was automatically washed with 6× SSPE at room temperature and 0.5× SSPE at 45°C within the Geniom One instrument (febit biomed gmbh). Each array was subsequently washed with SSPE-based febit stringent wash buffers 1 and 2 at room temperature. For elution of the enriched samples, arrays were each filled with 10 μL of febit elution reagent in a febit hybridization holder and incubated at 70°C for 30 min. Solution was manually transferred into an Eppendorf tube and dried by vacuum centrifugation in a Speed-Vac at 65°C. After an amplification step according to the Illumina library preparation procedure using paired-end primers for 18–35 cycles, the sample was treated like the original library and subjected to a second round of enrichment under the same conditions as before. After enrichment, hundreds of picograms of DNA library are typically recovered from each array depending on the array template as judged by qPCR using the Illumina adaptor primers and SYBRgreen quantitation (data not shown).

NGS using Illumina technology

Eluted samples were subjected to 10 cycles of PCR according to Illumina paired-end library preparation kit and purified by a MinElute PCR purification column (Qiagen). Quantification of samples was done by the Quant-It Picogreen assay (Invitrogen) using the Nanodrop 3300 instrument. Sequencing was performed using an Illumina GAII system using the paired-end mode and read lengths of 36 bp according to the manufacturer's protocol.

Data analysis

Paired-end sequencing reads were first filtered by removing reads with ambiguous nucleotide calls (three or more N) and reads with 34 or more A (or T or C or G). Reads from File 1 and File 2 of the two paired-end sequencing runs were aligned to target genes by using RazerS (Weese et al. 2009), which is part of SeqAn, an open-source C++ library of efficient algorithms and data structures for the analysis of biological sequences (Doring et al. 2008). The parameters used were "-gn 1 -f -r -i 94 -rr 100 -m 10," which allows up to two mismatches. The output alignment files were matched for each pair of reads: The two reads were mapped to opposite strands and in correct orientation and the length between the two reads (inclusive) was within 100–500 bp. The paired reads were matched to the ROI to obtain the reads for analysis of coverage depth. For the 1000 SNP loci experiment, the HR (being all loci of 500 bp) with extensions of ±500 bp for each locus was defined as ROI. The fold coverage for each base within the probe regions was calculated. For unique amplicon analysis, each pair of read sequences was counted only once, and duplicates were ignored. For visualization, reads on the HR obtained by paired-end mapping were mapped with the CLC genomics workbench using single-end mode and default conditions. For SNP analyses, base representations for each target position were calculated in percent. For positions with one base represented >90%, position was called homozygous. If no position was represented >90%, but two bases were represented >10%, position was called heterozygous for these two bases.

Acknowledgments

We thank Jack Leonard and Sonja Vorwerk for helpful discussions and critically reading the manuscript. We thank Andreas Keller for his assistance in setting up razerS for efficient alignment. We thank Anthony Caruso and Marcel Kränzle for assistance in data analysis.

References

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat Methods* **4:** 903–905.

Bau S, Schracke N, Kranzle M, Wu H, Stahler PF, Hoheisel JD, Beier M, Summerer D. 2009. Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays. *Anal Bioanal Chem* **393:** 171–175.

Baum M, Bielau S, Rittner N, Schmid K, Eggelbusch K, Dahms M, Schlauersbach A, Tahedl H, Beier M, Guimil R, et al. 2003. Validation of a novel, fully integrated and flexible microarray benchtop facility for gene expression profiling. *Nucleic Acids Res* **31:** e151. doi: 10.1093/nar/gng151.

Bentley DR. 2006. Whole-genome re-sequencing. *Curr Opin Genet Dev* **16:** 545–552.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53–59.

Coligan JE, Dunn BM, Speicher DW, Wingfield PT, Ploegh HL, ed. 2008. *Current Protocols in Protein Science*. John Wiley & Sons, Hoboken, NJ.

Dahl F, Stenberg J, Fredriksson S, Welch K, Zhang M, Nilsson M, Bicknell D, Bodmer WF, Davis RW, Ji H. 2007. Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc Natl Acad Sci* **104:** 9387–9392.

Doring A, Weese D, Rausch T, Reinert K. 2008. SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics* **9:** 11. doi: 10.1186/1471-2105-9-11.

Garber K. 2008. Fixing the front end. *Nat Biotechnol* **26:** 1101–1104.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27:** 182–189.

Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, et al. 2008. Single-molecule DNA sequencing of a viral genome. *Science* **320:** 106–109.

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, et al. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39:** 1522–1527.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein–DNA interactions. *Science* **316:** 1497–1502.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.

Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. 2007. Microarray-based genomic selection for high-throughput resequencing. *Nat Methods* **4:** 907–909.

Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, et al. 2007. Multiplex amplification of large sets of human exons. *Nat Methods* **4:** 931–936.

Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5:** 1005–1010.

Shendure J, Mitra RD, Varma C, Church GM. 2004. Advanced sequencing technologies: Methods and goals. *Nat Rev Genet* **5:** 335–344.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309:** 1728–1732.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, et al. 2008. The diploid genome sequence of an Asian individual. *Nature* **456:** 60–65.

Weese D, Emde A-K, Rausch T, Döring A, Reinert K. 2009. RazerS—fast read mapping with sensitivity control. *Genome Res* (this issue). doi: 10.1101/gr.088823.108.

Received February 5, 2009; accepted in revised form June 18, 2009.

Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing

Daniel Summerer, Haiguo Wu, Bettina Haase, et al.

Genome Res. 2009 19: 1616-1621 originally published online July 28, 2009
Access the most recent version at doi:10.1101/gr.091942.109

Genomics 95 (2010) 241–246

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods

Targeted high throughput sequencing of a cancer-related exome subset by specific sequence capture with a fully automated microarray platform

Daniel Summerer [a,*], Nadine Schracke [a], Haiguo Wu [b], Yang Cheng [a], Stephan Bau [a], Cord F. Stähler [a], Peer F. Stähler [a], Markus Beier [a]

[a] *febit biomed gmbh, Im Neuenheimer Feld 519, 69120 Heidelberg, Germany*
[b] *febit Inc., 99 Hayden Ave., Lexington, MA 02421, USA*

ARTICLE INFO

Article history:
Received 13 July 2009
Accepted 30 January 2010
Available online 6 February 2010

Keywords:
Exome Sequencing
Genomics
Next-Generation Sequencing
Sequence capture
Microarrays
Microfluidics

ABSTRACT

Sequence capture methods for targeted next generation sequencing promise to massively reduce cost of genomics projects compared to untargeted sequencing. However, evaluated capture methods specifically dedicated to biologically relevant genomic regions are rare. Whole exome capture has been shown to be a powerful tool to discover the genetic origin of disease and provides a reduction in target size and thus calculative sequencing capacity of >90-fold compared to untargeted whole genome sequencing. For further cost reduction, a valuable complementing approach is the analysis of smaller, relevant gene subsets but involving large cohorts of samples. However, effective adjustment of target sizes and sample numbers is hampered by the limited scalability of enrichment systems. We report a highly scalable and automated method to capture a 480 Kb exome subset of 115 cancer-related genes using microfluidic DNA arrays. The arrays are adaptable from 125 Kb to 1 Mb target size and/or one to eight samples without barcoding strategies, representing a further 26 – 270-fold reduction of calculative sequencing capacity compared to whole exome sequencing. Illumina GAII analysis of a HapMap genome enriched for this exome subset revealed a completeness of >96%. Uniformity was such that >68% of exons had at least half the median depth of coverage. An analysis of reference SNPs revealed a sensitivity of up to 93% and a specificity of 98.2% or higher.

Introduction

The enormous capacity of Next Generation Sequencing (NGS) instruments has dramatically changed the scope and comprehensiveness of genomics studies [1–8]. Beside current large scale studies like the 1000 genomes project that are mainly addressed by a limited number of genome centers, the possibility of sequencing relevant subsets of a genome with high sample throughput and at low cost has become a major interest of numerous researchers.

Several new concepts for sequence enrichment have been reported recently that have started to provide a means for efficient, targeted NGS projects. However, these methods still suffer from various drawbacks like limited scalability in terms of sample numbers, poor uniformity resulting in partial dropout of target coverage and time-consuming and complicated workflows [9–11]. Three basic principles of solution phase sequence capture have been reported so far, with each having its own advantages and drawbacks. Molecular inversion probes (MIP) or Selector probes have been used for enrichment of multiple discontinuous target regions with partially high grade of multiplexing and completeness, i.e. percent of target covered [12,13]. However, relatively low uniformity of coverage was also reported and part of the sequencing information was attributed to artifical probe sequence introduced during the enrichment workflow [12–14].

Solution phase enrichment with very long, biotinylated RNA probes has been reported recently [15,16]. However, a drawback of the method was a multi-step capture probe library construction with the potential to introduce bias. Moreover, the length of probes resulted in overrepresentation of off-target reads for short end sequencing that could only be overcome by complicated construction of shotgun libraries or more expensive long read sequencing [15]. Finally, PCR in microdroplets has been demonstrated for sequence enrichment [17], but flexibility of this approach is limited by the requirement of individually synthesized primers and suffers from the fact that primer binding sites have to be designed outside of the actual target regions to avoid nonsense reads from primers incorporated into enriched amplicons. This reduces the amount of relevant information within the sequencers base output and might complicate amplification of regions surrounded by repetitive sequence.

The majority of sequence enrichment methods reported so far was based on solid phase capture using *in situ* synthesized DNA

* Corresponding author. Fax: +49 6221 6510 329.
E-mail address: daniel.summerer@febit.de (D. Summerer).

0888-7543/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2010.01.006

microarrays with flexible content [18–27]. Overall, these methods have relatively short and simple workflows compared to solution phase capturing. A reported drawback was the need for relatively long hybridization times compared to solution phase capture. Three array formats have been used for targeted NGS to date, all allowing for *in situ* synthesis of capture probes and thus providing high flexibility of targeted sequences.

However, for all enrichment approaches, setups dedicated to selected subsets of biologically meaningful genomic loci have been rare. Two very recent studies described the enrichment of the whole human exome with a target size of 26.6 - 34 Mb using microarray capture with two different formats [25,27]. This approach has proven to be a powerful discovery tool, i.e. to reveal the genetic origin of disease by comprehensive exome sequencing of a limited number of individuals [24]. However, owing to the comprehensiveness of the method, significant capacity of not scalable microarrays had to be used for enrichment per sample and multiple sequencing instrument compartments were needed to achieve good coverage depths and completenesses [25,27].

A valuable complementing approach would be the analysis of a smaller subset of relevant genes but involving large cohorts of samples. This would for example allow for an efficient follow-up of genome wide association studies involving whole genome or whole exome sequencing or for other focused studies involving gene sets known to be involved in e.g. cancer development, cardiovascular diseases or drug response. From an economic point of view, such projects would greatly benefit from enrichment systems that are highly scalable to achieve effective further downsizing of targets and increase of sample numbers. Compared to untargeted sequencing, whole exome enrichment approaches represent a drastic reduction in calculative sequencing capacity of 94 – 120-fold. Consequently, focused analysis of relevant genomic subsets with target sizes in the range of several hundred Kb to 1 Mb represent a further reduction in the same order of magnitude.

We report a scalable approach termed HybSelect for selective capturing of focused exome subsets using compartmentalized, microfluidic biochips. The biochips can be processed with up to eight samples in parallel without barcoding strategies and are applicable to target sizes between 125 Kb and 1 Mb. This represents a reduction of calculative sequencing effort of 26 – 270-fold compared to current whole exome approaches. We demonstrate selective capture and sequencing of 115 cancer-related genes with a target size of 0.48 Mb resulting in a capacity of 2 samples per biochip without barcoding strategies. Moreover, the method uses a very simple workflow and is highly automated with potential benefits for cost, reproducibility and contamination risk.

Materials and Methods

Microarray Design and Synthesis

Light-activated *in situ* oligonucleotide synthesis on Geniom Biochips was performed as described previously [28]. One Biochip holds eight individual, microfluidic channels each containing an array of 15.624 individual DNA probe features of which ~120.000 are available for custom probes.

Exon sequences of 115 cancer-related genes from the cancer genome project were downloaded from NCBI and 55.589 50mer probes were tiled across the exon targets of the full region with an average probe density of 9 bp targeting sense and antisense strand in an alternating manner. Each exon was covered by at least 17 probes, i.e. small exons were extended to fit the tiling scheme. The full region of interest (ROI) was 9.2 Mb, corresponding to a core target actually containing exonic sequence of 0.48 Mb. Calculated for the whole biochip, this corresponds to a total capacity of ~20 Mb ROI or >1 Mb target size.

DNA sample preparation

The human genomic DNA sample NA18507 was obtained from Coriell repositories. 5 µg were dissolved in 80 µl of water and fragmented 2 times for 15 min by sonication at medium intensity (Bioruptor, Diagenode, Liége, Belgium). An end repair was performed using T4 DNA polymerase, Klenow Fragment of *E. coli* DNA polymerase I and T4 PNK in T4 DNA ligase buffer for 30 min at 20 °C (all NEB, Ipswich, USA). After purification using the MinElute PCR purification protocol (Qiagen, Hilden, Germany), A deoxynucleotides were added to polished doublestrands using the Klenow fragment (3´-5´- exo$^-$, Qiagen) in presence of 200 µM dATP in Klenow fragment reaction buffer for 30 min at 37 °C. After another MinElute PCR purification, Illumina paired end sequencing adaptors were ligated according to the manufactures protocol. After a Qiaquick PCR purification (Qiagen), ligation mixture was loaded onto a 2% agarose TBE gel and a library band of 200 - 400 bp was excised. Gel slice was purified with the Qiaquick gel extraction kit and 1 of 30 µL eluate was used for a 50 µL amplification reaction using Phusion HF Mastermix (Finnzymes, Espoo, Finland) and 0.2 µM of each primer of pairs pairs AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC and CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG ATC or ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TC and CTC GGC ATT CCT GCT GAA CCG CTC TTC CGA TC. Cycling conditions were: 30 s, 98 °C, then 18 times 10 s, 98 °C; 30 s, 65 °C; 30 s, 72 °C; then 300 s, 72 °C. Purification was performed using the Qiaquick PCR purification protocol. Libraries were analyzed by Bioanalyzer analysis (Agilent, Santa Clara, USA), quantified by Nanodrop 1000 UV measurement (Thermo Scientific, Waltham, USA) and stored in water at -20 °C until use.

Sequence capture protocol

For four arrays, 6 µg adaptor-ligated gDNA library were dissolved in febit Hybmix-4, heated to 95 °C for 5 min and placed on ice. Sample mixture was placed into the sample loading station of the Geniom RT Analyzer and automatically injected into the microfluidic channels of the biochip. Sample was denatured within the chip at 80 °C for 10 min and hybridized for 16 h at 42 °C with active movement of the sample. After hybridization, each array was automatically washed with 6x SSPE at room temperature and 0.5x SSPE at 45 °C. Each array was subsequently washed with SSPE-based febit stringent wash buffers 1 and 2 at room temperature. All protocol steps were carried out in a completely automated fashion by the Geniom RT Analyzer instrument without manual interference. For elution of the enriched samples, arrays were filled with 10 µL of 90% formamide in water each using an elution holder and incubated at 70 °C for 30 min in an oven. Solution was manually transfered into an Eppendorf tube and dried by vacuum centrifugation in a Speed-Vac at 65 °C. After an amplification step as described under DNA sample preparation for 35 cycles, the sample was treated like the original library and subjected to a second round of enrichment under the same conditions as before.

Eluted samples were subjected to 10 cycles of PCR according to the conditions described under DNA sample preparation and purified by Qiagen MinElute PCR purification (Qiagen, Hilden, Germany). Quantification of samples was done by the Quant-It Picogreen assay (Invitrogen, Carlsbad, USA) using the Nanodrop 3300 instrument (Thermo Scientific).

Data analysis

Paired-end Solexa reads (32.878.698 reads with 36 bp length for replicate 1 or 20.700.622 reads with 50 bp in length for replicate 2) were first filtered by removing reads with ambiguous nucleotide calls (3 or more N) and reads with 34 or more A (or T or C or G). This resulted in 15.816.258 or 10.954.170 reads usable for mapping for the

two replicates, respectively. Reads from File 1 and File 2 of the two paired end sequences were aligned with target genes by using razerS, which is part of SeqAn, an open source C++ library of efficient algorithms and data structures for the analysis of biological sequences [29]. The parameters used were "-gn 1 -f -r -i 94 -rr 100 -m 10" which allows up to 2 (36 bp reads) or 3 (50 bp reads) mismatches. The output alignment files were matched for each pair of reads: the two reads were mapped to opposite strands and in correct orientation and the length between the two reads (inclusive) was within 100-500 bp. The paired reads were further matched to extended regions covered by probes (consensus) to get the reads on target. The fold coverage for each base within the probe regions was calculated for unique reads. For SNP calling, individual base fractions for each position having a coverage of 5-fold or higher were calculated and positions were called homozygous if one base accounted for at least 80% and all other bases accounted for less than 10%. If two bases accounted for at least 20% each, the position was called heterozygous. Each called base was compared with UCSC genome hg18 (dbSNP130 masked version). If a difference was found, this position was identified as SNP. SNPs existed in dbSNP were separated from those new ones to calculate the percentages of known vs. novel SNPs.

Results and Discussion

General Workflow for Exome Subset Capture and Sequencing

The overall HybSelect workflow makes use of two key hardware components. The microfluidic Geniom Biochip containing eight individual channels each harboring an array of 15624 freely programmable DNA capture probes is used as sequence enrichment matrix (Fig. 1A). This biochip is processed by the Geniom RT Analyzer which allows for automated sample injection, hybridization with temperature control and active mixing, washing protocols and imaging (Fig. 1B). The HybSelect workflow consists of three basic steps: preparation of a standard genomic DNA library for sequencing, capturing of desired library fragments on the microfluidic arrays including stringent washing to remove unwanted fragments and elution followed by next-generation-sequencing (Fig. 1C). Application of the capture step after library preparation thereby allows facile adaption to different NGS platforms, since all current platforms use adaptor ligated libraries. Thus, no changes to suppliers linker mediated PCR protocols are necessary to adjust library amounts when needed.

We designed an exome subset capture array for enrichment of 115 genes identified in the Cancer Genome project of the Wellcome Trust Sanger Institute as a set highly relevant to the onset of various cancer types. Genes from the cancer gene census list excluding genes known for translocation mutations were used. The final array design contained genes ranging in size from 2.8 to 73.0 Kb with 1819 exons having a minimal, maximal and median size of 2 bp, 8686 bp and 134 bp, respectively. The design covered a total genomic region of interest (ROI) of 9.2 Mb which corresponds to a core exonic region of 0.48 Mb covered by probes. ~56.000 50-mer tiling probes targeting sense and antisense strands in an alternating manner were synthesized with the Geniom One instrument using ~44% of the capacity of a biochip.

Two individual human DNA libraries of the well-characterized Yoruban HapMap sample NA18507 [7,25,30] with length distributions of 200-400 bp and adaptors for Illumina paired-end sequencing were prepared, hybridized for 16 h on two different biochips, and the arrays were washed to remove weakly bound library fragments. The enriched, single stranded samples were eluted, amplified using Illumina paired end primers and subjected to a second cycle of hybridization and washing. After elution, samples were made double stranded by a limited number of PCR cycles.

Sequencing on one lane of a flowcell of an Illumina GA II instrument for each sample using the paired-end mode yielded a total of 15.8 and 11.0 million individual paired end reads after filtering for homopolymeric or ambiguous reads and removal of reads not mapping uniquely to the human genome.

Completeness and Uniformity of Target Coverage

Paired end reads were mapped against the genomic region covered with capture probes and coverage was analyzed. For the

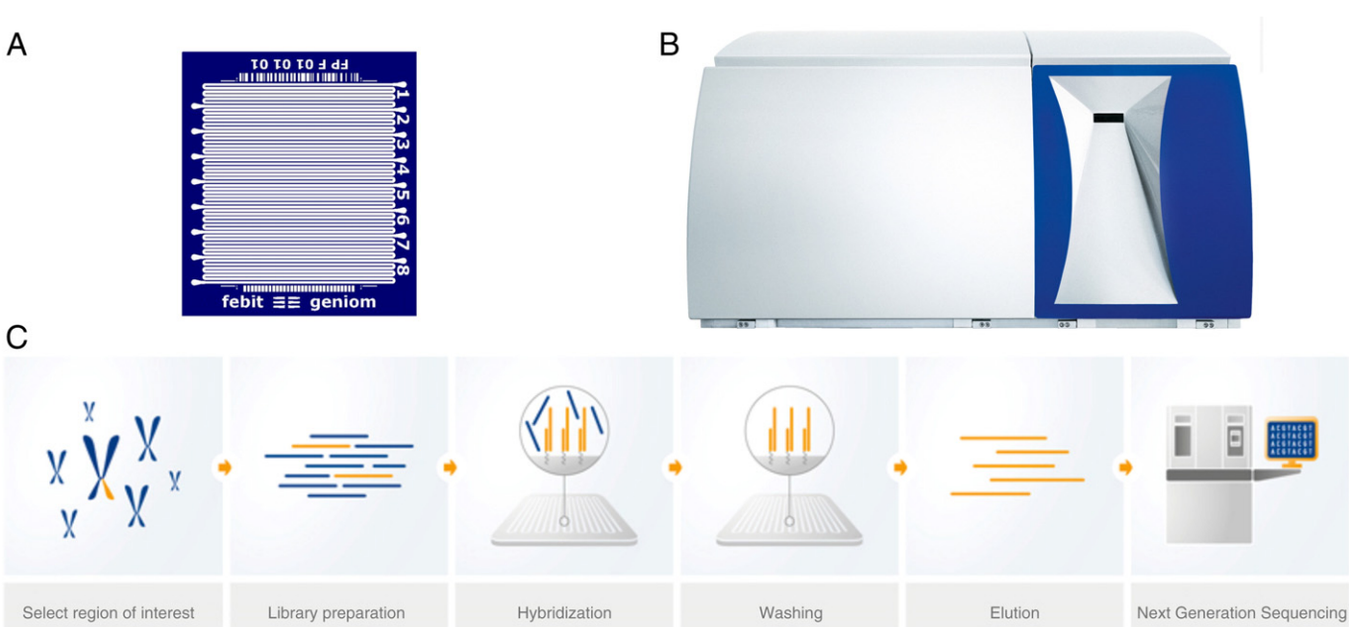


Fig. 1. Hardware and workflow used in the HybSelect process. A: Top view of the microfluidic Geniom Biochip with 8 individual channels each containing an array of 15624 DNA oligonucleotide probes. B: Front view of the Geniom RT Analyzer, a fully integrated microarray processing station allowing for automated sample injection, hybridization with mixing, temperature control, fluidic control and fluorescence detection. C: Workflow of the HybSelect process. Genomic DNA (1) is fragmented and a next generation sequencing library is constructed (2). Library is hybridized to a biochip containing capture probes for the desired target sequences (3) and washed to remove unwanted fragments (4). Desired library fragments are eluted (5) and used for next generation sequencing (6).

two independent replicate experiments, completenesses, i.e. percentages of the target covered at least once, were >96% for both samples (Table 1). For percentages of exon- and gene-wise median coverages, numbers increased to >97% and 100%, respectively. This completeness is in line with previous studies and shows only negligible dropout of target sequence (For a detailed, gene-wise analysis of on-target reads, average target coverages, and percentages of target covered ≥1-, 5- and 10-fold, see Supplementary Table 1).

Beside completeness of coverage, the uniformity of coverage depth is an important parameter of a sequence capture method, since even coverage avoids redundant reads in over-captured regions.

Analysis for all 115 genes revealed that 96% of all genes were in a range of coverage depth of <1 log. This indicates a low dependence of capture efficiency on individual genes and suggests wide applicability of the method to various sequence contexts. A more detailed analysis of coverage uniformity is shown in Fig. 2. The individual median target coverages of all 1819 exons for both replicates were normalized by dividing them by the median target coverage of all exons. By plotting the fraction of total exons exhibiting a specific normalized target coverage, it is possible to analyze and compare coverage uniformity of experiments independently of e.g. platform-dependent effects or overall sequence yield [15,22]. Of all exons, 46.9% and 48.8% exhibited the median target coverage or more, respectively. 69.7% and 68.1% had a normalized target coverage of 0.5 and 84.3% and 85.0% of 0.2. This data indicates similar or better uniformity compared to recently reported studies for solution-phase exonic capture experiments combined with Illumina NGS technology [15–17].

For further improvements, we sought to elucidate the origin of target coverage variability for individual exons. Fig. 3 shows the actual median target coverages of replicate 1 either for all exons (A) or in dependence of GC content of exons (B). A clear trend is visible that comparably low target coverage is obtained for GC contents outside of an optimum range with a lower limit of ~40% and a higher limit of ~60%. This trend is more dominant for exons with low GC contents compared to high GC contents. Overall, 58.5% of all exons fell into the optimum range of 40-60%. For these exons, an excellent completeness of 99.2% was obtained with 98.5% of exons having a target coverage of 5-fold or higher. These data suggest that applying more stringent GC-content criteria during probe design might substantially improve performance of the approach.

Another aspect for further improvement is the dependence of target coverage and exon size. Since sizes of targeted exons span a large range between 2 – 8686 bp, we were interested in dependence of exon-wise median target coverage and exon size. A histogram analysis revealed low variation of target coverage between exons of middle and larger sizes (Supplementary Fig. 1). However, it also pointed at a possibility for a facile further improvement of the method. Very small exons (1-30 bp) exhibited relatively low median target coverage of only 9.5-fold whereas exons 31-60 bp in size were covered at a median of 73-fold with a trend for even higher target coverages for larger exons. Hence, overall performance could also be increased by using denser tiling schemes for extended regions around very small exons.

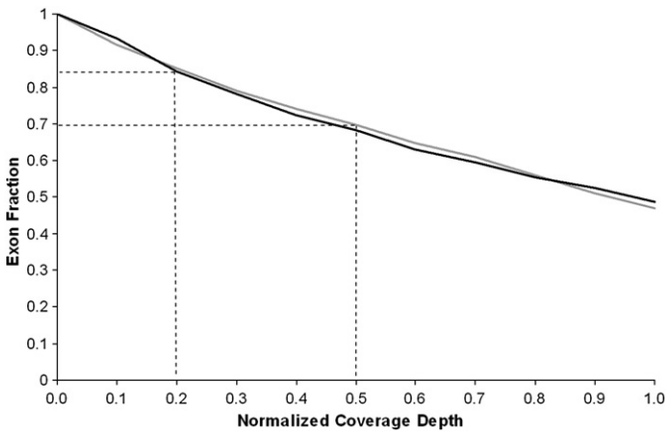


Fig. 2. Uniformity of per base coverage visualized by a normalized coverage distribution plot. Graph shows fraction of targeted exons exhibiting a target coverage equal or higher than the normalized target coverage shown on the x-axis. Normalized target coverage was calculated by dividing individual median target coverages of exons by the median target coverage for all exons. For 0.5- and 0.2-fold of normalized target coverage, exon fractions are indicated as dotted lines.

Detection of Single Nucleotide Polymorphisms (SNP)

Since resequencing for variant discovery is currently the most important application of NGS platforms, a crucial parameter of any sequence enrichment method for NGS is its potential to detect and correctly call novel SNPs. For such an analysis, we included all exon bases of Yoruban HapMap sample NA18507 with coverages of 5-fold or higher which has been used as quality criterion for SNP calling previously [16]. This corresponds to a SNP calling sensitivity (percent of target sufficiently covered for SNP detection) of 88.6 – 93% (Table 1). In these regions, 4998 and 4702 coding SNPs (cSNPs) were detected in the two samples, respectively. A comparison with dbSNP revealed that 89.2% and 91.0% of these SNPs were matching previous database entries. This compares to 74% matches recently obtained for a genome-wide comparison of

Table 1
Statistics of mapping of sequencing reads obtained from Illumina paired end sequencing of two replicate samples enriched for exons of 115 cancer genes. Shown are the sizes of the ROI (region of interest), the target (exonic region covered by capture probes), the number of on-target reads obtained by the two individual sequencing runs of one lane each, average target coverages (fold) and percentages of target covered at a depth of at least 1-fold, 5-fold, 10-fold and 20-fold. Percentages are shown base-wise, exon-wise and gene-wise.

General Metrics:	ROI	Target	On Target Reads	Average Target Coverage
Replicate 1	9345045	482093	2663643	183.82
Replicate 2	9345045	482093	817614	74.05
Percent of Bases covered:	@ ≥1-fold	@ ≥5-fold	@ ≥10-fold	@ ≥20-fold
Replicate 1	97.2	93	89.4	83.3
Replicate 2	96.5	88.6	80.5	68.9
Percent of Exons covered:	@ ≥1-fold	@ ≥5-fold	@ ≥10-fold	@ ≥20-fold
Replicate 1	98.5	95.8	93.2	86.4
Replicate 2	97.8	93.4	84.3	71.2
Percent of Genes covered:	@ ≥1-fold	@ ≥5-fold	@ ≥10-fold	@ ≥20-fold
Replicate 1	100	100	100	100
Replicate 2	100	100	100	93.9

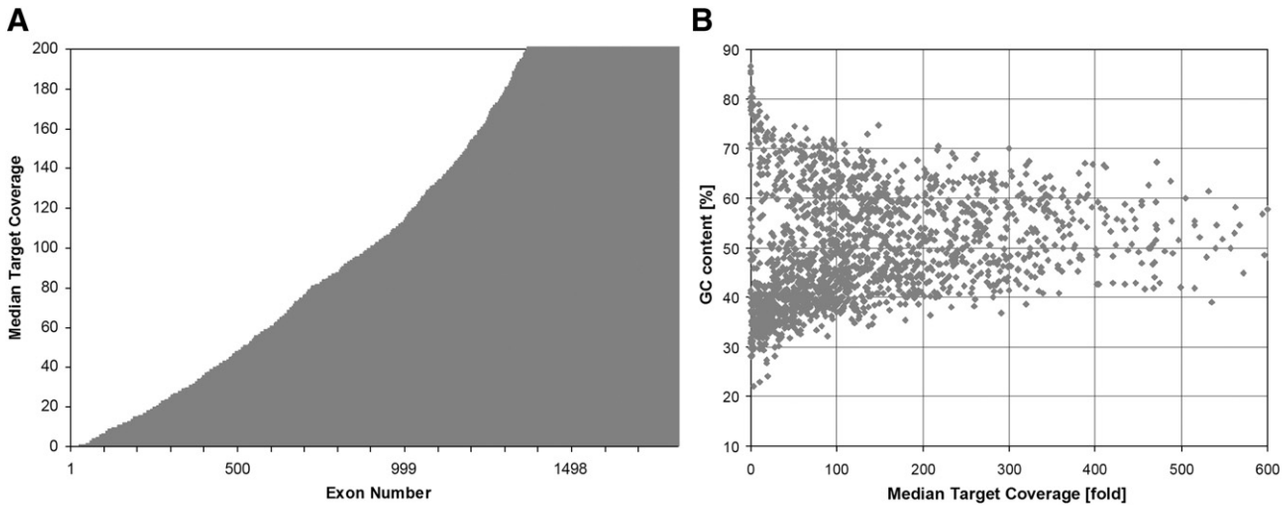


Fig. 3. Exon-wise analysis of median target coverages obtained from mapping of paired end reads of an Illumina GAII sequencing run with sample replicate 1 enriched for 115 cancer related genes. A: Shown is the median fold target coverage for the 1819 individual exons. X-axis shows the individual exon number, y-axis shows the median target coverage for individual exons. B: Median per base coverage for 1819 exons in dependence on exons GC content. X-axis shows the individual median target coverage for exons, y-axis shows the GC content of individual exons in percent.

Illumina sequencing data of the identical HapMap sample [7]. However, in a recent whole exome sequencing project of this sample using Illumina technology, 89.1% concordance was obtained for cSNPs only, which closely mirrors the concordance obtained for our solely exonic target [25].

This indicates a low potential of the approach for false positive calls that could originate from e.g. suboptimal conditions of enrichment, sequencing or mapping methods and would cause an excess of newly identified SNPs vs. previously known database entries.

We next analyzed the percental nucleotide representations of all HapMap reference SNP positions contained in the targeted exons. 836 SNPs with reference data were present in the captured regions that were used for further analysis. Of 836 SNPs, 790 (94.5%) and 754 (90.2%) SNPs were thereby covered 5-fold or higher for the two replicates. Nucleotide analysis and comparison to HapMap reference data (HM-All, data from all HapMap project phases) revealed an overall concordance of 98.2% and 99.1% for all SNPs, similar to specificities reported previously for array based sequence capture [14,22,27] and other enrichment methods [15–17]. Generally, specificity could be further enhanced by increasing the minimum depths of coverage used for filtering of callable positions, however, for the cost of decreasing sensitivities [7,8].

To further understand the origin of SNP calling discrepancies between targeted Illumina sequencing and HapMap genotyping results, we made a follow-up analysis for all non-concordant SNP positions. Different types of discrepancies thereby may hint at different error sources. For example, heterozygous sequencing calls for homozygous HapMap genotypes may hint at accidental base substitutions generated by PCR during library preparation or the HybSelect process when present in one replicate. Presence in both replicates may rather hint at a systematic error e.g. in sequencing, read mapping or HapMap genotyping, since random PCR artifacts in both samples seem unlikely. However, a systematic error that could be associated with a hybridization-based sequence capture method may be loss of heterozygousity due to preferential binding of capture probes to the complementary allele. In our study, there were 21 non-concordant calls found at 14 different positions within the total 1544 calls for SNPs with coverage at 5-fold or higher for both replicates (see Supplementary Table 2). Of these, only 6 (5 positions) were missed heterozygote alleles of which only two occurred in both replicates. In contrast, the majority of discrepancies (12 at 6 positions) were called in both replicates of the sample with almost identical base fractions, suggesting systematic errors that are independent of the sequence capture process. Three positions had relatively low coverage of $\leq$8-fold and one position had coverage of $\geq$5-fold in only one of the replicates.

These data suggest that the majority of non-concordant calls are due to systematic errors in process steps aside from the actual HybSelect procedure and that the actual calling specificity is substantially higher than stated above. Additionally, specificity might increase even further with higher coverage depth of SNPs that were covered poorly.

Conclusion

Taken together, we present a highly scalable method to enrich focused, biologically relevant exome subsets with increased sample numbers. The method provides excellent completeness of coverage with similar or better coverage uniformity than previously reported for exonic targets. This is reflected by high sensitivity and specificity of SNP calling. Our data further suggest that this performance could be even further increased by relatively simple alterations of protocol parameters, i.e. probe design algorithms in terms of GC content and tiling density for very small exons. Microfluidic array architecture with associated short hybridization times and a high level of automation throughout the procedure thereby enables fast processing and easy handling with potential benefits for cost, reproducibility and contamination.

The method efficiently amends technologies involved in large-scale discovery studies such as whole genome or whole exome sequencing. For efficient follow-up projects involving massive sample numbers, scalability of enrichment methods becomes crucial to reduce needed capacities of enrichment and sequencing instrumentation. The architecture of the presented biochip features eight individual array channels with free scalability between 0.125 and 1 Mb and/or one and eight samples. Depending on target size, a throughput of eight samples per two days is the current throughput without barcoding strategies. However, since coverage of most target bases obtained is significantly higher than the threshold of $\geq$5-fold used for SNP calling, it is reasonable to assume that a severalfold increase in throughput could be achieved by barcoding with limited loss in sensitivity. We envision that current efforts for improvement of probe design along the parameters identified in this study as well as further increase in read lengths and numbers of NGS instruments will again strongly increase the potential for massive multiplexing with high numbers of barcoded samples.

Beside the cancer-related biochip presented here, we currently design further pre-evaluated sub-exome biochips for various fields such as neurodegenerative or cardiovascular disease, drug response or human aging.

Acknowledgments

We thank Jack Leonard for helpful discussions and critically reading the manuscript. We thank Andreas Keller for his assistance in setting up razerS for efficient alignment.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2010.01.006.

References

[1] D.R. Bentley, Whole-genome re-sequencing, Curr. Opin. Genet. Dev. 16 (2006) 545–552.
[2] T.D. Harris, P.R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. Dimeo, J.W. Efcavitch, et al., Single-molecule DNA sequencing of a viral genome, Science 320 (2008) 106–109.
[3] D.S. Johnson, A. Mortazavi, R.M. Myers, B. Wold, Genome-wide mapping of in vivo protein-DNA interactions, Science 316 (2007) 1497–1502.
[4] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, et al., Genome sequencing in microfabricated high-density picolitre reactors, Nature 437 (2005) 376–380.
[5] J. Shendure, R.D. Mitra, C. Varma, G.M. Church, Advanced sequencing technologies: methods and goals, Nat. Rev. Genet. 5 (2004) 335–344.
[6] J. Shendure, G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra, G.M. Church, Accurate multiplex polony sequencing of an evolved bacterial genome, Science 309 (2005) 1728–1732.
[7] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, K.P. Hall, D.J. Evers, C.L. Barnes, H.R. Bignell, et al., Accurate whole human genome sequencing using reversible terminator chemistry, Nature 456 (2008) 53–59.
[8] J. Wang, W. Wang, R. Li, Y. Li, G. Tian, L. Goodman, W. Fan, J. Zhang, J. Li, J. Zhang, et al., The diploid genome sequence of an Asian individual, Nature 456 (2008) 60–65.
[9] K. Garber, Fixing the front end, Nat. Biotechnol. 26 (2008) 1101–1104.
[10] D. Summerer, Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing, Genomics 94 (2009) 363–368.
[11] E.H. Turner, S.B. Ng, D.A. Nickerson, J. Shendure, Methods for genomic partitioning, Annu. Rev. Genomics Hum. Genet. 10 (2009) 263–284.
[12] F. Dahl, J. Stenberg, S. Fredriksson, K. Welch, M. Zhang, M. Nilsson, D. Bicknell, W.F. Bodmer, R.W. Davis, H. Ji, Multigene amplification and massively parallel sequencing for cancer mutation discovery, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 9387–9392.
[13] G.J. Porreca, K. Zhang, J.B. Li, B. Xie, D. Austin, S.L. Vassallo, E.M. LeProust, B.J. Peck, C.J. Emig, F. Dahl, et al., Multiplex amplification of large sets of human exons, Nat. Methods 4 (2007) 931–936.
[14] E.H. Turner, C. Lee, S.B. Ng, D.A. Nickerson, J. Shendure, Massively parallel exon capture and library-free resequencing across 16 genomes, Nat. Methods 6 (2009) 315–316.
[15] A. Gnirke, A. Melnikov, J. Maguire, P. Rogov, E.M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, C. Russ, et al., Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing, Nat. Biotechnol. 27 (2009) 182–189.
[16] R. Tewhey, M. Nakano, X. Wang, C. Pabon-Pena, B. Novak, A. Giuffre, E. Lin, S. Happe, D.N. Roberts, E.M. Leproust, et al., Enrichment of sequencing targets from the human genome by solution hybridization, Genome Biol. 10 (2009) R116.
[17] R. Tewhey, J.B. Warner, M. Nakano, B. Libby, M. Medkova, P.H. David, S.K. Kotsopoulos, M.L. Samuels, J.B. Hutchison, J.W. Larson, et al., Microdroplet-based PCR enrichment for large-scale targeted sequencing, Nat. Biotechnol. 27 (2009) 1025–1031.
[18] T.J. Albert, M.N. Molla, D.M. Muzny, L. Nazareth, D. Wheeler, X. Song, T.A. Richmond, C.M. Middle, M.J. Rodesch, C.J. Packard, et al., Direct selection of human genomic loci by microarray hybridization, Nat. Methods 4 (2007) 903–905.
[19] S. Bau, N. Schracke, M. Kranzle, H. Wu, P.F. Stahler, J.D. Hoheisel, M. Beier, D. Summerer, Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays, Anal. Bioanal. Chem. 393 (2009) 171–175.
[20] E. Hodges, Z. Xuan, V. Balija, M. Kramer, M.N. Molla, S.W. Smith, C.M. Middle, M.J. Rodesch, T.J. Albert, G.J. Hannon, et al., Genome-wide in situ exon capture for selective resequencing, Nat. Genet. 39 (2007) 1522–1527.
[21] D.T. Okou, K.M. Steinberg, C. Middle, D.J. Cutler, T.J. Albert, M.E. Zwick, Microarray-based genomic selection for high-throughput resequencing, Nat. Methods 4 (2007) 907–909.
[22] D. Summerer, H. Wu, B. Haase, Y. Cheng, N. Schracke, C.F. Staehler, M.S. Chee, P.F. Stahler, and M. Beier, Microarray-based Multicycle-Enrichment of Genomic Subsets for Targeted Next-Generation-Sequencing, Genome Res. 19 (2009) 1616–1621.
[23] N. Schracke, T. Kornmeyer, M. Kranzle, P.F. Stahler, D. Summerer, M. Beier, Specific sequence selection and next generation resequencing of 68 E. coli genes using HybSelect, New Biotechnol. 26 (2009) 229–233.
[24] S.B. Ng, K.J. Buckingham, C. Lee, A.W. Bigham, H.K. Tabor, K.M. Dent, C.D. Huff, P.T. Shannon, E.W. Jabs, D.A. Nickerson, et al., Exome sequencing identifies the cause of a mendelian disorder, Nat. Genet. 42 (2010) 13–14.
[25] S.B. Ng, E.H. Turner, P.D. Robertson, S.D. Flygare, A.W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E.E. Eichler, et al., Targeted capture and massively parallel sequencing of 12 human exomes, Nature 461 (2009) 272–276.
[26] D.T. Okou, A.E. Locke, K.M. Steinberg, K. Hagen, P. Athri, A.C. Shetty, V. Patel, M.E. Zwick, Combining microarray-based genomic selection (MGS) with the Illumina Genome Analyzer platform to sequence diploid target regions, Ann. Hum. Genet. 73 (2009) 502–513.
[27] M. Choi, U.I. Scholl, W. Ji, T. Liu, I.R. Tikhonova, P. Zumbo, A. Nayir, A. Bakkaloglu, S. Ozen, S. Sanjad, et al., Genetic diagnosis by whole exome capture and massively parallel DNA sequencing, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 19096–19101.
[28] M. Baum, S. Bielau, N. Rittner, K. Schmid, K. Eggelbusch, M. Dahms, A. Schlauersbach, H. Tahedl, M. Beier, R. Guimil, et al., Validation of a novel, fully integrated and flexible microarray benchtop facility for gene expression profiling, Nucleic Acids Res. 31 (2003) e151.
[29] A. Doring, D. Weese, T. Rausch, K. Reinert, SeqAn an efficient, generic C++ library for sequence analysis, BMC Bioinformatics 9 (2008) 11.
[30] K.J. McKernan, H.E. Peckham, G.L. Costa, S.F. McLaughlin, Y. Fu, E.F. Tsung, C.R. Clouser, C. Duncan, J.K. Ichikawa, C.C. Lee, et al., Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding, Genome Res. 19 (2009) 1527–1541.

New Biotechnology • Volume 27, Number 2 • May 2010 RESEARCH PAPER

Research Paper

A flexible and fully integrated system for amplification, detection and genotyping of genomic DNA targets based on microfluidic oligonucleotide arrays

Daniel Summerer[1,3], Dona Hevroni[2,3], Amit Jain[2,3], Olga Oldenburger[1], Jefferson Parker[2], Anthony Caruso[2], Cord F. Stähler[1], Peer F. Stähler[1] and Markus Beier[1]

[1] febit biomed gmbh, Im Neuenheimer Feld 519, 69120 Heidelberg, Germany
[2] febit Inc., 99 Hayden Avenue, Lexington, MA 02421, USA

A strategy allowing for amplification, detection and genotyping of different genomic DNA targets in a single reaction container is described. The method makes use of primer-directed solution-phase amplification with integrated labeling in a closed, microfluidic oligonucleotide array. Selective array probes allow for subsequent detection and genotyping of generated amplicons by hybridization. The array contains up to 15,624 programmable features that can be designed, *de novo* synthesized and tested within 24 hours using an automated benchtop microarray synthesizer. This enables rapid prototyping and adaptation of the system to newly emerging targets such as pathogenic bacterial or viral subtypes. The system was evaluated by amplifying and detecting different loci of viral (HPV), bacterial (*Bacillus* sp.) and eukaryotic (human) genomes. Multiplex PCR and semi-quantitative detection with excellent detection limits of <100 target copies is hereby demonstrated. The high automation grade of the system reduces contamination risk and workload and should enhance safety and reproducibility.

Introduction

The ever increasing understanding of organism complex nucleic acid repertoire such as genome structure and stability, microbial diversity or transcriptional dynamics has called for highly parallel detection, identification and quantification of nucleic acids. Owing to its excellent sensitivity and accuracy, PCR has hereby been a central target for the development of highly multiplexed assay technologies.

To increase throughput of PCR assays, several strategies have been followed. One strategy is the combination of multiple homogenous PCR systems using several specific or degenerate primer pairs within one reaction vessel. Target detection is then often achieved by using fluorescent signaling probes, such as labeled oligonucleotides that undergo changes in fluorescence behavior owing to nucleolytic cleavage or conformational changes during PCR product formation [1–5]. Many of these methods are very mature, allow for quantitative real-time analysis of samples for multiple targets and have found widespread application in research and molecular diagnostics. However, a limitation of homogenous multiplex PCR systems with fluorescent signaling has so far been the relatively low multiplexing grade owing to spectral overlap and resulting cross-talk of fluorophores.

A second strategy for increased throughput that circumvents cross-talk by spatial separation is the parallelization of individual, homogenous PCR setups with limited or no primer-pair multiplexing within single reactions. However, scaling PCR to analyze larger numbers of targets and samples simultaneously is limited by the logistics and cost of the assay when performed in traditional multiwell-plate formats. Consequently, recent developments have focused on the miniaturization of individual PCR reactions leading to parallel methods with high to very high throughput [6–14]. Although multiplexing grade in terms of amplification targets within a given sample can be increased by higher parallelization

Corresponding author: Summerer, D. (daniel.summerer@febit.de)
[3] These authors contributed equally to this work.

1871-6784/$ - see front matter doi:10.1016/j.nbt.2010.03.005

of individual reactions, this approach is limited. Partitioning of a sample has to be compensated by increased sensitivity of individual assays. Moreover, this strategy is inapplicable in cases where the required assay numbers for a sample exceed copy numbers of individual targets due to overdilution. However, assays in molecular diagnostics may often require the detection of low abundance nucleic acids in the range of <100 copies per sample, which imposes an intrinsic limit to the approach. Hence, there is increasing demand for methods that combine high multiplexing grade of targets without sample partitioning during PCR. One attractive option is to apply heterogenous detection systems to multiplex PCRs conducted in single vessels. This allows for spatial separation of detection events and individual readout without signal cross-talk as in homogenous detection systems. This strategy has, for example, been used in a method that applies beads with individual, target-specific receptors such as oligonucleotide probes that are simultaneously added to an amplicon mixture for binding. Up to 100 different bead types can thereby be optically identified by fluidic separation and color coding and bound PCR products quantified by fluorescence [15,16].

Even higher multiplexing grades can be achieved by the use of oligonucleotide arrays that can simultaneously detect hundreds of thousands or millions of different nucleic acid sequences in parallel [17]. Numerous microarray-based assays have been described for the detection of different target types such as viruses, bacterial pathogens or human genetic variants [18–23]. However, workflows of such methods have been rather labor- and time-intensive with partially separated amplification, labeling, microarray hybridization, washing and detection, often involving purification and individual hardware for the various processing steps.

Here, we describe a method using only one processing station and a single, microfluidic oligonucleotide array that serves as a low volume compartment for all steps of a nucleic acid detection and typing process. This includes amplification and labeling of nucleic acid targets, array hybridization, washing, fluorescent staining and detection of individual PCR products. The method is evaluated with viral, bacterial and human nucleic acid targets in multiplexing mode and a detection limit of <100 copies is demonstrated. By using a fully automated platform for *de novo* array synthesis, probe content is highly flexible with a prototyping iteration cycle of probe design, microarray synthesis, experimental testing and microarray redesign of less than 24 hours. This allows the rapid development of novel assay formats to adapt the system to novel target sequences such as emerging viral or bacterial pathogenic subtypes.

Materials and methods

DNA samples and oligonucleotides

Plasmid containing the entire genome of HPV 6b (ATCC-45150D) was obtained from LGC Promochem. Bacterial genomic DNA was obtained from ATCC. Used species were *B. cereus* (ATCC 14579), *B. subtilis str. 168* (ATCC 23857) and *B. thuringiensis ser. israelensis* (ATCC 35646). Oligonucleotides were purchased from Sigma Genosys.

Amplification protocols

For HPV PCR experiments, varying copy numbers of pHPV6b were amplified using Absolute Mastermix (ABgene) in the presence of 100 μM Biotin-16-dUTP and 0.5 μM primer pool MY09 (CGTCCMARRGGAWACTGATC), 0.5 μM primer pool MY11 (GCMCAGGGWCATAAYAATGG), varying amounts of human genomic DNA (Promega) and/or 0.25 μM primer β-Glob_fwd (CAACTTCATCCACGTTCACC) and 0.25 μM primer β-Glob_rev (GAAGAGCCAAGGACAGGTA). PCRs from a single mastermix were conducted as control in parallel in tubes using a Mastercycler (Eppendorf) and in a microfluidic Geniom Biochip using an Amplispeed microarray slide thermocycler (Advalytix) or a Geniom RT Analyzer instrument. Cycling conditions were as follows: 15 min 95°C, then 10 times: (1 min 95°C, 1.5 min 57°C, 1 min 72°C), then 25 times: (1 min 95°C, 1 min 55°C, 1 min 72°C), then 5 min 72°C. Product mixtures were either analyzed by agarose gel electrophoresis or immediately hybridized in microfluidic channels for reactions carried out in a Geniom biochip (see below). After PCR establishment, identities of all PCR products were confirmed by Sanger sequencing.

6-Plex PCRs targeting four *Bacillus* strains, HPV 6b and human β-globin contained the same concentrations of primers MY09, MY11, β-Glob_fwd and β-Glob_rev as well as 0.5 μM of each of the primers B.cereus_446_F (CCTACTATAATCCATGCA), B.cereus_446_R (GGAGAAGATAGAATTGCT), B.Sub_395_F (CCTTCTATTTCTAACGCA), B.Sub_395_R (CGATAATCATTGATCCGT), B.Thu.Isr_407_F (CCATTCATGATAACTGCT), B.Thu.Isr_407_R (GGTACCGTAATTATTGGA). Reactions further contained 1× ThermoStart buffer (Abgene), 500 μM of each dATP, dGTP and dCTP, 125 μM TTP, 18.75 μM Biotin-16-dUTP, 0.03125 U/μL ThermoStart DNA polymerase (Abgene) and varying amounts of different templates as specified in the 'Results' section. Cycling conditions were as described above. Product mixtures were either analyzed by agarose gel electrophoresis or immediately hybridized in microfluidic channels for reactions carried out in a Geniom biochip (see below).

Microarray design and synthesis

Light-activated *in situ* oligonucleotide synthesis was performed essentially as described [24] using a digital micromirror device (DMD, Texas Instruments). This allows for light-directed activation on a microfluidic array consisting of a glass–silicon–glass sandwich within the Geniom instrument (febit biomed). Depending on the number of DMD micromirrors used for one feature and for the spacing between features, each chip consists of eight arrays with 6776 (2 × 2 mirrors for each feature with 1 mirror spacing) or 15,624 (1 mirror for each feature with 1 mirror spacing) individual features.

For selective detection, 21mer tiling probes for the targeted PCR products of β-globin (248 probes), HPV 6b (880 probes, including surrounding region of target sequence) were designed with a 1 bp resolution to analyze the complete sequences for the specific regions. For further array designs after initial experimental validation of sensitivity and selectivity of probes, a 6776-feature setup was used.

Bacillus detection probes were designed against forward and reverse strands of amplicon regions with a target size of 25 bp using a 5 bp tiling offset. Probes were further selected to eliminate ambiguous bases, runs of four or more identical bases (polyN) and to maximize 5′-terminal uniqueness across the target genomes. Each experimental amplicon generated 60–70 probes per strand and PCR product target.

Microarray hybridization, detection and data analysis

For experiments using purified, labeled PCR products, samples were dissolved in febit hybridization mix-1, including febit control oligo mix, heated to 95°C for 5 min and placed on ice. Geniom biochip was denatured by washing with water at 80°C and incubated with febit prehybridization buffer for 15 min at room temperature. Buffer was removed and sample was injected into microfluidic arrays and incubated for four hours at 45°C. Samples were removed and the biochip was automatically washed consecutively with 6× SSPE at room temperature and 0.5× SSPE at 45°C within the Geniom device. Streptavidin–(R)-phycoerythrin (SAPE, Invitrogen) in 6× SSPE was injected and the biochip was incubated for 15 min at room temperature and then washed with 6× SSPE. Fluorescence image was acquired using the integrated detection system of the Geniom device.

For experiments using integrated hybridization of PCR product mixes, the biochip was directly used for PCR without denaturation or incubation with febit prehybridization buffer. PCRs were conducted using the Amplispeed microarray slide thermocycler (Advalytix) or a Geniom RT analyzer instrument. PCR mastermix as described above was injected into microfluidic channels and biochip was subjected to the PCR program. Chip was denatured for 5 min at 95°C immediately after the PCR reaction, cooled to 45°C and incubated for 16 h. Afterwards the target solution was removed from arrays and analyzed by electrophoresis on a 2.5% agarose gel. Biochip was washed and stained as described above. After the first wash step after SAPE-incubation, a protocol for signal amplification was performed. For 6-plex PCRs, incubation with an antibody-solution (1× MES, 0.925 M NaCl, 0.05% Tween-20, 1 mg/ml BSA) containing multi-biotinylated anti-streptavidin antibody (Vector Laboratories; BA-0500, 1:167 diluted) and goat IgG (Sigma; I5256, 1:100 diluted) as second antibody was conducted. After a second incubation with SAPE and washing, amplified fluorescence was detected. For PCRs targeting HPV 6b and human β-globin only, signal amplification was conducted using the Anti-Biotin Oyster 550 (900) signal amplifier antibody (Genisphere) at 10 ng/μL under the conditions described above. For the analysis of fully integrated microarray experiments, raw fluorescence intensities were recorded, medians of probe replicates were calculated and medians of background features (consisting of a single T residue) were subtracted.

Results and discussion

We aimed at developing an integrated system for multiplex detection and subtyping of various DNA targets using a single reaction vessel with a simple and automatable workflow. The established approach makes use of the closed, microfluidic Geniom biochip that contains eight individual microchannels with a volume of ~3 μL each presenting an array of 6776 or 15,624 DNA capture probe features on its inner surface. The overall workflow is outlined in Fig. 1. A purified genomic DNA sample is mixed with a PCR mastermix containing all components necessary for efficient amplification of targeted loci including biotin-16-dUTP for integrated random labeling. The mixture is injected into the chip and

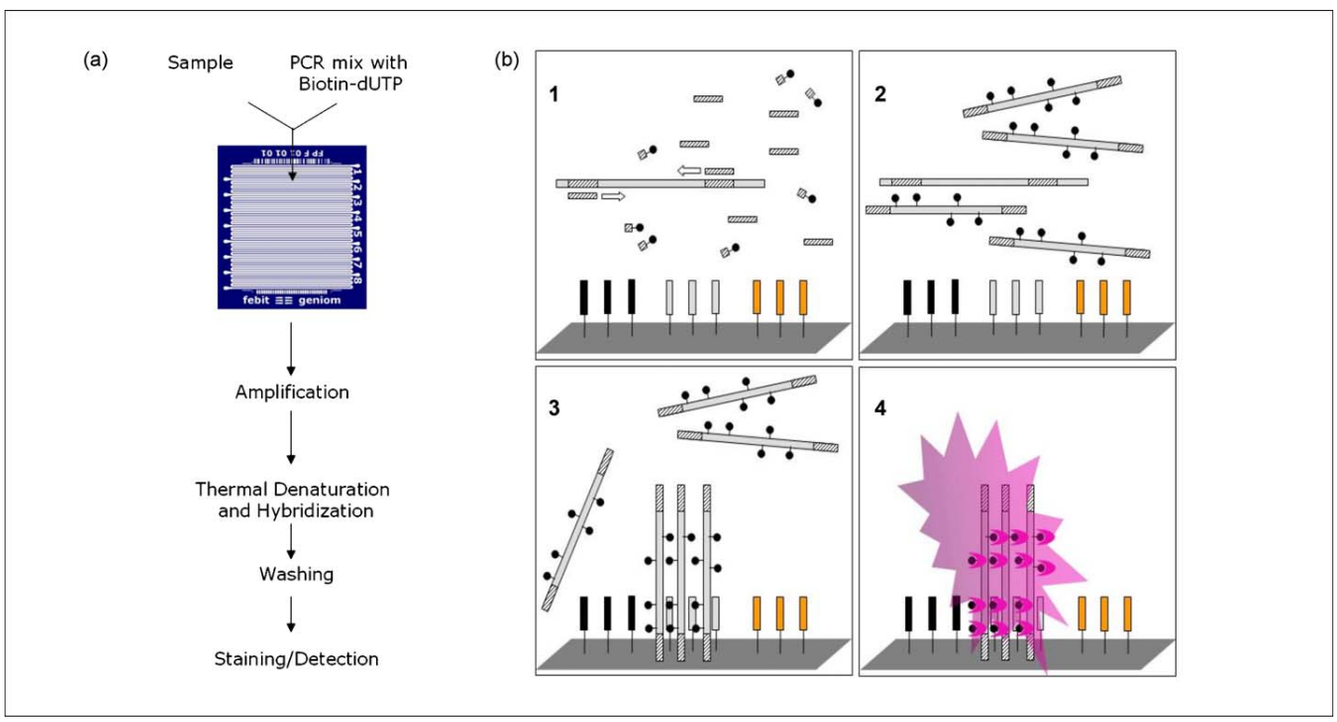


FIGURE 1

Overview of fully integrated amplification, detection and typing of genomic DNA targets. **(a)** General workflow. Purified nucleic acid sample containing target nucleic acid is combined with PCR mastermix. Mixture is injected into a microchannel of a Geniom biochip containing selective DNA probes for binding of formed PCR products. Amplification protocol including integrated labeling, hybridization, washing, fluorescent staining and detection is automatically conducted in the employed processing platform. **(b)** Scheme of in-chip process. Sample DNA molecules are amplified in the presence of biotin-16-dUTP (1) leading to randomly biotinylated PCR products (2). Products are hybridized to selective DNA capture probes and non- or weakly bound DNA is washed away (3). Chip is incubated with Streptavidin–(R)-phycoerythrin, washed and fluorescence is recorded (4).

amplification is performed within the chip followed by a thermal denaturation and a hybridization step at 45°C for selective binding of the probes. After stringent washing, the chip is incubated with a fluorescent streptavidin conjugate (Streptavidin–(R)-phycoerythrin, SAPE), washed and an antibody-based signal amplification process is conducted. Fluorescence is recorded and analysis reveals the presence of individual amplicons.

For the detection of viral genomes, a generic PCR for the amplification of an L1 region fragment of various human papillomavirus (HPV) subtypes based on the known degenerate primer system MY09/11 was established [25]. The MY09/11 primer set is especially well suited for subtyping using high density DNA microarrays, because PCR products have sufficient size to allow for the design of probes against multiple L1 loci, thereby enhancing reliability of subtype discrimination. Other generic primer systems like the GP5+/GP6+ or SPF mixes yield much smaller products and therefore do not sufficiently match the potential of microarray probe content [26–28]. To verify the presence of human genomic DNA within a sample, PCRs were performed as duplex PCRs containing a second primer pair targeting the human β-globin gene, which is widely used as positive control marker in human detection assays. Additionally, a multiplex PCR represents a more challenging test-case for the microchannel environment.

For subtyping of PCR products, capture probes were designed for the β-globin control product and the L1 PCR product of primer pair MY09/11. 21mer tiling probes with a resolution of 1 bp were designed to cover the whole region of all PCR products and a prototype array with a 15,624 feature density was synthesized. In that way, a maximal number of probes per sequence can be experimentally validated by the hybridization of individual PCR products from which a fraction with desired sensitivities and selectivities can be chosen for the design of an optimized subtyping array. Purified, biotinylated PCR products of HPV 6b and β-globin were individually hybridized to two arrays containing all designed probes and relative binding efficiencies as well as cross-hybridization tendencies were determined. All probes exhibiting a median discrimination of the noncognate PCR product of >6 were used for the design of a second generation subtyping array. This resulted in 168 and 78 probes specific for β-globin and HPV 6b, respectively.

This selection was further evaluated in a second round of cross-hybridization experiments for the selection of a minimal number of highly specific probes with a 6776 probe array. Overall, 8 probes for HPV 6b and 3 probes for β-globin were used for the final array design. Using 8 probes per target thereby results in a theoretical multiplexing level of >260 or >600 targets when array densities of 6776 or 15,624 features/array are used and each probe is included in 3 replicates.

PCRs targeting HPV and the human β-globin gene were conducted in the microchannels of a Geniom biochip with a volume of ~3 μL per channel using the Geniom RT Analyzer as processing station. This platform facilitates the workflow of the in chip PCR by featuring integrated PCR temperature cycling, hybridization, washing routines, fluorescence staining and detection. Plasmids containing genomes of HPV subtype 6b were spiked in varying concentrations into a background of human genomic DNA. This allows for controlled titration of virus copies in a typical complexity of a patient sample.

As a first test, PCR product formation was analyzed by agarose gel analysis after the removal of the PCR mixture from the channels to assess product purity and detection limit with a standard technique. In singleplex mode, of HPV L1 PCR product was clearly visible with a detection limit of ~620 copies per 3 μL PCR reaction within a background of 0.5 ng human genomic DNA per reaction with good reproducibility (Fig. 2).

When targeting β-globin in a singleplex PCR, product formation was visible when starting with 0.25 ng (~75 genome copies) of human genomic DNA. Moreover, both the β-globin and HPV L1 products could be detected by agarose gel electrophoresis when using 310 copies of HPV 6b and 0.25 ng human genomic DNA as starting amounts in multiplex PCR mode (Fig. 2). This demonstrates that the employed microchannels can be used as PCR reaction containers that allow for excellent sensitivity. Importantly, these results were obtained using previously known standard PCR primer systems and standard reaction conditions without special adaptation of the applied conditions to the microchannels.

Next, PCR product typing performance of the microarray was tested in the fully integrated workflow with PCR conducted in the microchannels. Chips used for multiplex PCRs containing different starting amounts of human gDNA and HPV 6b with a non-template control PCR were conducted with integrated hybridization, washing, staining and detection. Fluorescence data of probe features are shown in Fig. 3. No significant fluorescence was observed in negative controls for HPV 6b-specific probes, whereas

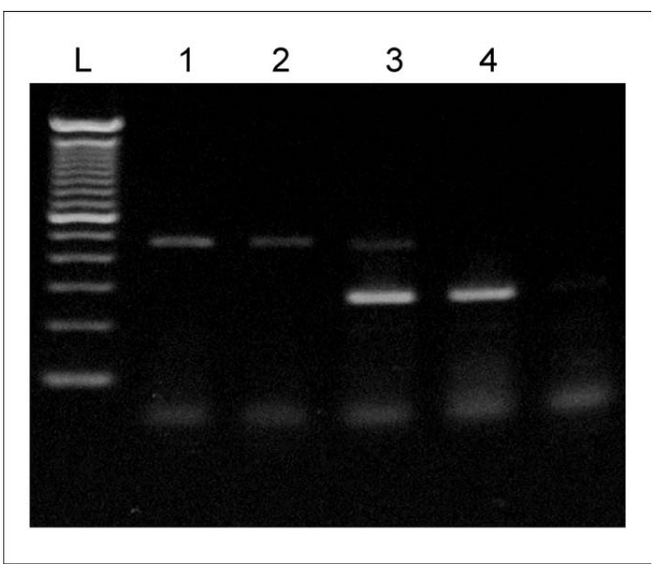


FIGURE 2

Multiplex PCR for generic amplification of HPV and specific control amplification of β-globin from samples containing human genomic DNA and HPV genomic DNA. PCR was conducted using the primer pool MY09/11 for generic amplification of the HPV L1 region, specific primers for human β-globin and biotin-16-dUTP for integrated labeling. All PCRs were performed in the microfluidic channels of a Geniom biochip. L: Ladder. 1: PCR conducted with 620 copies HPV 6b and 0.5 ng human gDNA as template with primers targeting HPV only. 2: Replicate PCR of lane 1. 3: PCR conducted with 310 copies HPV 6b and 0.25 ng human gDNA as template with primers targeting HPV and human β-globin. 4: PCR conducted with 310 copies HPV 6b and 0.25 ng human gDNA as template with primers targeting human β-globin only. 5: negative control.

some background fluorescence was visible for the β-globin probes in the absence of template. However, observed fluorescence for all tested probes clearly exhibited dependence on starting amount of template for both targets. This allows a semi-quantitative analysis of target nucleic acids and could be employed for the determination of viral load and amount of host cell material within a patient sample. Importantly, the detection limit using fluorescence of probe-bound amplicons is lower than the limit obtained for agarose gel electrophoresis and allows the detection of ~100 copies of HPV 6b, ~75 human genome copies and a viral load of 1.3 HPV 6b copies per human genome copy.

To test further targets and more demanding PCR complexities, a 6-plex PCR system was next established. Four primer pairs targeting individual loci on three different Bacillus strains (*Bacillus cereus*, *Bacillus subtilis* and *Bacillus thuringiensis ser. israelensis*) were designed for maximal specificity between these strains. Primer pairs were individually tested in PCR tubes against their respective target genomic DNA (data not shown). For a 6-plex PCR, the three primer pairs were used in combination with primers specific for HPV 6b and β-globin. PCRs targeting *B. cereus* (3000 genome copies/array), *B. subtilis* (3000 copies) and *B. thuringiensis ser. israelensis* (2300 copies) in the presence or absence of human genomic DNA (1 ng/array) were conducted in parallel using a regular PCR tube or the microchannels of a biochip as reaction container. All microchannels contained identical sets of capture probes specific for the targeted amplicons. PCR products were formed in both the presence and absence of human genomic DNA with no clear resolution of *Bacillus*-related products, presumably owing to similar amplicon lengths (Fig. 4a). No significant formation of byproducts was observed and no product was present in negative control PCR without genomic DNA. PCRs carried out in microchannels with integrated hybridization and fluorescence imaging afforded a collection of capture probes with sufficient intensity for further analysis (Fig. 4b). In contrast to agarose gel electrophoresis, the presence of all three *Bacillus*-related products was indicated by fluorescence signals in positive reactions (Fig. 4b, 1–2), whereas no significant fluorescence was observed for most probes in the negative control PCR (Fig. 4b, 3). PCR reactions in the absence of human genomic DNA were carried out in duplicate using two arrays and individual capture probes within the two

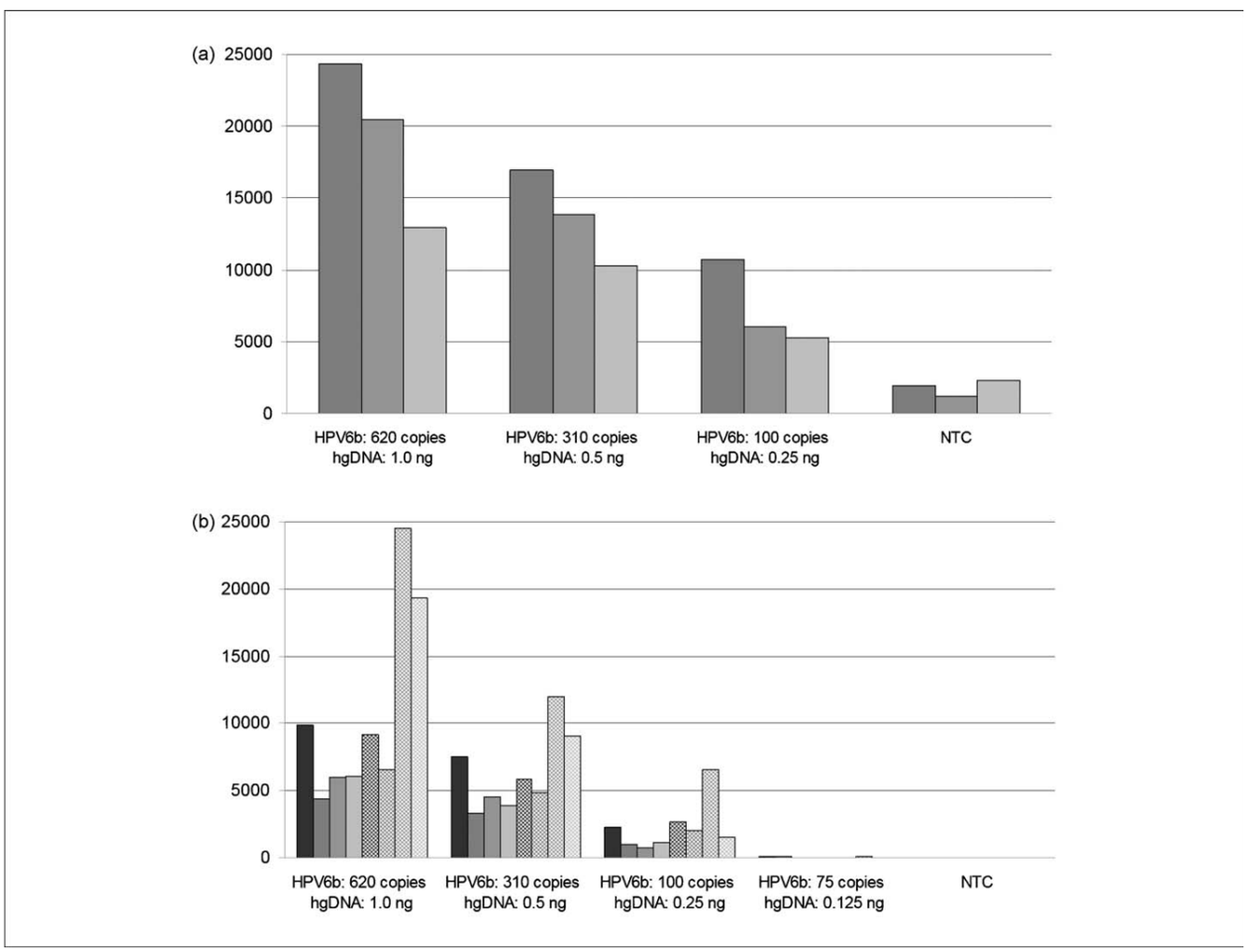


FIGURE 3

Fluorescence data of hybridization of PCR products from integrated amplification, labeling and selective detection of HPV subtype 6b and human β-globin. All PCRs were performed in the microfluidic channels of a Geniom biochip with primer pairs targeting HPV 6b and human β-globin with starting template amounts as depicted in the diagram. **(a)** Fluorescence data of probes specific for human β-globin. **(b)** Fluorescence data of probes specific for HPV 6b. NTC = Non-template control PCR.

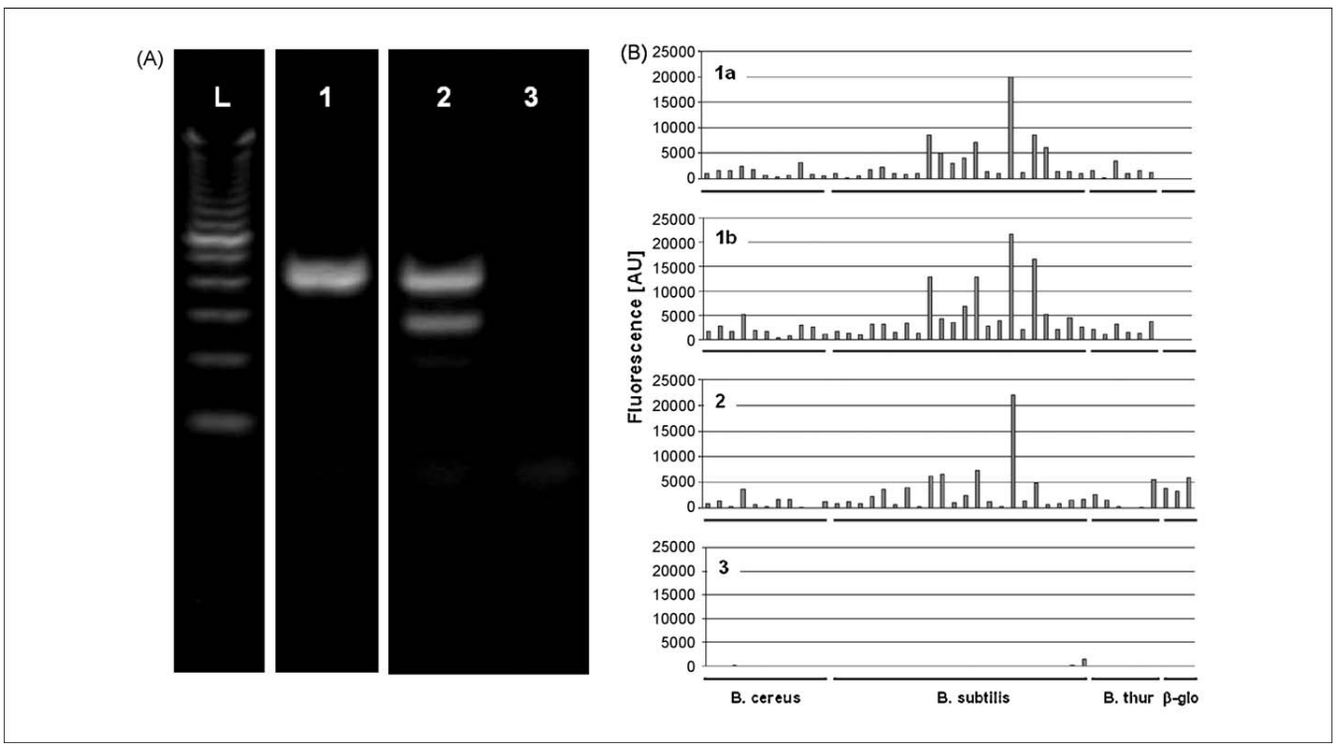


FIGURE 4

6-Plex PCR targeting different *Bacillus* species and human β-globin. **(a)** Agarose gel analysis of PCRs carried out in PCR tubes using six individual primer pairs and different genomic DNA templates. Expected amplicon lengths for targets are *Bacillus cereus*: 446 bp, *Bacillus subtilis*: 395 bp, *Bacillus thuringiensis ser. israelensis*: 407 bp, *human* β-globin: 300 bp. L = Ladder. 1: PCR conducted in the presence of genomic DNA of *Bacillus cereus*, *Bacillus subtilis* and *Bacillus thuringiensis ser. israelensis*. 2: PCR conducted in the presence of genomic DNA of *Bacillus cereus*, *Bacillus subtilis* and *Bacillus thuringiensis ser. israelensis* and human genomic DNA. 3: Control PCR conducted in the absence of template DNA. **(b)** Diagram of fluorescence intensities of different array capture probes obtained after integrated PCR, hybridization, fluorescence staining and detection in microchannels of a biochip. Identical PCR mixtures as described in A were used. 1a, b: Two replicates conducted under conditions as used in A, Lane 1. 2, 3: PCRs conducted under conditions as in A, lanes 2 and 3, respectively.

replicates exhibited similar fluorescence intensities (Fig. 4b, 1a,b). No fluorescence was observed for probes specific for human β-globin. For PCR in the presence of human genomic DNA (Fig. 4b, 2), fluorescence intensities seemed to differ slightly from PCRs in the absence of human gDNA. This might reflect an impact of the additional product formation on efficiency of PCRs targeting *Bacillus* species. However, probes correctly indicated the presence of individual *Bacillus*-related products. Additionally, the presence of the human β-globin PCR product was clearly indicated by the respective probes. These data show that PCRs involving multiple genomes and up to six primer pairs can be conducted within the microchannels and products can be selectively detected with excellent sensitivity.

In summary, we have developed a fully integrated system to combine all steps of a typical protocol for detection and typing of genomic targets from non-amplified samples. The method allows for semi-quantitative detection and typing of viral, pro- and eukaryotic targets covering different complexities with high sensitivity and selectivity. Owing to the fast cycles of microarray design and testing, the system can be quickly adapted to new genetic variants and provides a theoretical capacity for selective detection of >600 target nucleic acids when prototyping is employed in the way presented here. In contrast to previous studies that employ separated amplification and subtyping methods, often involving several further steps for labeling and purification, the highly automated workflow results in low hands-on-time and should be beneficial for reproducibility and contamination risk.

We are currently developing increasingly multiplexed detection arrays for both RNA and DNA targets and evaluate probe collections for detection and subtyping of various viral, as well as pro- and eukaryotic pathogens that should enable high throughput screening approaches for multiple disease types.

Acknowledgements

We thank Joanna Grigas for her contributions to the *Bacillus* studies. This work was funded by the Federal Ministry of Education and Research (BMBF) under contract 0315181.

References

1 Kaltenboeck, B. and Wang, C. (2005) Advances in real-time PCR: application to clinical laboratory diagnostics. *Adv. Clin. Chem.* 40, 219–259
2 Wittwer, C.T. *et al.* (2001) Real-time multiplex PCR assays. *Methods* 25, 430–442
3 Heid, C.A. *et al.* (1996) Real time quantitative PCR. *Genome Res.* 6, 986–994
4 Marras, S.A., Kramer, F.R. and Tyagi, S. (1999) Multiplex detection of single-nucleotide variations using molecular beacons. *Genet. Anal.* 14, 151–156

5 Solinas, A. *et al.* (2001) Duplex Scorpion primers in SNP analysis and FRET applications. *Nucleic Acids Res.* 29, E96
6 Nagai, H. *et al.* (2001) High-throughput PCR in silicon based microchamber array. *Biosens. Bioelectron.* 16, 1015–1019
7 Nagai, H. *et al.* (2001) Development of a microchamber array for picoliter PCR. *Anal. Chem.* 73, 1043–1047
8 Marcus, J.S., Anderson, W.F. and Quake, S.R. (2006) Parallel picoliter rt-PCR assays using microfluidics. *Anal. Chem.* 78, 956–958
9 Kalinina, O. *et al.* (1997) Nanoliter scale PCR with TaqMan detection. *Nucleic Acids Res.* 25, 1999–2004
10 Morrison, T. *et al.* (2006) Nanoliter high throughput quantitative PCR. *Nucleic Acids Res.* 34, e123
11 Brenan, C.J., Roberts, D. and Hurley, J. (2009) Nanoliter high-throughput PCR for DNA and RNA profiling. *Methods Mol. Biol.* 496, 161–174
12 Ottesen, E.A. *et al.* (2006) Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. *Science* 314, 1464–1467
13 Spurgeon, S.L., Jones, R.C. and Ramakrishnan, R. (2008) High throughput gene expression measurement with real time PCR in a microfluidic dynamic array. *PLoS One* 3, e1662
14 Weisenberger, D.J. *et al.* (2008) DNA methylation analysis by digital bisulfite genomic sequencing and digital MethyLight. *Nucleic Acids Res.* 36, 4689–4698
15 Spiro, A. and Lowe, M. (2002) Quantitation of DNA sequences in environmental PCR products by a multiplexed, bead-based method. *Appl. Environ. Microbiol.* 68, 1010–1013
16 Schmitt, M. *et al.* (2006) Bead-based multiplex genotyping of human papillomaviruses. *J. Clin. Microbiol.* 44, 504–512
17 Hoheisel, J.D. (2006) Microarray technology: beyond transcript profiling and genotype analysis. *Nat. Rev. Genet.* 7, 200–210
18 Albrecht, V. *et al.* (2006) Easy and fast detection and genotyping of high-risk human papillomavirus by dedicated DNA microarrays. *J. Virol. Methods* 137, 236–244
19 Iqbal, J. *et al.* (2008) Fabrication and evaluation of a sequence-specific oligonucleotide miniarray for molecular genotyping. *Indian J. Med. Microbiol.* 26, 13–20
20 Klaassen, C.H. *et al.* (2004) DNA microarray format for detection and subtyping of human papillomavirus. *J. Clin. Microbiol.* 42, 2152–2160
21 Uttamchandani, M. *et al.* (2009) Applications of microarrays in pathogen detection and biodefence. *Trends Biotechnol.* 27, 53–61
22 Call, D.R. (2005) Challenges and opportunities for pathogen detection using DNA microarrays. *Crit. Rev. Microbiol.* 31, 91–99
23 Shen, Y. and Wu, B.L. (2009) Microarray-based genomic DNA profiling technologies in clinical molecular diagnostics. *Clin. Chem.* 55, 659–669
24 Baum, M. *et al.* (2003) Validation of a novel, fully integrated and flexible microarray benchtop facility for gene expression profiling. *Nucleic Acids Res.* 31, e151
25 Chan, S.Y. *et al.* (1995) Analysis of genomic sequences of 95 papillomavirus types: uniting typing, phylogeny, and taxonomy. *J. Virol.* 69, 3074–3083
26 Bulkmans, N.W. *et al.* (2004) POBASCAM, a population-based randomized controlled trial for implementation of high-risk HPV testing in cervical screening: design, methods and baseline data of 44,102 women. *Int. J. Cancer* 110, 94–101
27 Chan, P.K. *et al.* (2006) Biases in human papillomavirus genotype prevalence assessment associated with commonly used consensus primers. *Int. J. Cancer* 118, 243–245
28 Gravitt, P.E. *et al.* (2000) Improved amplification of genital human papillomaviruses. *J. Clin. Microbiol.* 38, 357–361

nature biotechnology

High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing

Mark Matzas[1,5], Peer F Stähler[1,5], Nathalie Kefer[1], Nicole Siebelt[1], Valesca Boisguérin[1], Jack T Leonard[1], Andreas Keller[1], Cord F Stähler[1], Pamela Häberle[1], Baback Gharizadeh[2], Farbod Babrzadeh[2] & George M Church[3,4]

The construction of synthetic biological systems involving millions of nucleotides is limited by the lack of high-quality synthetic DNA. Consequently, the field requires advances in the accuracy and scale of chemical DNA synthesis and in the processing of longer DNA assembled from short fragments. Here we describe a highly parallel and miniaturized method, called megacloning, for obtaining high-quality DNA by using next-generation sequencing (NGS) technology as a preparative tool. We demonstrate our method by processing both chemically synthesized and microarray-derived DNA oligonucleotides with a robotic system for imaging and picking beads directly off of a high-throughput pyrosequencing platform. The method can reduce error rates by a factor of 500 compared to the starting oligonucleotide pool generated by microarray. We use DNA obtained by megacloning to assemble synthetic genes. In principle, millions of DNA fragments can be sequenced, characterized and sorted in a single megacloner run, enabling constructive biology up to the megabase scale.

Current *de novo* gene construction[1–4] rests on 1990's technology for chemical oligonucleotide synthesis, which is costly and has error rates of 1 in 300 base pairs (bp). Errors are typically avoided by manually selecting the best Sanger sequences using electrophoretic automation. Recent innovations in programmable array technology[5–8] offer the possibility to synthesize pools of thousands to millions of sequences per array with lengths comparable to conventional synthesis. The technology thus provides an extremely rich source of DNA oligonucleotides with great flexibility and superior efficiency regarding throughput and cost per bp. However, the error rate of microarray-derived oligonucleotides is typically higher compared to conventional synthesis, making error avoidance or correction necessary. Furthermore it is challenging to divide the derived oligonucleotide pools, containing vast amounts of species, into subpools—necessary, for example, to guide the assembly of synthetic genes, chromosomal regions or whole pathways in synthetic biology.

Megacloning turns NGS from a previously purely analytical method into a preparative tool, and represents a tremendous source of sequence-verified DNA where the yield from one NGS run is equivalent to that from hundreds to thousands of Sanger-sequence runs. It therefore addresses the challenge of error reduction for both conventional and microarray-derived DNA oligonucleotides. The method yields high-quality DNA libraries containing perfect parts with desired and correct sequences in adjustable ratios useful for a wide range of (bio-)technological applications.

Here we present a proof-of-concept study aimed at the retrieval of clonal DNA with known sequence from an NGS platform after sequencing (**Fig. 1**). The workflow comprises the input of DNA of short length, an NGS run to generate sequence-verified DNA clones, the identification of DNA with desired sequence on the sequencer's substrate and the retrieval of the clones of choice. The sources for the input DNA are for the most part independent of the megacloning step. For the present work, input DNA was derived from conventional oligonucleotide synthesis and from DNA microarrays. We used the NGS platform GS FLX from Roche 454 Life Sciences[9,10]. Owing to its open-top architecture, accessibility of the beads and the bead size, this platform is well suited for a pick-and-place approach using micropipettes to retrieve specific beads from the 454-Picotiterplate (PTP) and transfer them into conventional multi-well plates for further processing.

First, we established a technical setup for the controlled extraction of beads. The PTP at this stage contained a natural sample from human DNA, and extraction was done using a micropipette controlled by a microactuator device (**Supplementary Data**). To assess the fidelity of our setup, we compared the reads coming from the GS FLX platform with Sanger-derived sequences of DNA amplified from extracted beads. The alignment of Sanger sequences to the NGS reads matched 99.9%. Only two mismatches were obtained in 2,410 bp. Both were putative insertions in the GS FLX reads occurring at homopolymer stretches and therefore have a high likelihood of being platform-specific, base-calling artifacts[9] (**Supplementary Data**).

Next we collected a set of 319 beads with DNA clones from a microarray-derived pool initially containing 3,918 sequences. The beads for extraction were selected to ensure that their GS FLX reads perfectly matched sequences in our starting pool. The obtained DNA and the

[1]febit group, Heidelberg, Germany. [2]Stanford Genome Technology Center, Stanford University, Palo Alto, California, USA. [3]Harvard Medical School, Boston, Massachusetts, USA. [4]Wyss Institute for Biologically Inspired Engineering, Boston, Massachusetts, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to G.M.C. (gmc@harvard.edu).

Received 8 June; accepted 19 October; published online 28 November 2010; doi:10.1038/nbt.1710

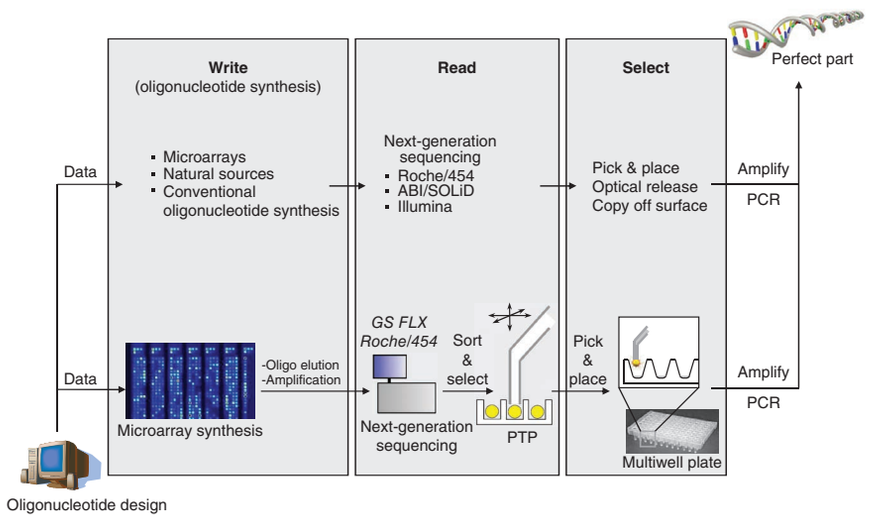


Figure 1 Coalescence of DNA reading and writing. The general approach begins with DNA from a variety of sources. Here we used oligonucleotides synthesized from microarrays as well as from conventional sources. Then, next-generation sequencing is used to read and identify oligonucleotides with desired sequences. Here we used the GS FLX platform (454/Roche). Finally, the DNA is sorted and retrieved selectively, in this case with a microactuator-controlled micropipette guided by two microscope cameras. The technologies used for retrieval depend on the sequencing platform.

untreated pool were compared after being sequenced independently on a Genome Analyzer II (Illumina GAII). We mapped 3.1% of reads from the initial (nonenriched) DNA pool without errors to the set of 319 selected sequences. In the enriched pool the fraction of reads mapping perfectly to the target sequences was 84.3%. The increase by a factor of 27.2 shows clearly a successful enrichment of selected and correct sequences (**Fig. 2a**,**b**). Also the analysis of reads on the level of single-target sequences shows that for 94% of the sequences in the selected pool, 50% or more of the reads were correct (**Fig. 2c**). Error-prone sequences contained a high number of different species likely to be caused by known sequence variations on the GAII, as reported previously[11].

To test the assembly of gene fragments based on megacloned oligonucleotides stemming from a microarray, we assembled two gene fragments, each ~220 bp in length, combining either nine or ten megacloned, bead-derived amplicons in a PCR-based gene assembly reaction[12,13]. The obtained assemblies were cloned and Sanger sequenced. Seven out of eight clones matched the target sequence perfectly. Interestingly, one clone showed insertions and deletions all located within a region 23 bp wide. Errors in assemblies originating from inaccuracies in the starting material could be expected to be distributed evenly over the entire construct. As this sequence was otherwise free of errors, these defects were likely caused by misassembly rather than errors in the building blocks used (**Supplementary Data**).

To further evaluate the capabilities of the megacloning approach to generate biologically functional genes, we applied the method to DNA fragments 274–394 bp in length and extracted 32 beads from the PTP carrying putatively correct sequences. These DNA fragments were the product of gene assembly reactions[12] using overlapping 40-mer oligonucleotides synthesized using conventional phosphoramidite chemistry and could be assembled into a model gene encoding β-D-glucuronidase (*uidA*)[14] (2,080 bp).

Three Sanger sequences obtained from the bead DNA were totally unrelated to the expected sequence and were probably caused by wrong bead extraction or contamination. The remaining 29 sequences covered 7,195 bp and matched without errors to the expected target sequences (**Supplementary Data**).

We then assembled the model gene out of nine DNA fragments from the set of 29 matching beads. The full-length gene construct was again checked by Sanger sequencing for absence of errors, and the biological functionality of the gene was tested in an enzymatic assay based on the conversion of X-Glc (5-bromo-4-chloro-3-indolyl-β-glucoside) substrate into blue dye[15] (**Supplementary Data**). Besides the proof of feasibility of generating biological functional genes, this experiment further mimics other applications of our technology, such as the use of sheared natural DNA and its subsequent sorting and reordering.

The absence of errors in 7,195 bp of DNA obtained from 29 extracted beads raised the question of achievable error rates from the megacloner process. Therefore we explored the potential of megacloning using a statistical model. This model considers two main sources of error—namely, wrong sequencing calls and polymerase errors during DNA amplification[16]. The calculations estimated the chance of finding one error in our extracted sequence space of ~7,200 bp to be 29%, which is in line with our experimental findings. The theoretical error rate of bead amplicons after megacloning using the setup employed in this study was estimated to be 1 error in 21 kbp (**Supplementary Data**). Compared with the error rate in the starting material of 1 error in 40 bp (determined from GAII data of the initial microarray pool), this equals a 500-fold error reduction.

We further calculated the expected amount of reads from NGS that match the target sequences of a given pool without errors. These numbers are crucial to estimate the complexity of pools that can be processed in one megacloner run. The resulting efficiency and cost structure are influenced mainly by three parameters: the error rate of the starting pool, the sequencing accuracy and the length of the variable sequence (**Supplementary Data**). With an error rate of 1 error in 40 bp and an average sequencing accuracy of 99.9% in the GS FLX, we expect a five- to tenfold cost reduction in producing DNA fragments (compared to conventional oligonucleotide synthesis) that can be achieved now with the prototype device (**Supplementary Data**). Because these fragments are largely free of errors, further savings can be expected in gene synthesis because the cost of subsequent sequencing for final quality control will be lower.

In this work we demonstrated the targeted retrieval of bead-bound DNA from a high-throughput sequencer without major modifications to the sequencing process. Previous methods for error correction in DNA pools[7,17–21] do not adequately handle collections of closely related oligonucleotide sequences that occur during assembly of repetitive sequences or multi-gene family libraries. They also do not enable hierarchical assembly strategies, which are made possible by the ordered selection and physical separation of clonal DNA described here.

The megacloner process has been proven to be useful for retrieval and sorting of correct and functional sequences and to increase the portion of error-free sequences in a sample substantially. This technology allows the processing of DNA from microarrays but also from a variety of other sources, such as conventional oligonucleotide synthesis or natural DNA fragments.

Megacloning could be optimized beyond the estimates in this work of one error in 21 kbp from input DNA having an error rate of 1 in 40 bp. Although such raw material can be obtained by state-of-the-art microarray technologies, the quality of input DNA could be increased further by addressing the amplification step of bead-bound DNA—for example, with higher fidelity polymerases, as the predicted contribution of the polymerase to the error rate is 4.7-fold higher than the expected error rate of the megacloner itself (**Supplementary Data**). Another accessible parameter for optimizing the overall process in terms of error rates is improvement in the quality of the DNA starting material. Also, optimization of sequencing accuracy could be a way to improve the ability to select correct parts after NGS. This is, however, the subject of ongoing optimization in the scope of NGS development, including ligase-based methods with improved accuracy[22].

The pool used in our conceptual study contained ~4,000 sequences. According to our results and extrapolations, this can be increased to ~30,000 sequences per pool with the described setup. As the bead extraction is generally independent of the pool complexity, it is mainly limited by the NGS platform and the quality of the starting material (**Supplementary Data**). More advanced microarray formats are able to deliver libraries with even higher complexity and of sufficient quality to fit into a gene assembly process[23]. Therefore, with an appropriate degree of automation that reaches an extraction frequency of two or three beads per minute, which is achievable with state-of-the-art robotics, the work-up of one PTP becomes possible within days, resulting in > 10^6 bp per plate. Hence, the downstream process (amplification, cleanup, assembly) will represent the next bottleneck.

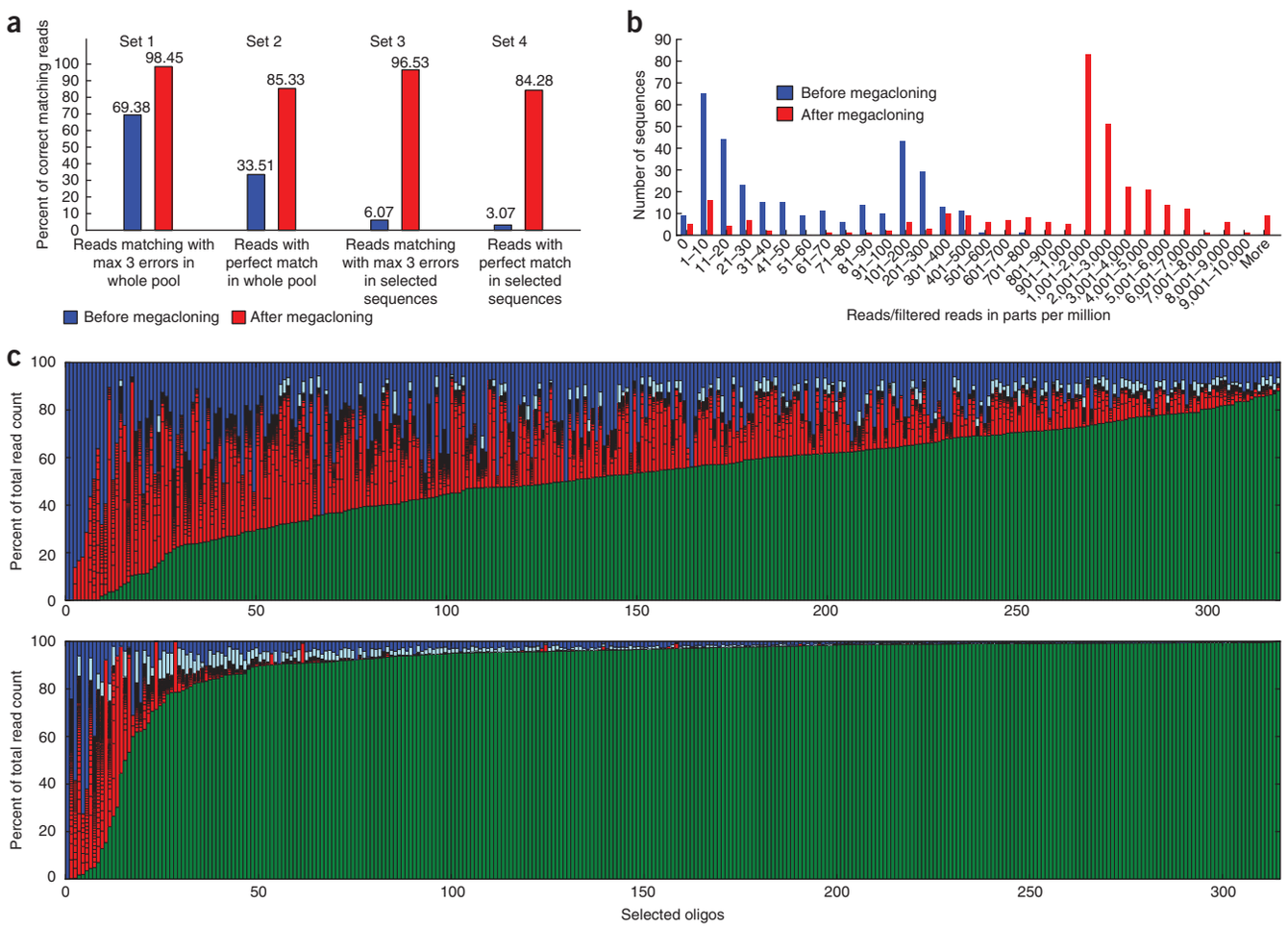


Figure 2 NGS-based comparison of untreated and megacloned oligonucleotide pools from microarray. (**a**) Comparison of the initial microarray oligonucleotide pool (blue) and the pool enriched with the megacloner technology (red) based on the results of the Illumina GAII runs. The bars in set 1 represent the fraction of reads that could be mapped allowing up to three errors. Bars in set 2 show the fractions of perfectly matching reads to the sequence set of the initial pool (3,918 sequences). The difference between the blue and the red bar in set 2 represents the enrichment of correct sequences by megacloning. The bars in set 3 and set 4 show the fractions of reads mapping to sequences from the selected pool of 319 sequences. The difference between blue and red bars in set 3 shows the enrichment of a selected 319 sequences before megacloning compared with after. Blue and red bars in set 4 represent the enrichment of sequences that are in the set of 319 selected sequences and that are correct. (**b**) Histogram of read counts in the Illumina GAII data of the initial pool (blue) and the enriched megacloned sample (red). Only reads mapping without errors to one of the 319 selected target sequences have been taken into account. To compare the two NGS runs on the basis of read counts, we converted the numbers into parts-per-million (p.p.m.) from the total number of filtered reads. (**c**) Composition of reads from the Illumina GAII data including 319 selected sequences in the initial pool (top) and the enriched pool (bottom). The oligonucleotides are sorted by the fraction of correct reads. Green, correct reads; red, error-prone reads (compartments in the red bars represent single sequences with a read count of 0.1% or more of total reads for the particular sequence); light blue, sum of nonunique error-prone reads where each sequence represents less than 0.1% of total reads for the particular sequence; blue, unique reads. In the Illumina GAII data set from the enriched sample, just 315 out of 319 selected sequences could be detected.

Our next focus in the present context is improvement and automation of physical bead extraction. The workflow used in this study still involved a considerable number of manual steps and some human intervention, which was identified as the most important source of error in terms of extraction of unwanted beads. Therefore, the success rate of ~90% (29 beads out of 32) has to be increased for the bead localization and retrieval process.

The method described here holds the potential to decrease production cost for synthetic DNA by one or more orders of magnitude. This source of high-quality DNA could aid the field of synthetic biology, as well as the production of libraries for antibodies or enzyme variants. In addition to synthetic sources, the sorting of natural DNA could enable the quick reconstruction or combination of DNA fragments to assemble genes, chromosomes or genomes, while simultaneously including synthetic parts of DNA.

The principle that we applied here using the GS FLX technology should also be generally applicable to other available NGS platforms such as Illumina's GAII, SOLiD, the Polonator or others. In the present context, the advantage of the GS FLX platform is the robot-accessible platform architecture and the comparably large size of the beads. Owing to different architectures of the other platforms, such as partially closed systems and substantially smaller DNA carriers, harvesting DNA from those will require a different mechanism, such as optical approaches including photosensitive and cleavable linker-molecules. The advantage of these platforms is a considerably higher number of DNA clones, which potentially could increase the capacity and throughput of the technology up to the gigabase level.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank B.A. Roe, F.Z. Najar and D.D. White for sequencing support, J. Jäger for technical consulting, and D. Summerer, T. Brefort, S. Kosuri and D. Levner for discussions and comments.

AUTHOR CONTRIBUTIONS

M.M., P.F.S. and G.M.C. conceptualized the megacloning method and wrote the manuscript; M.M. designed and lead the study, wrote all algorithms for sequence design, data analysis, image conversion, image processing and microactuator control; M.M., N.K., N.S. acquired the used technology, set up the microactuator device and optical systems; N.S. designed the *uidA* genetic model; M.M., N.K., N.S., V.B. and P.H. designed and optimized molecular biological methods; C.F.S. and J.T.L. contributed to bead picking and engineering concepts; A.K. set up the statistical models and calculations; J.T.L. contributed to the design of molecular biological steps and the acquisition of sequencing samples; B.G. and F.B. evaluated and implemented necessary changes into the sample preparation and the sequencing process on the 454/Roche platform.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Endy, D. Foundations for engineering biology. *Nature* **438**, 449–453 (2005).
2. Menzella, H.G. *et al.* Combinatorial polyketide biosynthesis by *de novo* design and rearrangement of modular polyketide synthase genes. *Nat. Biotechnol.* **23**, 1171–1176 (2005).
3. Gibson, D.G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–56 (2010).
4. Carr, P.A. & Church, G.M. Genome engineering. *Nat. Biotechnol.* **27**, 1151–1162 (2009).
5. Gao, X. *et al.* A flexible light-directed DNA chip synthesis gated by deprotection using solution photogenerated acids. *Nucleic Acids Res.* **29**, 4744–4750 (2001).
6. Singh-Gasson, S. *et al.* Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat. Biotechnol.* **17**, 974–978 (1999).
7. Tian, J. *et al.* Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* **432**, 1050–1054 (2004).
8. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
9. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
10. Wicker, T. *et al.* 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* **7**, 275 (2006).
11. Willenbrock, H. *et al.* Quantitative miRNA expression analysis: comparing microarrays with next-generation sequencing. *RNA* **15**, 2028–2034 (2009).
12. Stemmer, W.P., Crameri, A., Ha, K.D., Brennan, T.M. & Heyneker, H.L. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **164**, 49–53 (1995).
13. Richmond, K.E. *et al.* Amplification and assembly of chip-eluted DNA (AACED): a method for high-throughput gene synthesis. *Nucleic Acids Res.* **32**, 5011–5018 (2004).
14. Jefferson, R.A., Burgess, S.M. & Hirsh, D. beta-Glucuronidase from *Escherichia coli* as a gene-fusion marker. *Proc. Natl. Acad. Sci. USA* **83**, 8447–8451 (1986).
15. Couteaudier, Y., Daboussi, M.J., Eparvier, A., Langin, T. & Orcival, J. The GUS gene fusion system (*Escherichia coli* beta-D-glucuronidase gene), a useful tool in studies of root colonization by Fusarium oxysporum. *Appl. Environ. Microbiol.* **59**, 1767–1773 (1993).
16. Cline, J., Braman, J.C. & Hogrefe, H.H. PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res.* **24**, 3546–3551 (1996).
17. Carr, P.A. *et al.* Protein-mediated error correction for de novo DNA synthesis. *Nucleic Acids Res.* **32**, e162 (2004).
18. Smith, J. & Modrich, P. Removal of polymerase-produced mutant sequences from PCR products. *Proc. Natl. Acad. Sci. USA* **94**, 6847–6850 (1997).
19. Bang, D. & Church, G.M. Gene synthesis by circular assembly amplification. *Nat. Methods* **5**, 37–39 (2008).
20. Fuhrmann, M., Oertel, W., Berthold, P. & Hegemann, P. Removal of mismatched bases from synthetic genes by enzymatic mismatch cleavage. *Nucleic Acids Res.* **33**, e58 (2005).
21. Binkowski, B.F., Richmond, K.E., Kaysen, J., Sussman, M.R. & Belshaw, P.J. Correcting errors in synthetic DNA through consensus shuffling. *Nucleic Acids Res.* **33**, e55 (2005).
22. McKernan, K.J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).
23. Kosuri, S. *et al.* Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat. Biotechnol.* advance online publication, doi:10.1038/nbt.1716 (28 November 2010).

ONLINE METHODS

Oligo synthesis, sequence design, adaptors. Oligonucleotides used for this work were synthesized on programmable microarray synthesizers using light-directed synthesis methods[5]. Conventional oligonucleotides used for gene assembly were obtained from Sigma Aldrich. Harvesting of oligonucleotides from microarray surfaces was performed by chemical cleavage of succinate-ester bonds using ammonia hydrochloride solution.

Amplification of microarray-derived oligonucleotide pools by emulsion PCR. Microarray-derived oligonucleotide pools were amplified before NGS using emulsion PCR[24]. Therefore universal terminal sequences were attached during synthesis and served as primer regions. Amplification primers contained adaptors for sequencing on the Illumina GAII platform and/or the 454 GS FLX (**Supplementary Data**).

Sequencing on the 454 GS FLX. The sample preparation for the PCR-amplified oligonucleotides was done according to the manufacturer's protocols (Roche/454). To keep the DNA intact after sequencing, we exchanged the bleaching cleaning buffer with TE buffer before the sequencing run to avoid degradation of DNA during the final cleaning steps of the Roche sequencer.

Data analysis of 454 data and image conversion. NGS reads obtained from the GS FLX sequencer were aligned to the target sequences in the oligonucleotide pool to find the best matching sequence for every read and to perform further analysis, such as error rate estimation. Perfect matching sequences were selected and localized in the sequencer image by using the coordinates attached to every read sequence. For sequence data analysis, we used various Python scripts using the BioPython package. The images from the GS FLX sequencer were converted into the TIFF standard format using the Python Imaging Library.

Bead localization and extraction. After aligning the GS FLX reads to the set of target sequences, we selected reads that perfectly matched one of the desired oligonucleotide sequences in the pool. For localization of beads we located the corresponding chemiluminescent signals in the converted raw image from the GS FLX platform using the *x*- and *y*-coordinates that were included in the NGS raw data. To locate beads in the PTP, we identified reference points in the raw image and their corresponding positions in the PTP using suitable patterns of light signals. Based on these reference points the bead positions on the PTP were calculated using an algorithm for scaling and rotation. The extraction was performed with a micropipette with an outer diameter of 28 µm. For pipette handling we used a three-axis microactuator (**Supplementary Data**). Before extraction of beads the PTP was stored under a water layer to prevent desiccation and shrinking of beads. After picking, the beads were transferred immediately into a PCR vial and stored under water until further processing.

Amplification of DNA from beads. Amplification of bead-bound DNA was performed with the primers 454-A and 454-B, targeting the Roche/454 adaptors, or 'slx-fw-long' and 'slx-rev-long' for Illumina adaptors. For amplification of fragments with 40-mer variable regions, primers were 5′-biotinylated to facilitate subsequent removal of primer regions on a streptavidin matrix. PCR conditions: 20 mM Tris-HCl (pH 8.8), 10 mM ammonium-sulfate, 10 mM potassium chloride, 2 mM magnesium-sulfate, 0.1% Triton X-100, 200 µM each dNTP, 2% (vol/vol) DMSO, 1 µM each primer, 50 U/ml native pfu polymerase (Fermentas). Cycling: initial denaturation 96 °C (2 min); then 30 cycles of 96 °C (30 s), 63 °C (30 s), 72 °C (30 s) and final elongation 72 °C (3 min). After amplification, all PCR products were analyzed on PAGE (**Supplementary Data**) to check specificity and yield. For generation of the subpool containing 319 sequences, we estimated the concentration on the basis of the gel analysis and mixed the amplicons in equimolar concentrations.

Illumina sequencing and data analysis. As the sample contained suitable adaptors all steps regarding adaptor ligation have been omitted. All other steps were done according to the protocols from Illumina.

The NGS raw data obtained from Illumina GAII were processed by the following steps.

1. Truncation of reads to the length of the variable regions (40 bp).
2. Filtering out reads containing ambiguities (filtered reads).
3. Group reads with similar sequences (bins).

Subsequently for each read we identified the best matching target sequence from the oligonucleotide pool by mapping all reads to a pseudo-genome using rapid alignment of small RNA reads (razerS) (http://www.seqan.de/projects). The pseudo-genome was generated by concatenation of the variable parts of pool sequences separated by 40-mer poly-T stretches. The corresponding target sequence could then be determined by the matching position in the pseudo-genome. Alignments from the razerS output were used to determine insertions, deletions and substitutions. To compare the two GAII runs based on the number of correct reads, we converted the read counts into parts-per-million units (p.p.m.), taking the number of filtered reads before the matching procedure (after step 2) as a basis.

Assembly of gene fragments from conventional oligonucleotides. Gene fragments > 200 bp were assembled from conventionally synthesized 40-mer oligonucleotides having a constant overlap region of 20 nucleotides to the adjacent oligomer. Primer regions for 454 sequencing and restriction sites for primer removal were included during assembly. The assembly reaction contained 5 nM of each construction oligonucleotide and 200 nM of terminal primers. PCR conditions: 1× KOD polymerase buffer (Novagen), 1.25 mM $MgSO_4$, 40 µM each dNTP, 5 U/ml KOD Hot Start Polymerase (Novagen). Cycling for gene assembly: initial denaturation 96 °C (4 min); then 30 cycles of 96 °C (10 s), 55–40 °C touchdown (30 s), 72 °C (10 s). For subsequent amplification: 96 °C (10 s), 55 °C (30 s), 72° (30 s), final elongation 72 °C (3 min).

Assembly of genes from >200 bp fragments. Gene assembly up to 2 kbp were performed according to the protocol used for assembly of > 200 bp from oligonucleotides.

Primer removal and cleanup of bead amplicons before gene assembly. For removal of primer regions amplicons were incubated with LguI restriction endonuclease in 1× Tango buffer (Fermentas) for 3 h at 37 °C. For > 200 bp fragments, small restriction fragments containing primer regions were removed by PCR purification columns (GenElute PCR Clean-Up, Sigma Aldrich). For cleanup of microarray-derived fragments, we used 40-mer variable region biotinylated primers during bead DNA amplification and removed restriction products containing biotin residues using streptavidin matrix. The 40-mer fragments were ethanol precipitated and dissolved in water before further processing.

Assembly of genes from 40-mer double-stranded DNA fragments. For the assembly of genes from 40-mer dsDNA we used a two-stage assembly protocol including a primerless PCR followed by a PCR for amplification of the resulting products described previously[13].

24. Williams, R. *et al.* Amplification of complex gene libraries by emulsion PCR. *Nat. Methods* **3**, 545–550 (2006).

Toward the blood-borne miRNome of human diseases

Andreas Keller[1,2,21], Petra Leidinger[2,21], Andrea Bauer[3], Abdou ElSharawy[4], Jan Haas[5], Christina Backes[2], Anke Wendschlag[6], Nathalia Giese[7], Christine Tjaden[7], Katja Ott[7], Jens Werner[7], Thilo Hackert[7], Klemens Ruprecht[8], Hanno Huwer[9], Junko Huebers[10], Gunnar Jacobs[4], Philip Rosenstiel[4], Henrik Dommisch[11], Arne Schaefer[4], Joachim Müller-Quernheim[12], Bernd Wullich[13], Bastian Keck[13], Norbert Graf[14], Joerg Reichrath[15], Britta Vogel[5], Almut Nebel[4], Sven U Jager[16], Peer Staehler[6], Ioannis Amarantos[6], Valesca Boisguerin[6], Cord Staehler[6], Markus Beier[6], Matthias Scheffler[6], Markus W Büchler[7], Joerg Wischhusen[17,18], Sebastian F M Haeusler[17], Johannes Dietl[17], Sylvia Hofmann[4], Hans-Peter Lenhof[19], Stefan Schreiber[4,20], Hugo A Katus[5], Wolfgang Rottbauer[5], Benjamin Meder[5], Joerg D Hoheisel[3], Andre Franke[4,21] & Eckart Meese[2,21]

In a multicenter study, we determined the expression profiles of 863 microRNAs by array analysis of 454 blood samples from human individuals with different cancers or noncancer diseases, and validated this 'miRNome' by quantitative real-time PCR. We detected consistently deregulated profiles for all tested diseases; pathway analysis confirmed disease association of the respective microRNAs. We observed significant correlations ($P = 0.004$) between the genomic location of disease-associated genetic variants and deregulated microRNAs.

MicroRNAs (miRNAs) can regulate hundreds of genes post-transcriptionally and appear to regulate virtually all cellular processes. Owing to these properties, miRNAs have a critical role not only in physiological but also in pathological processes[1]. Although most reported miRNA expression profiles have been generated from solid tissues, there is growing evidence that miRNA profiles are readily accessible from body fluids, such as blood[2,3]. The aim of our multicenter study was to elucidate and compare blood expression profiles of 863 miRNAs for different human diseases to test for disease-specific alterations. The generated blood-based 'miRNome' data have been deposited in the Gene Expression Omnibus and updated versions are available at http://genetrail.bioinf.uni-sb.de/wholemirnomeproject/. We applied identical standardized experimental and biostatistical procedures to the 454 analyzed blood samples from individuals with lung cancer, prostate cancer, pancreatic ductal adenocarcinoma, melanoma, ovarian cancer, gastric tumors, Wilms tumor, pancreatic tumors, multiple sclerosis, chronic obstructive pulmonary disease (COPD), sarcoidosis, periodontitis, pancreatitis or acute myocardial infarction and from unaffected individuals (controls). All participating centers had to contribute samples to the control group (**Supplementary Table 1**). The different control cohorts had a high degree of reproducibility between the centers (**Supplementary Fig. 1**).

The platform we used is a highly specific primer extension–based microarray that shows a very small degree of cross-hybridization and can be used to distinguish between members of the *let-7* family[4]. To test for technical variance, we repeated the measurements on four samples (two blood samples and two tissue samples) and found a median correlation of 0.97. The correlation between different samples was significantly lower as shown by two-tailed unpaired Wilcoxon Mann-Whitney test ($P < 0.05$) (**Supplementary Fig. 2**). To estimate the biological variance, we analyzed blood samples taken from a healthy individual at three different time points during the day (9 a.m., 12 noon and 3 p.m.), with duplicate measurements at each time. Median correlation between the time points was 0.98 and between duplicates it was 0.99 (**Supplementary Fig. 3**).

On average, we found for each disease 103 deregulated miRNAs ($P < 0.05$; *t*-test after Benjamini-Hochberg adjustment). A total of 62 miRNAs (7.18% of all 863) were deregulated in at least six diseases in comparison to controls (**Supplementary Table 2**), and 24 miRNAs (2.78%) were deregulated in >50% of the 14 analyzed diseases. One miRNA (hsa-miR-320d) was deregulated in 11 diseases and three miRNAs (hsa-miR-423-5p, hsa-miR-146b-3p and hsa-miR-532-3p) were deregulated in nine of the tested

[1]Biomarker Discovery Center, Heidelberg, Germany. [2]Institute of Human Genetics, Saarland University, Medical Faculty, Homburg, Germany. [3]German Cancer Research Center, Functional Genome Analysis, Heidelberg, Germany. [4]Institute of Clinical Molecular Biology, Christian Albrechts University, Kiel, Germany. [5]Department of Internal Medicine, University of Heidelberg, Heidelberg, Germany. [6]febit group, Heidelberg, Germany. [7]Department of General Surgery, University of Heidelberg, Heidelberg, Germany. [8]Department of Neurology, Charité University Hospital, Berlin, Germany. [9]Department of Cardiothoracic Surgery, Voelklingen Heart Center, Voelklingen, Germany. [10]Department of Pneumology, Voelklingen Lung Center, Voelklingen, Germany. [11]Department of Periodontology, Operative and Preventive Dentistry, University Hospital Bonn, Bonn, Germany. [12]Department of Pneumology, University Hospital Medical Center, Freiburg, Germany. [13]Department of Urology, University Clinic, Friedrich-Alexander University, Erlangen-Nuremberg, Nuremberg, Germany. [14]Department of Pediatric Hematology and Oncology, Saarland University, Medical Faculty, Homburg, Germany. [15]Clinic for Dermatology, Venerology and Allergology, Saarland University, Medical Faculty, Homburg, Germany. [16]Praxis für Dermatologie, Sulzbach, Germany. [17]Department of Obstetrics and Gynecology, Medical School, University of Wuerzburg, Wuerzburg, Germany. [18]Interdisciplinary Center for Clinical Research, Junior Research Group 'Tumor progression and immune escape', Medical School, University of Wuerzburg, Wuerzburg, Germany. [19]Center For Bioinformatics, Saarland University, Saarbruecken, Germany. [20]First Medical Department, University Clinic Schleswig-Holstein, Kiel, Germany. [21]These authors contributed equally to this work. Correspondence should be addressed to A.K. (ack@bioinf.uni-sb.de).

RECEIVED 9 MAY; ACCEPTED 27 JULY; PUBLISHED ONLINE 4 SEPTEMBER 2011; DOI:10.1038/NMETH.1682

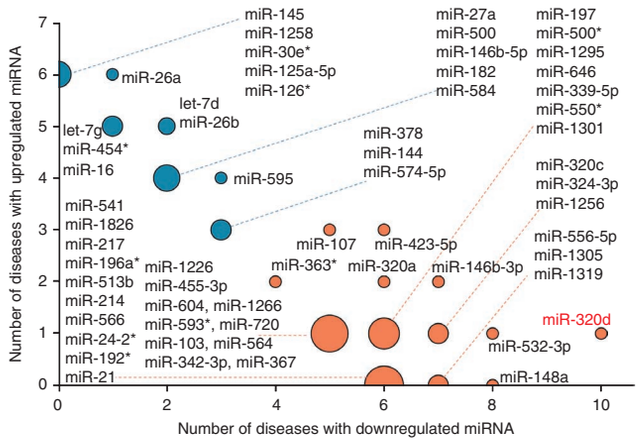


Figure 1 | Bubble plot of miRNAs that are up- or downregulated in several diseases. Bubble sizes correspond to the number of deregulated miRNAs. Orange bubbles denote miRNAs that are more often significantly down-regulated ($P < 0.05$) than upregulated. Blue bubbles denote miRNAs that are either more often upregulated or equally frequent up- and downregulated. *Homo sapiens* (hsa)-miR-320d was significantly deregulated ($P < 0.05$) in 11 diseases.

diseases. Known properties of these miRNAs are listed in **Supplementary Table 2**. Most miRNAs were consistently deregulated, that is, they were either up- or downregulated in the majority of diseases (**Fig. 1**). Analysis of the human microRNA disease database[5] revealed that only a few of the miRNAs deregulated in blood were also previously reported as deregulated in solid tissues derived from individuals with the same diseases (**Supplementary Table 3**). A total of 121 miRNAs (14%) were not deregulated in any of the 14 analyzed diseases.

We carried out pathway analysis of putative target genes for miRNAs that were deregulated in at least six of 14 diseases ($n = 62$) and for miRNAs that were not deregulated in any disease ($n = 121$). We extracted the targets with $P < 0.001$ for both miRNA sets using GeneTrail[6,7]. We found a total of 7,598 target genes for both miRNA sets. Of these genes, 27% were targets of miRNAs in both sets, 21% were targets of miRNAs that were frequently deregulated and 52% were targets of miRNAs that were not deregulated in our study. We applied an over-representation analysis relying on the hypergeometric distribution using GeneTrail to find significantly enriched ($P < 0.05$) biochemical pathways. For the set of frequently deregulated miRNAs, we found several disease-associated pathways (**Supplementary Table 4**) including 'pathways in cancer'. We did not detect any enriched pathway for the target genes of the 121 miRNAs that were not significantly deregulated in any disease. Pathways with significantly fewer ($P < 0.05$) targets than expected are indicated in **Supplementary Table 4**.

To explore whether the significantly deregulated miRNAs are in close genomic physical proximity to known susceptibility variants, we extracted 3,495 published single-nucleotide polymorphisms (SNPs) from the US National Institutes of Health genome-wide association study catalog (accessed 28 July 2010) and searched for the coding sequence of miRNAs in a genomic window of 250 kilobases (kb) around these SNPs. We detected 241 cases of physical proximity between SNPs and miRNAs. Of these, seven were related to diseases included in our study, representing interesting candidates for testing the hypothesis that miRNA deregulation depends on nearby genetic variants. Of the seven SNPs, four are associated with heart diseases, including cardiac structure and function (rs7910620) and mean platelet volume (rs2393967, rs10914144 and rs10506328), two with multiple sclerosis (rs703842 and rs17445836) and one with melanoma. Notably, the relevant miRNA was significantly deregulated ($P < 0.05$) in the same disease, in six of the seven cases. To test whether these results could occur by chance, we carried out 10^6 non-parametric permutation tests. The proximity of genetic variants and deregulated miRNAs was significant ($P = 0.004$). All pairs of SNPs and adjacent miRNAs are summarized in **Supplementary Table 5** and one representative example is presented in **Figure 2**.

To distinguish individuals with disease from controls or from individuals with other diseases by miRNA profiling, we applied machine-learning techniques. Each of the 14 diseases was separated from controls with an average accuracy of 88.5%, ranging from at least 81.3% to up to 100% (**Supplementary Table 6**). By using only two miRNAs, we obtained an average accuracy of 72.5%, whereas the use of ten miRNAs resulted in an average accuracy of 80.6% ($P = 0.0002$, two-tailed unpaired Wilcoxon Mann-Whitney test) (**Supplementary Fig. 4**). Next, we performed pair-wise classification analyses between different diseases using samples collected at the same site to exclude between-institution bias. For the separation between pancreatic cancer and other pancreatic diseases, the accuracy was not significant ($P > 0.05$). However, this result does not necessarily imply a general similarity between miRNA profiles of malignant and nonmalignant diseases of the same organ. For example, we could distinguish lung cancer from COPD with an accuracy of 91.7%, corresponding to a highly significant classification ($P < 10^{-6}$). COPD is a common co-morbidity of lung cancer and also precedes tumors in 50–90% of cases[8]. Thus, a biomarker separating individuals with lung cancer from those with COPD but without cancer may prove useful.

We performed an independent validation of the miRNA profiles using different technologies and cohorts of individuals. In previous studies, we had confirmed 474 deregulated miRNAs in different diseases by performing quantitative real-time PCR (qRT-PCR) on samples from several individuals with lung cancer, melanoma, glioma and acute myocardial infarction[8–11]. Here we additionally performed a large-scale validation for a larger dataset including data for 44 individuals with lung cancer and 41 with COPD. We selected 18 significantly deregulated ($P < 0.05$) miRNAs that separate both diseases in quadruplicate by qRT-PCR using the SmartChip

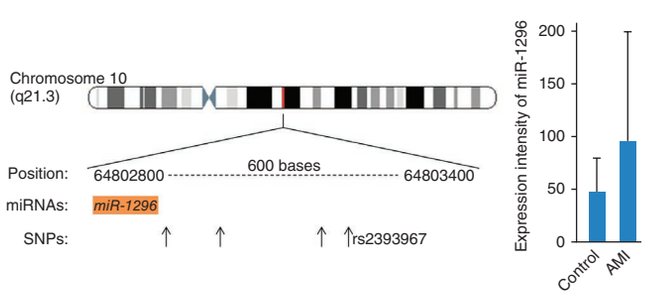


Figure 2 | Representative example for the physical proximity of a significantly deregulated miRNA and a known SNP. A schematic of the human chromosome 10q21 with *hsa-miR-1296* (magenta) and four SNPs (arrows) including SNP rs2393967 (SNP database (dbSNP) accession number) that is associated with heart diseases. The plot shows expression and s.d. of hsa-miR-1296 in the blood of individuals with acute myocardial infarction (AMI, $n = 20$) compared to that in healthy controls ($n = 70$). $P = 0.006$.

Real-Time PCR System (WaferGen Biosystems). Of those 18 miRNAs, we validated 14, that is, these miRNAs were deregulated in a comparable manner in array and qRT-PCR experiments. The remaining four miRNAs were only rarely expressed as indicated by mean threshold cycle (Ct) values >28.5. In **Supplementary Table 7** we list raw qRT-PCR data and the variance for the replicates. The overall correlation of the quantile normalized qRT-PCR and array results for the 45 analyzed miRNAs (27 miRNAs of previous studies and 18 miRNAs in the present study) was as high as 0.86 (**Supplementary Fig. 5**). We provide scatter plots and fold changes for all tested miRNAs (**Supplementary Table 8**).

We developed the concept of disease probability plots (DPPs) to determine the probability that a miRNA expression profile correctly indicates that an individual has one or several of the analyzed diseases. We computed the probabilities via a regression approach for each individual sample. Analyzing all DPPs, we predicted the correct disease in 67.45% of all individuals (exemplary DPPs are available in **Supplementary Fig. 6**). Assuming that all diseases are almost equally frequent in our dataset, this translates into an over eightfold increased accuracy of disease prediction by miRNA profiling as compared to random guessing.

Although our study supports the idea that blood cells have an miRNA pattern that varies between different diseases, there are several points to be considered when blood miRNA patterns are associated with diseases. Any association between a miRNA pattern and a disease can be confounded by co-morbidity for another disease. Furthermore, blood cells may not contribute equally to an miRNA pattern, with expression variation in a few cell types accounting for most of the pattern. Indeed, as recently shown for 27 different cell populations isolated from normal mouse hematopoietic tissues, different blood cell types have specific miNA expression patterns[12]. Distribution of the complete blood count (CBC) is known to vary in disease, for instance owing to cancers or diseases of the blood[13] or bone marrow, cancers that spread to the bone marrow, autoimmune disease or side effects of medications. There are also variations in CBC in healthy individuals. It is possible that changes in miRNA profile in disease reflect shifts in the distribution of different blood-cell types. We tested this possibility using principal-component analysis; specifically, we carried out standard principal-component analysis on the expression matrix (http://genetrail.bioinf.uni-sb.de/wholemirnomeproject/) and computed for each principal component the fraction of the overall data variance. Although it is likely that shifts in cell populations affect the overall miRNA profiles, we observed that even 27 different cell populations, represented by the first 27 principal components with highest variance, can account for only about 60% of the total variance in the miRNA profiles. Taken together, the ability to recognize systematic features in human blood cells and the relatively small normal CBC variation in healthy individuals provides support for the feasibility of using miRNA expression patterns in peripheral blood as the basis for detection of disease[13].

Identifying the complex relationships between disease and changes in miRNA expression patterns in blood cells could contribute not only to an understanding of the mechanism behind the pattern and of disease associations but provide insight into the pathological processes because miRNAs in turn influence the expression of thousands of genes.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Accession codes. Gene Expression Omnibus: GSE31568.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank F. Flachsbart, B. Noack and B. Loos for support. This work was financially supported by the German Ministry of Research Education (Bundesministerium für Bildung und Forschung 01EX0806), Hedwig Stalter Foundation, Homburger Forschungsförderungsprogramm and by the Deutsche Forschungsgemeinschaft (LE2783/1-1). Infrastructure support was received from the Deutsche Forschungsgemeinschaft cluster of excellence 'Inflammation at Interfaces'.

AUTHOR CONTRIBUTIONS

A.K. initiated the study; E.M., P.R., J.M-Q., A.B., P.S., V.B., C.S., M.B., M.W.B., J.W., S.F.M.H., J.D., S.S., H.A.K., W.R., B.M., J.D.H. and A.F. designed the study; A.K., P.L., A.E., H.A.K., W.R., B.M., J.D.H., A.F., E.M., S.S. and B.V. wrote the manuscript; A.K., J.H., C.B., A.W., I.A., B.V. and H-P.L. analyzed data; P.L., A.B., C.T., A.E., N.G., K.O., J.W., T.H., G.J., H.D., A.S., B.W., B.K., N.G., A.N., V.B., B.V., S.H. and B.M. performed experiments; C.T., K.O., T.H., K.R., H.H., J.H., G.J., H.D., A.S., B.W., B.K., J.R., S.U.J., N.G., M.S., M.W.B., J.W. and S.F.M.H. collected samples.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Taft, R.J., Pang, K.C., Mercer, T.R., Dinger, M. & Mattick, J.S. *J. Pathol.* **220**, 126–139 (2010).
2. Mitchell, P.S. *et al. Proc. Natl. Acad. Sci. USA* **105**, 10513–10518 (2008).
3. Otaegui, D. *et al. PLoS ONE* **4**, e6309 (2009).
4. Vorwerk, S. *et al. New Biotechnol.* **25**, 142–149 (2008).
5. Lu, M. *et al. PLoS ONE* **3**, e3420 (2008).
6. Backes, C. *et al. Nucleic Acids Res.* **35**, W186–W192 (2007).
7. Backes, C., Meese, E., Lenhof, H.P. & Keller, A. *Nucleic Acids Res.* **38**, 4476–4486 (2010).
8. Keller, A. *et al. BMC Cancer* **9**, 353 (2009).
9. Leidinger, P. *et al. BMC Cancer* **10**, 262 (2010).
10. Meder, B. *et al. Basic Res. Cardiol.* **106**, 13–23 (2011).
11. Roth, P. *et al. J. Neurochem.* **18**, 449–457 (2011).
12. Petriv, O.I., Hansen, C.L., Humphries, R.K. & Kuchenbauer, F. *Cell Cycle* **10**, 2–3 (2011).
13. Garzon, R. *et al. Blood* **111**, 3183–3189 (2008).

ONLINE METHODS

Blood samples. The blood samples were collected and processed from five different institutions working closely together with the Heidelberg Biomarker Discovery Center (http://www.bdc-heidelberg.com/biomarker-discovery/index.cfm). The participating centers were the German Cancer Research Center (Deutsches Krebsforschungszentrum), Saarland University, Heidelberg University, Kiel University and Wuerzburg University. Groups at each of these centers provided samples from individuals with disease and from healthy individuals. Blood was extracted using PAXgene Blood RNA tubes (BD).

All blood donors participating in this study gave their informed consent. A complete list of screened samples is provided in **Supplementary Table 1**.

miRNA extraction and microarray screening. A total of 2.5 ml to 5 ml of blood were extracted in PAXgene Blood RNA tubes. The PAXgene Blood RNA tubes ensure stabilization of RNA and hence stabilization of the expression profiles. Blood cells were obtained by centrifugation at 5,000*g* for 10 min at room temperature (18–25 °C). The miRNeasy kit (Qiagen) was used to isolate total RNA including miRNA from the resuspended blood cell pellet according to the manufacturer's instructions. The eluted RNA was stored at −70 °C.

All samples were shipped overnight on dry ice and analyzed with the fully automated Geniom RT Analyzer (febit biomed) at febit's in-house genomic service department using the Geniom Biochip miRNA *Homo sapiens* version v12 to v14. Geniom biochips consist of a meandering microchannel that forms the so-called 'biochip'. Each biochip can be used to analyze eight different samples independently. The flexible oligomer synthesis is done *in situ* inside the microchannels using a light-directed process. The probes were designed as the reverse complements of the mature miRNA sequences as published in miRBase plus nucleotides at the 5′-end of the capture oligonucleotide as needed for the enzymatic extension (microfluidic primer extension assay; MPEA). For conventional miRNA hybridization assays the reverse complement of the miRNA sequences as published in the miRBase releases version 12.0 to 14.0 (ref. 14) (in total 863 mature miRNAs and miRNA star sequences) were synthesized with seven intraarray replicates[4].

We mixed 250 ng of total RNA with 1 µl of 5 pM miRNA spike-in mix and dried it in a tabletop speedvac (Univapo 100H). Each RNA pellet was fully resuspended in 25 µl of hybridization buffer and denatured for 3 min at 95 °C. Until the hybridization, the denatured samples were kept on ice. Microarray hybridization was performed using the Geniom RT Analyzer and Geniom miRNA biochips *Homo sapiens*. The samples were loaded automatically and hybridization of unlabeled sample has been carried out for 16 h. On-chip sample labeling with biotin was carried out by MPEA[4]. Therefore, streptavidin R-phycoerythrin conjugate (SAPE) solution, antibody solution, equilibration buffer (1× NEB 2; New England Biolabs), stop buffer (6× SSPE; Applied Biosystems) and enzyme solution were placed into the RT Analyzer. The array equilibration was followed by incubation with enzyme solution. Enzyme incubation was stopped with stop buffer. SAPE staining, signal amplification and detection proceeded fully automated within the Geniom RT Analyzer. All steps from sample loading to miRNA detection were processed fully automatic inside the machine. As internal control standards five different probes labeled with Cy3 or biotin (bio) were included: 5′-[Cy3]TCACTCATGGTTATGGCAGCACTGC-3′ (80 nM), 5′-[bio]GTAGTTCGCCAGTTAATAGTTTGCG-3′ (12 nM), 5′-[bio]TCTTACCGCTGTTGAGATCCAGTTC-3′ (4 nM), 5′-[bio]CCCACTCGTGCACCCAACTGATCTT-3′ (0.4 nM) and 5′-[bio]CCATCCAGTCTATTAATTGTTGCCG-3′ (0.04 nM).

The enzymatic MPEA together with the fully automated handling ensured a high degree of specificity as well as excellent reproducibility.

The detection pictures were evaluated using the Geniom Wizard Software. For each feature, the median signal intensity was calculated. Following a background correction step, the median of the seven replicates of each miRNA was computed. To normalize the data across different arrays, quantile normalization[15] was applied and all subsequent analyses were carried out using the normalized and background subtracted intensity values. Since the miRBase has been upgraded twice in the past year from version 12.0 to version 14, we used for the final data analysis the 863 miRNAs that were consistently present in all three versions. The whole miRNome data are available for download from the project homepage (http://genetrail.bioinf.uni-sb.de/wholemirnomeproject/) and in the Gene Expression Omnibus[16].

Statistical analysis. Single miRNA analyses were carried out using *t*-tests (unpaired, two-tailed) after verifying approximate normal distribution using Shapiro-Wilk test. The resulting *P* values were adjusted for multiple testing using Benjamini-Hochberg's adjustment[17]. In addition, the area under the receiver characteristic curve was computed.

Supervised classification of samples was carried out using support vector machines (SVM)[18] as implemented in the R e1071 package[19]. As parameters of the SVM, we evaluated different kernel methods including linear, polynomial (degree 2 to 5), sigmoid and radial basis function kernels.

To detect miRNAs that contribute most diagnostic information and thus lead to accurate classifications, a subset selection technique has been applied. Specifically, an iterative filter approach based on the *t*-test was carried out. In each iteration, the *s* miRNAs with lowest *P* values were computed on the training set in each fold of a standard tenfold cross-validation, where *s* was sampled in regular intervals between 2 and 500 miRNAs. The respective subset was used to train the SVM and to carry out the prediction of the test samples in the cross validation. To compute probabilities for classes, a regression approach based on the output of the support vectors has been applied. To test for overtraining, nonparametric permutation tests were applied. All computations were carried out using the publicly available R statistical language[19].

To evaluate the classification, we computed accuracy, specificity and sensitivity.

Pathway analysis. To detect biochemical networks that are putatively regulated by disease miRNAs, we carried out a so-called overrepresentation analysis. For a set of miRNAs, we extracted the targets using Genetrail (http://genetrail.bioinf.uni-sb.de/) via MicroCosm V5 (http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/) that uses the miRanda algorithm. To reduce the number of false positive miRNA targets, we applied a significance value threshold of 0.001 (ref. 6). The set of putative mRNA targets of disease relevant miRNAs was used as input for the web-based gene set analysis tool GeneTrail to find Kyoto Encyclopedia

doi:10.1038/nmeth.1682

of Genes and Genomes (KEGG) pathways that are significantly enriched with targets of disease relevant miRNAs[20]. All significance values were corrected for multiple testing by Benjamini-Hochberg adjustment.

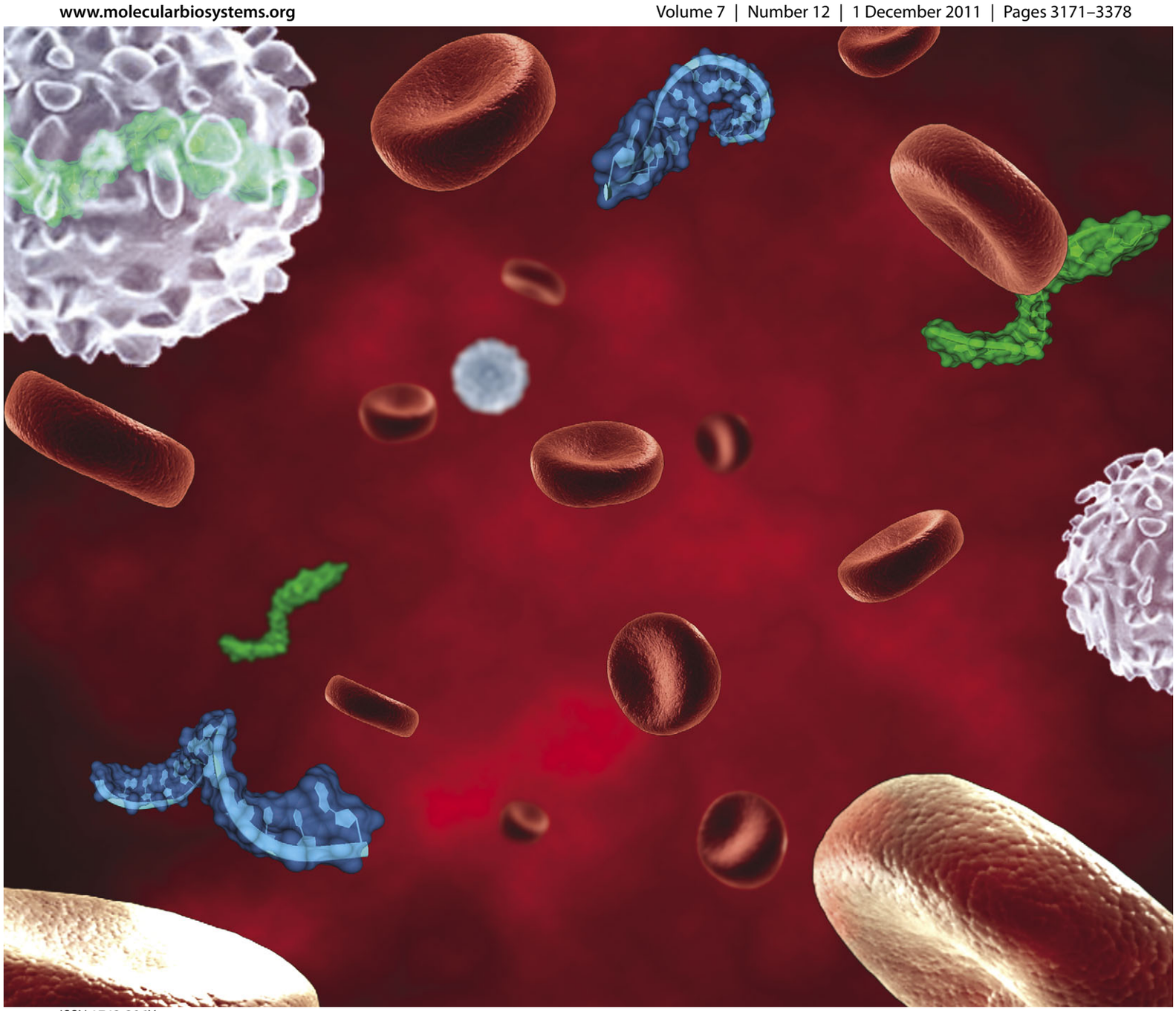
14. Griffiths-Jones, S. *Methods Mol. Biol.* **342**, 129–138 (2006).
15. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. *Bioinformatics* **19**, 185–193 (2003).
16. Edgar, R., Domrachev, M. & Lash, A.E. *Nucleic Acids Res.* **30**, 207–210 (2002).
17. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. *Behav. Brain Res.* **125**, 279–284 (2001).
18. Vapnik, V. *The Nature of Statistical Learning Theory*. 2nd edn. (Springer, New York, 2000).
19. Team, R. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2008).
20. Kanehisa, M. & Goto, S. *Nucleic Acids Res* **28**, 27–30 (2000).

Indexed in MEDLINE!

Molecular BioSystems

www.molecularbiosystems.org

Volume 7 | Number 12 | 1 December 2011 | Pages 3171–3378



ISSN 1742-206X

RSC Publishing

PAPER

Eckart Meese *et al.*

Next-generation sequencing identifies novel microRNAs in peripheral blood of lung cancer patients

1742-206X(2011)7:12;1-1

Molecular BioSystems

Dynamic Article Links

Cite this: *Mol. BioSyst.*, 2011, **7**, 3187–3199

www.rsc.org/molecularbiosystems

PAPER

Next-generation sequencing identifies novel microRNAs in peripheral blood of lung cancer patients†

Andreas Keller,‡[abc] Christina Backes,‡[ab] Petra Leidinger,[b] Nathalie Kefer,[a] Valesca Boisguerin,[a] Catalin Barbacioru,[d] Britta Vogel,[e] Mark Matzas,[a] Hanno Huwer,[f] Hugo A. Katus,[e] Cord Stähler,[ac] Benjamin Meder§[e] and Eckart Meese§*[b]

Received 29th August 2011, Accepted 10th October 2011
DOI: 10.1039/c1mb05353a

MicroRNAs (miRNAs) are increasingly envisaged as biomarkers for various tumor and non-tumor diseases. MiRNA biomarker identification is, as of now, mostly performed in a candidate approach, limiting discovery to annotated miRNAs and ignoring unknown ones with potential diagnostic value. Here, we applied high-throughput SOLiD transcriptome sequencing of miRNAs expressed in human peripheral blood of patients with lung cancer. We developed a bioinformatics pipeline to generate profiles of miRNA markers and to detect novel miRNAs with diagnostic information. Applying our approach, we detected 76 previously unknown miRNAs and 41 novel mature forms of known precursors. In addition, we identified 32 annotated and seven unknown miRNAs that were significantly altered in cancer patients. These results demonstrate that deep sequencing of small RNAs bears high potential to quantify miRNAs in peripheral blood and to identify previously unknown miRNAs serving as biomarker for lung cancer.

Introduction

For many human diseases there is still a lack of peripheral biomarkers for efficient disease detection, therapy monitoring, and estimation of prognosis. Especially in patients with lung cancer, timely diagnosis and early specific treatment is crucial to improve patients' individual outcome. This is often difficult since today's diagnostic procedures only allow comparatively late diagnosis and hence treatment. Novel biomarkers for lung cancers, regardless of the underlying histological differences, that allow specific discrimination between patients from healthy individuals could markedly improve clinical care.

MiRNAs regulate a manifold of biological processes through negative regulation of gene expression. This reveals their high potential to influence almost every—physiological or pathophysiological—molecular pathway. Recent evidence also suggests that miRNAs impact on the development of human diseases including cancer. Most recently, miRNAs were furthermore recognized as promising non-invasive biomarkers for diverse human disorders.[1–6]

While array-based technologies or quantitative real-time PCR (qRT-PCR) have commonly been used to characterize the annotated human miRNome known at the time of these studies, next-generation sequencing (NGS) approaches now offer the option of getting an even deeper understanding of miRNA profiles in human diseases. However, only a few studies examined miRNA profiles in human blood or other body fluids including serum and plasma by NGS. Most of the published NGS studies focus on the analysis of already known miRNAs but less on the identification of novel miRNAs. For example, for non-small cell lung cancer (NSCLC) a four-miRNA serum signature was identified using Solexa sequencing.[7] In addition, miRNA signatures derived from serum of patients with esophageal squamous cell carcinoma and gastric cancer were identified by high-throughput sequencing.[8,9] Furthermore, a 13-miRNA-based biomarker was identified that discriminates between HBV cases from controls and HCV cases, and also HBV-positive hepatocellular carcinoma cases from controls and HBV cases.[10] Ge *et al.*[11] revealed the potential of NGS of miRNAs circulating in maternal plasma for non-invasive prenatal diagnostics and Luo *et al.*[12] already identified placenta-specific miRNAs in pregnant women. One study that provides novel miRNA was from Vacz *et al.* who predicted 370 novel miRNAs in PBMC of healthy individuals.[13]

[a] *Biomarker Discovery Center Heidelberg, Germany*
[b] *Department of Human Genetics, Saarland University, Homburg, Germany. E-mail: hgemee@uks.eu; Tel: +49-(0)6841-1626038*
[c] *Siemens Healthcare, Strategy, Erlangen, Germany*
[d] *Life Technologies, Foster City, USA*
[e] *Department of Internal Medicine III, University of Heidelberg, Heidelberg, Germany*
[f] *Department of Cardiothoracic Surgery, Voelklingen Heart Center, Germany*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c1mb05353a

‡ Authors contributed equally as first authors.

§ Authors contributed equally as senior authors.

Here, we performed NGS of small RNAs in human peripheral blood of patients with lung cancer and of healthy control individuals. By using SOLiD sequencing technology and DNA barcoding we generated over 25 million sequencing reads per sample and identified numerous known and novel miRNAs specific for lung cancer. The results of this study are integrated in the "Whole Disease miRNome" project,[6] which aims to improve our understanding of the human miRNome in a wide range of human pathogenic processes.

Materials and methods

Study population

For the NGS approach, we obtained whole blood samples from ten patients with non-small cell lung cancer (NSCLC) and ten healthy individuals (Table 1). We collected 2.5–5 ml whole blood in PAXgene™ Blood RNA tubes (PreAnalytiX) and stored the samples at −20 °C until extraction of total RNA. Lung cancer patients and healthy individuals showed a non-significant difference in gender distribution (Fishers Exact test *p*-value of 0.36).

For quantitative real time PCR (qRT-PCR) we obtained lung cancer tissue from 16 different patients. Lung cancer tissue samples were stored at −80 °C after resection until RNA isolation. We combined the isolated RNA from those tissues to four pools, *i.e.*, one squamous cell lung cancer pool, one adenocarcinoma pool, one large cell lung cancer pool, and one small cell lung cancer pool.

Considering the ethnic groups, all individuals were Caucasians with except of one Persian among the healthy blood donors. The enclosed lung cancer patients did not undergo any radio- or chemotherapy before blood drawing and tumor resection. All tumor patients were smokers or former smokers with 7 to 80 pack years.

Local ethics committee has approved the analysis of blood and tissue from patients and controls and participants have given their informed consent.

Isolation of total RNA from blood cells and tissue

The RNA isolation of the PAXgene™ Blood RNA Tubes was performed as previously described.[4] The RNA was stored at −70 °C until use. For the isolation of RNA from tissue, samples were homogenized in 2 ml QIAzol lysis reagent and incubated for 5 min at RT. Then 200 µl chloroform were added, vortexed for 15 s, and incubated for 2–3 min at RT. Subsequently, we followed the same protocol as applied for blood.

Library preparation

1.5 µg of total RNA was enriched for the fraction of small RNAs (10–40 nt) using Ambion's flashPAGE Fractionator, followed by sodium acetate precipitation. SOLiD internal adapters were ligated using 100 ng enriched fraction. After ligation, smallRNAs were transcribed into cDNA with Reverse Transcriptase. cDNA fragments between 60 and 80 nt (small RNAs + adaptors) were isolated from a 10% TBE Urea Gel (Novex-System, Invitrogen). RNA from gel slices was amplified with 15 PCR cycles using the same 5′-Primer for each sample and ten different 3′-Primers including the barcode sequences (SOLiD Multiplexing Barcoding Kit 01-16). A total of ten purified and barcoded DNA libraries was analyzed with a HS-DNA Chip in the Agilent Bioanalyzer 2100 and subsequently pooled in equimolar amounts.

Next generation sequencing

The pooled libraries were diluted to a concentration of 41 pg μl^{-1}. DNA was amplified monoclonally on magnetic beads in an emulsion PCR. Emulsions were broken with butanol and the remaining oil was washed off the templated double-stranded beads. DNA on the bead surface was denatured to allow hybridization of the enrichment beads to the single stranded DNA. Using a glycerol cushion the null beads can be separated from the templated beads. After centrifugation, the enriched magnetic beads were in the supernatant. The enrichment-beads were separated from the magnetic beads by denaturation. The 3′-end was enzymatically modified for deposition on the

Table 1 Characteristics of blood donors

Sample	Age	Gender	Tumor classification	TNM classification	Clinical staging	Therapy	Ethnicity
Lung cancer 715	76	Male	Squamous cell lung cancer	T2bN1	IIB	No	Caucasian
Lung cancer 721	57	Male	Squamous cell lung cancer	T3N0	IIB	No	Caucasian
Lung cancer 731	71	Male	Lung adenocarcinoma	T1bN0	IA	No	Caucasian
Lung cancer 735	65	Female	Squamous cell lung cancer	T2bN1	IIB	No	Caucasian
Lung cancer 739	59	Female	Lung adenocarcinoma	T3N0	IIB	No	Caucasian
Lung cancer 742	67	Male	Squamous cell lung cancer	T2aN1	IIA	No	Caucasian
Lung cancer 744	56	Male	Lung adenocarcinoma	T2aN2	IIIA	No	Caucasian
Lung cancer 746	72	Male	Lung adenocarcinoma	T2aN1	IIA	No	Caucasian
Lung cancer 747	61	Male	Lung adenocarcinoma	T3N1	IIIA	No	Caucasian
Lung cancer 748	69	Female	Squamous cell lung cancer	T2aN0	IB	No	Caucasian
Control 1	30	Female	Healthy	—	—	—	Caucasian
Control 2	53	Male	Healthy	—	—	—	Caucasian
Control 3	25	Female	Healthy	—	—	—	Caucasian
Control 4	29	Male	Healthy	—	—	—	Caucasian
Control 5	29	Female	Healthy	—	—	—	Caucasian
Control 6	60	Male	Healthy	—	—	—	Caucasian
Control 7	43	Female	Healthy	—	—	—	Caucasian
Control 8	36	Male	Healthy	—	—	—	Caucasian
Control 9	51	Female	Healthy	—	—	—	Caucasian
Control 10	29	Female	Healthy	—	—	—	Persian

sequencing slide. 700 Million Beads were loaded onto a Full Slide and sequenced on a SOLiD 4 analyzer.

Mapping of reads

Mapping of SOLiD sequencing reads against known miRNAs and the genome was done using the RNA2MAP tool (version 0.5) from Applied Biosystems (http://solidsoftwaretools.com/gf/project/rna2map/). To use the default parameters of this mapping pipeline, we first trimmed the reads to a size of 35 nt. To reduce the overhead of computation, we reduced the amount of reads per sample to those being unique in the sample. The RNA2MAP pipeline included three steps: (1) reads are filtered against tRNAs, rRNAs, and other repetitive elements; (2) the remaining reads are mapped against the predicted precursor sequences of miRNAs from miRBase (version 16[14–16]); (3) the remaining reads are mapped against the human genome (hg19). The mapped genome reads served as input for the prediction of novel miRNAs with miRDeep.[17] The predicted novel miRNA precursor sequences were added to the precursor sequences of miRBase and step 2 of the RNA2MAP pipeline was repeated to retrieve the counts for both the known and novel predicted precursor sequences.

Prediction of novel miRNAs

For the prediction of novel miRNAs, we used a probabilistic model of miRNA biogenesis in combination with the frequency of RNA reads along the secondary structure of the miRNA precursor as implemented in miRDeep.[17] Previously, we transformed the output of the alignments of RNA2MAP to the so-called 'blastparsed' format of miRDeep. To this end, we removed the sequencing adaptor, converted the colorspace mapping into bases, re-counted the mismatches, adjusted the alignment length, and computed a bit score and an *E*-value as described previously.[18] The miRDeep pipeline itself was run with default parameters using Randfold (v 2.0,[19]) and a fasta file containing the mature miRNA sequences from miRBase v16 (without human sequences) to improve accuracy and sensitivity. To reduce the number of false positive predictions, we ran 100 permutation tests and excluded a predicted novel miRNA if found in any of the permutation runs. The remaining putative novel miRNAs (p-value < 0.01) were mapped with BLAST (v 2.2.24[20]), against known ncRNA and miRNA sequences from diverse sources (miRBase v16, snoRNA-LBME-db[21]), ncRNAs from Ensembl "Homo_sapiens.GRCh37.59.ncrna.fa" ((ftp://ftp.ensembl.org/pub/release-59/fasta/homo_sapiens/ncrna/) NONCODE v 2.0[22]). We excluded sequences that aligned with more than 90% of their length (allowing 1 mismatch) to any of the ncRNA sequences.

Distribution of miRNA reads across the miRNA precursors

Since we performed a size selection we do not intent to measure the expression level of the miRNA-precursor but of the mature miRNAs. The mapping of mature miRNA reads to the respective precursor sequence, however, offers the option to understand how the mature miRNA reads distribute along the precursor. To consider the distribution of reads mapping to a miRNA precursor, we computed for each precursor separately the coverage of each base position for lung cancer samples and controls. Likewise, we also computed for each base position of each precursor a significance value using the Wilcoxon Mann–Whitney (WMW) test.

Downstream analysis

To further evaluate the NGS miRNA profiles, we carried out statistical computations using R.[23] The Shapiro–Wilk test has been applied to determine whether miRNA counts across all samples are normally distributed. To normalize samples standard quantile normalization has been applied to make the different sequencing runs comparable to each other. Expression of a miRNA i in a sample j has been measured as the normalized read count of this miRNA in the respective sample. The Grubbs test has been carried out for detecting outliers. The non-parametric WMW test has been performed for detecting differentially regulated miRNAs. To further assess the validity of the signature we carried out non-parametric permutation tests. Here, the class labels have been randomly shuffled 100 times and the same analyses as for the original class labels have been carried out. A p-value was computed as the fraction of random runs with a likewise significant result as the original computations.

In addition to WMW analysis, we performed an analysis considering the total length of a miRNA precursor to identify possible novel miRNAs that derive from this precursor. In detail, we computed for each precursor m and each base i the WMW significance value for the respective position in the precursor at position i, testing the hypothesis that read counts of miRNA m at position i are significantly higher for lung cancer samples as compared for normal controls. For each miRNA precursor, we then counted the number of bases with WMW significance values <0.05. Furthermore, the area under the receiver operator characteristics (AUC) curve has been computed for each miRNA. Cluster analysis has been done using the 'hclust' package.

For computing targets of deregulated miRNAs, the miRANDA algorithm has been applied and only miRNA–mRNA relations with p-values < 0.0001 have been considered.[24] To carry out gene set enrichment of target genes, we used GeneTrail and carried out a so-called over representation analysis.[25,26]

Validation of miRNA expression by qRT-PCR

To verify the accuracy of NGS-based miRNA quantification, expression levels of the newly identified miRNAs hsa-can-miR-49, hsa-can-miR-1040, hsa-can-miR-675, hsa-can-miR-213, hsa-can-miR-915, and hsa-miR-98_new were assessed using qRT-PCR (measured in duplicates) according to manufacturer's instructions (ABI, USA). We performed qRT-PCR with the RNA of the 20 blood samples from ten lung cancer patients and ten healthy individuals and with the RNA of the four lung cancer tissue pools.

qRT-PCR was done as follows: custom miRNA primers were synthesized by Qiagen (Hilden, Germany). The small nuclear RNA RNU6B-2 served as endogenous control. Hsa-miR-577 served as negative control and hsa-let-7g served as positive control. These two control miRNAs have been selected from our array-based study including 454 blood samples.[6] While hsa-let-7g

Table 2 Sequencing reads

Sample	Total reads	Unique reads	Mappable reads	Uniquely mapped reads	Uniquely mapped reads without mismatches to miRNAs
Control 1	21 546 906	6 483 422	15 622 637	8 702 194	2 983 422
Control 2	25 780 926	9 347 407	16 942 190	13 006 925	1 777 253
Control 3	27 351 543	8 199 034	19 697 784	12 352 288	1 626 120
Control 4	26 575 164	8 512 058	18 039 852	12 548 571	1 572 865
Control 5	25 621 021	9 831 745	16 609 732	15 823 914	1 256 992
Control 6	21 508 347	7 430 551	14 738 961	11 323 006	1 399 058
Control 7	21 667 199	6 770 030	15 103 061	11 940 299	1 097 079
Control 8	26 375 514	10 592 366	16 959 721	17 492 668	1 198 169
Control 9	26 510 814	9 508 500	16 289 720	6 923 836	1 803 276
Control 10	19 342 152	4 736 279	14 178 971	4 909 246	2 863 519
Lung cancer 715	39 838 662	12 057 004	28 320 865	18 976 323	2 219 591
Lung cancer 721	20 553 924	7 382 608	13 931 434	10 072 355	1 982 624
Lung cancer 731	29 427 176	10 086 103	20 025 354	16 223 858	1 837 469
Lung cancer 735	25 970 295	9 902 069	16 703 580	13 628 112	1 938 176
Lung cancer 739	17 517 290	5 864 764	12 274 764	8 527 749	1 693 297
Lung cancer 742	41 063 105	16 480 873	24 454 710	24 251 687	1 259 950
Lung cancer 744	16 378 241	6 369 296	10 235 069	8 255 511	1 645 217
Lung cancer 746	42 718 688	14 847 395	27 930 159	19 550 059	3 062 906
Lung cancer 747	18 571 825	6 795 712	12 133 853	9 606 735	1 432 632
Lung cancer 748	34 928 862	13 648 517	22 178 040	16 980 397	3 099 004
SUM	529 247 654	18 4845 733	352 370 457	261 095 733	37 748 619
Average	25 875 610	8 929 732	16 656 656	12 450 429	1 735 275
Std Dev	7 691 588	3 001 406	4 935 673	4 897 476	586 987

was highly expressed in most samples in this study, hsa-miR-577 was one of the lowest expressed miRNAs across this comprehensive cohort.

miRNA target prediction and functional analysis

For the prediction of targets of the newly identified miRNAs, we applied the miRanda algorithm (version 3.3a)[27] with default parameters to 3′ UTR sequences downloaded from the UCSC Table Browser.[28] The predicted mRNA targets were tested for functional enrichments using GeneTrail.[26]

Specificity of uniquely mapping reads and comparison to miRBase

Although the first step of RNA2MAP is a filtering step to remove reads that map against tRNAs, rRNAs, and other repetitive elements, we wanted to verify that potential reads that can be mapped to the mature forms of our novel miRNAs do not map to other noncoding RNAs. Therefore, we downloaded noncoding RNA sequences from Ensembl (http://www.ensembl.org/info/data/ftp/index.html), as well as mRNA exon sequences and intergenic sequences from the UCSC Table Browser.[28] To compare the results to those of already annotated miRNAs, we carried out the same analysis for all known miRNAs from the miRBase (v17). The mature forms of our novel miRNAs and the known miRNAs were used in a BLAST analysis against the downloaded fasta sequences. We extracted the novel/miRBase miRNAs, where at least one mature form matched without a mismatch with at least 90% of its length.

Results

High-throughput transcriptome sequencing results in high coverage of the human miRNome

We sequenced the small RNA fraction of 20 blood samples including ten samples of lung cancer patients and ten samples of healthy individuals. Details on the samples including tumor type and clinical staging are provided in Table 1. In total, we obtained 530 million reads including 185 million unique reads for the 20 samples. Of all reads, 352 millions were mappable to the human genome including 38 millions that were mappable to human miRNAs known at the time of this study (miRBase v16) without any mismatch. The prediction of novel miRNAs was based on the 352 million reads. All read counts of the 20 blood samples are summarized in Table 2 and presented as a bar-chart in Fig. 1. By using the uniquely mapped reads we detected 770 known miRNAs and known miRNA precursors representing 64% of the known human miRNome (miRBase v16).

Detection of novel miRNAs expressed in peripheral blood

For the prediction of novel miRNAs, we applied a probabilistic model of miRNA biogenesis that considers the frequency of

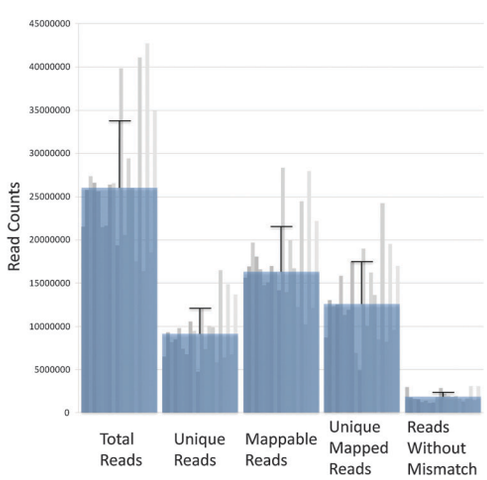


Fig. 1 Mapping statistics. The blue bars show the average of the mappable read counts of all 20 analyzed blood samples together with the respective standard deviations. The grey bars indicate the average of the mappable read counts of the 20 single blood samples.

RNA reads along the secondary structure of putative miRNA precursors. Initially, we detected 1081 putative novel miRNA precursor sequences. Next, we carried out two filter steps to reduce the number of false positives. First, we performed 100 permutation tests, eliminating 520 (48.1%) of the initially identified putative miRNA sequences. We blasted the remaining sequences against different data collections of small non-coding RNAs and found 351 (32.5%) sequences with at least one hit with already annotated small RNAs under the condition that one mismatch was allowed. After eliminating those 351 miRNAs, we obtained 210 putative novel miRNA sequences (unknown at the time of the study according to miRBase v16). To verify the specificity of putative reads mapping to the mature forms of the 210 novel miRNAs, we performed a BLAST analysis with the sequences of the mature forms against other non-coding RNA sequences (rRNA, tRNA, miRNA, snRNA, snoRNA, lincRNA), as well as mRNA and intergenic sequences. The same analysis was done for all known miRNAs from miRBase v17. As presented in Fig. 2, our newly identified miRNAs and the known miRNAs from miRBase showed a very similar distribution. Most of the miRNAs (55% of known miRNAs and 53% of newly identified miRNAs) did not map against any of the other RNA resources or intergenic regions. We found hits in intergenic regions for 38% of known and 43% of novel miRNAs. In both groups, 4% of miRNAs matched against mRNAs, while mapping against other non-coding RNA regions was insignificant.

Out of the 210 putative novel miRNA sequences, 30 miRNAs were identified with at least 25 reads that mapped uniquely to a precursor. As summarized in Table 3, each of these 30 miRNA sequences was detected in at least two blood samples, and two sequences were found in all 20 blood samples. On average, putative new miRNA sequences were detected in 16 out of 20 samples. Of the 30 putative novel miRNAs, four miRNAs have now been included in the recent miRBase release v17 and are highlighted in Table 3.

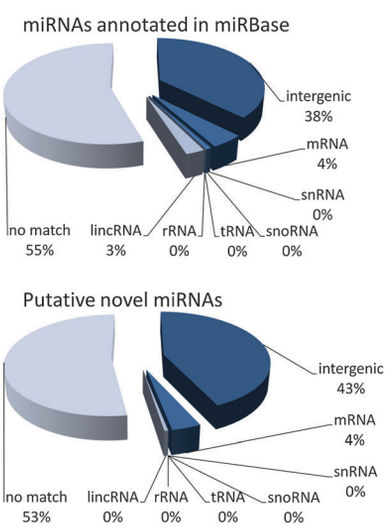


Fig. 2 Results of the BLAST analysis. We mapped the mature forms of miRBase v17 and our novel miRNAs to different groups of noncoding RNAs, mRNA, and intergenic sequences. The pie chart indicates the numbers of miRNAs mapping to the respective nucleic acid groups.

For all novel miRNAs as well as known miRNAs we computed a histogram plot (Fig. 3). While the median read count for miRNAs annotated in miRBase v16 was 24 reads, the median read count of novel miRNAs was still 18 reads per miRNA. For both, known and novel miRNAs the highest proportion of miRNAs lies in the range of up to 50 reads per miRNA. Considering all miRNAs covered by up to 150 reads the novel miRNAs were more frequent than the known miRNAs, providing evidence that the identified miRNA candidates are detected at a substantial level. The histogram plots for all single samples that essentially validate the general picture are provided in Fig. 1 (ESI†).

Next, we randomly selected five miRNA sequences from the newly identified 30 miRNAs that were identified with at least 25 reads that mapped uniquely to a precursor and carried out a qRT-PCR analysis. The qRT-PCR was performed with the ten different blood samples of healthy controls, the ten blood samples of lung cancer patients, and four pools of different types of lung cancer tumor tissues to compare the abundance of the respective miRNAs in blood and tumor tissue of lung cancer patients. In addition to these five miRNAs we also tested one miRNA as positive control and one miRNA as negative control. Based on previous array-based experiments[6,29] we selected miRNA hsa-let-7g that has usually been highly expressed in our previous experiments as positive control and miRNA hsa-miR-577 that has usually not been expressed as negative control. The ΔC_t-values of all novel miRNAs measured in blood samples fall in between the positive and negative control ΔC_t-values as shown in Fig. 4. The comparison between tissue samples and blood samples of lung cancer patients showed higher expression of all five miRNAs in the cancer blood samples as indicated by lower ΔC_t-values. In detail, the expression of three miRNAs (hsa-can-miR-1040, hsa-can-miR-675, and hsa-can-miR-915) was significantly lower in lung cancer tissue as compared to patients' blood and in two cases (hsa-can-miR-49 and hsa-can-miR-213) almost not detectable in lung cancer tissue as indicated by ΔC_t-values of 20 and 25. Notably, the latter miRNAs are more than one million less abundant in tissue than in blood of the tumor patients. The comparison between blood of lung cancer patients and blood of controls revealed higher expression of all five miRNAs in blood of cancer patients providing further evidence for an increase of specific miRNAs in blood of lung cancer patients. In summary, our qRT-PCR experiments validated the high-throughput sequencing experiments very well.

Functional target analysis of putative miRNAs

For known miRNAs validated and putative targets are known in the literature. These also show an enrichment in functional categories, *e.g.*, KEGG pathways.[30] As described in the "dictionary on microRNAs and their putative target pathways"[31] common targets of miRNAs are mRNAs involved in regulatory pathways as the "p53 signaling pathway" or the "TGF-beta signaling pathways" and disease related categories as "Pathways in cancer". As described in the Materials and methods section we carried out a search for putative targets of the novel putative miRNAs. The respective target gene set

Table 3 miRNA sequences with at least 25 reads that mapped uniquely to precursors. All analyses were performed using miRBase release v16. During the publication process miRBase v17 was released. Overlaps of previously unknown novel miRNAs (according to miRBase v16) with the recent miRBase v17 are indicated in bold and the official names and sequences are given in brackets

miRNA	Number of blood samples	Counts in blood of controls	Counts in blood of patients	Major sequence	Minor sequence
hsa-can-miR-163	20	4370	3189	TCGCATTGAACCTGAGAGGCA	CCTCCGGTATTCAAGCGATT
hsa-can-miR-277 (hsa-miR-4707)	18	514	493	GCCCGCCCCAGCCGAGGTT (hsa-miR-4707-3p: AGCCCGCCCCAGCCGAGGUUCU)	CCCCGGCGCGGGCGGGTTC (hsa-miR-4707-5p: GCCCCGGCGCGGGCGGGUUCUGG)
hsa-can-miR-811	20	688	262	GGGCCGTGGAGGTGGACTG	GTGCACAACTGCAGGGGTGTG
hsa-can-miR-915	19	64	86	CTCTTCATCTACCCCCCAG	GGAGGGTGTGGAAGACAT
hsa-can-miR-49 (hsa-miR-4659a)	18	53	91	CGTTGCCATGTCTAAGAAGAA (hsa-miR-4659a-5p: CUGCCAUGUCUAAGAAGAAAAC)	CTTCTTAGACATGGCAGCTTC (hsa-miR-4659a-3p: UUUCUUCUUAGACAUGGCAACG)
hsa-can-miR-473	19	49	60	GTCAGTTTGTCAAACTCTTT	GGAGTTGTGATCCTTTGGAGA
hsa-can-miR-571	17	27	74	CGCAACCCACACACGGTCTCA	AGACCGTGTGTGGGTTGCTGAG
hsa-can-miR-346	18	25	70	TTGGAATCCTCGCTAGAGCGT	GCTCTAGCGGGGATTCCAATA
hsa-can-miR-675	18	49	27	CCACAAACCTGCCAGCCCTG	GGGCGGCTATTGTGGGG
hsa-can-miR-275	16	46	30	TGGGTGTGGGCAGTGGGCGGGC CAAGGACA	GCAGTTGGCACCGTCCCCTGCG CCTACCCACT
hsa-can-miR-385	11	60	5	GGCGGGCAGCGGGTGAGGGGGTGG	GCGGGGCCCCGGACAAGGGT CCGCAGA
hsa-can-miR-213	17	28	33	TGCTCTTACATCTCAAACGAT	CGGTTGAGATGCAAGGGCTGC
hsa-can-miR-881	8	48	10	GCCCCTTTCTCAGACCCCCA	GGCCCTGGAAAGGGTCAG
hsa-can-miR-358	17	19	32	GCCCAGAGGATCACGGAGCCA	GCTCCTTGCACCTGTGGCTGC
hsa-can-miR-480	2	1	47	CTAGCAGTCTCAGGACACA	TGCCCTGAGACTGCTAAGT
hsa-can-miR-56	16	20	25	ATCACCACCAAACCTGTTCTTC	AGAACAGGTTTGGTGGGGATTC
hsa-can-miR-1040	17	20	19	GATTTCAGCGCTCTGCCCCT	GGGCAGAGCACTGTGTGTGG
hsa-can-miR-288 (hsa-miR-4688)	15	13	20	GGGGCAGCAGAGGACCTGGGC (hsa-miR-4688:UAGGGGCAGCAGAGGACCUGGG)	CCTGATCCTCAGCTGCCCTCTC
hsa-can-miR-1011	17	17	15	GTCTTTTGCCCTTTCAGCT	CTGGAAGGGCAAAAGACTG
hsa-can-miR-839	16	14	16	GTGCCTGTGCAGAGGGAGCT	CCCCCTCCGAGCAGGCACTG
hsa-can-miR-1065 (hsa-miR-4701-5p)	14	10	19	TTGGCCACCACACCTACCCCTT (hsa-miR-4701-5p: UUGGCCACCACACCUACCCCUU)	GGGTGATGGGTGTGGTGTCCACAGG (hsa-miR-4701-3p: AUGGGUGAUGGGUGUGGUGU)
hsa-can-miR-454	14	4	24	CCACCTTCAAAGGCACTCCG	GAGGCCTCTGCTGGTGCTG
hsa-can-miR-390	14	17	11	TCCTCTCCTCCCTGTGCCGAC	AAGCGCGGGGAGGGAGGATA
hsa-can-miR-23	11	14	14	ACCCACCTGATGCCCCGTCCCA	GGGAGGGGCAGGAGGGGTGGAATG
hsa-can-miR-152	3	25	2	CCTCTTCCCAGCACTCCCCT	GAGGGTTGCGGGAAGGGGGA
hsa-can-miR-555	16	15	11	AAAACAGGATAGGCACTAAA	TAGAGCCTATCCTGTTTTGC
hsa-can-miR-678	15	7	19	CGGTCCCTAACCCCCTCCGGA	CAGGGGAGGGAAGGGGAGCCGAG
hsa-can-miR-963	11	19	7	AGAAATTGGTTAAATTGGAGGG	GACCAATTTAACCAATTACTAT
hsa-can-miR-942	11	18	7	CTCTCCCCGCTTTTAACCCTA	GGGTTAAGAGTGGGGAGAAGA
hsa-can-miR-308	9	17	8	ACACCAAAACAATGAAAAC	TATCATTGTTTTAGTGTTT

has then been used as input for the gene set analysis tool GeneTrail[26] and compared with all human genes. We detected a total of 59 KEGG pathways being significantly enriched for targets of our miRNA candidates following multiple testing adjustment by the Benjamini Hochberg approach. The most significant pathways include "Metabolic pathways" (p-value of 7.88×10^{-9}) and "Pathways in cancer" (p-value of 1.1×10^{-6}) being already described as target pathways of known miRNAs.[31]

Besides these functional categories we also tried to identify known miRNAs that show a significant overlap in their target genes as compared to the novel miRNAs. Here, we found 39 known miRNAs that showed an enriched overlap in their targets and the targets of our novel miRNAs. These most prominent known miRNAs were hsa-miR-29c and hsa-miR-30c with significance values of 3.07×10^{-3} in both cases. The complete results are provided in Table S1 (ESI†).

miRNA biomarker signatures predict lung cancer

To identify possible disease discriminating miRNA signatures, we first performed unsupervised hierarchical clustering of miRNAs derived from all samples. Therefore, we excluded noisy and extremely abundant features and two clear outliers (Grubbs test p-value < 0.01). Using the Euclidian distance measure we identified separate clustering of lung cancer blood samples and control samples ($p = 0.00025$). Only one control blood sample clustered together with the ten lung cancer samples. This result was confirmed by a principle component analysis. The convex hulls of the first and second principle component of lung cancer samples and controls do not show any overlap (Fig. 2, ESI†).

Next, we determined and quantified differentially regulated single miRNAs. Here, we focused on the mature miRNAs but also tested whether the miRNA was significantly deregulated at all. To this end, we considered the expression of the precursor to be the sum of reads mapping to the precursor. The most abundant miRNA was hsa-miR-223 with a total of 8.6 million uniquely mapped reads in all 20 blood samples. The second most abundant miRNA was hsa-miR-425 with 0.7 million reads. The tenth most abundant miRNA, hsa-miR-339-5p, shows only 87 000 reads which are two orders of

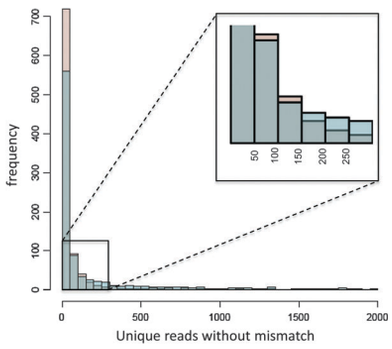


Fig. 3 Read frequencies for the known and the novel miRNAs. The histogram plot shows for each known and putative novel miRNA the frequency of unique read counts without mismatch. The rose shaded boxes indicate the putative novel miRNAs, the blue shaded boxes indicate the known miRNAs, and the green boxes indicate the overlap between both. In both cases, *i.e.*, novel and known miRNAs, highest proportion of miRNAs can be found in the area between 1 and 50 reads. In an intermediate range between 51 and 150 reads per miRNA the novel miRNAs are more frequent while in higher ranges the known miRNAs are more frequent.

magnitude less reads than the most abundant miRNA hsa-miR-223. The uniquely mapped read counts for the ten most abundant miRNAs are listed in Table 4. These numbers indicate that the total read counts of miRNAs are not normally distributed. This is also shown by the Shapiro–Wilk test with a significance value <0.05. Since many single miRNAs were not normally distributed we applied the non-parametric WMW test that detected 70 significantly deregulated miRNAs after adjustment for multiple testing including 50 miRNAs (71.4%) that were up-regulated in blood of lung cancer patients and 20 miRNAs (28.6%) that were down-regulated. After exclusion of precursor sequences, we still found 39 deregulated miRNAs, including 28 (71.8%) that were

Table 4 miRNAs with highest unique read count

miRNA	Unique read count
hsa-miR-223	8 646 130
hsa-miR-425	684 517
hsa-miR-185	509 690
hsa-miR-17	367 360
hsa-miR-25	297 148
hsa-miR-130a	267 587
hsa-miR-150	159 224
hsa-miR-93	149 657
hsa-miR-20a	112 031
hsa-miR-339-5p	86 571

up-regulated in blood of lung cancer patients and 11 (28.2%) that were down-regulated. Out of these 39 miRNAs, hsa-miR-140-3p, hsa-miR-130b*, and hsa-miR-181a* showed the lowest calculated AUC value of 0.03 (*i.e.* more abundant in control samples), and miR-99b, and miR-590-3p showed the highest AUC value of 1 (*i.e.* more abundant in lung cancer samples), demonstrating a high diagnostic potential of these miRNAs. Bar-plots of two representative miRNAs with maximal (hsa-miR-590-3p) and minimal (hsa-miR-140-3p) AUC values are given in Fig. 5. Interestingly, out of the 39 miRNAs, 32 have previously been annotated in miRBase including six miRNAs that have been associated with lung cancer, namely hsa-miR-140, hsa-miR-145, hsa-miR-30e, hsa-let-7d, hsa-let-7g, and hsa-miR-98. Out of the 32 miRNAs found in miRBase v16, 25 miRNAs were found to be differentially expressed in our previous study based on miRNA screening on microarrays.[4] Importantly, the direction of deregulation was identical for 21 of 25 miRNAs (84%) in the previous and in the present study. Besides the 32 known miRNAs we also found seven putative miRNAs being significantly deregulated. A cluster heatmap of these seven miRNAs and all 20 samples is provided in Fig. 6. As presented in the cluster dendrogram we found two clear clusters

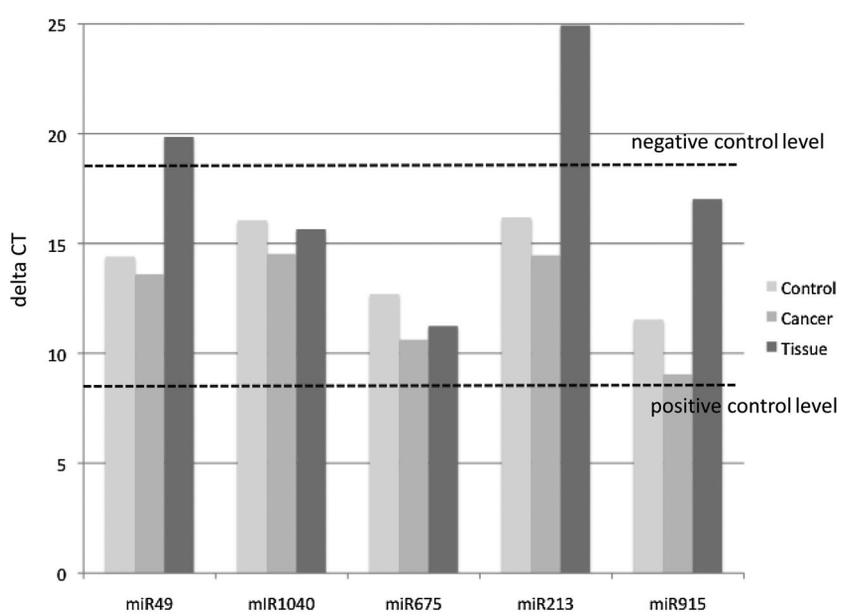


Fig. 4 qRT-PCR validation of novel miRNAs. The dashed lines denote the ΔC_t-values of the positive control hsa-let-7 and the negative control hsa-miR-577. The ΔC_t-values of the indicted five novel miRNAs are given for blood of controls, blood of patients and lung cancer tissues. High ΔC_t-values indicate low abundancy of miRNAs.

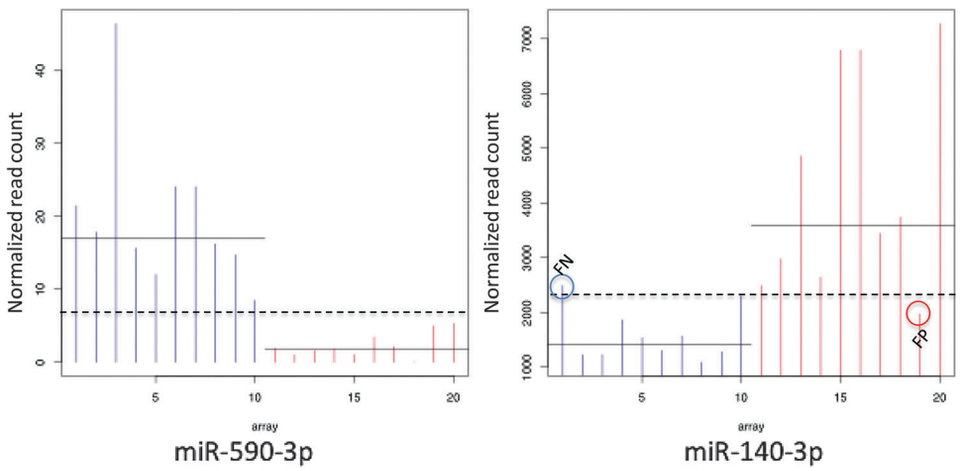


Fig. 5 Normalized read counts for two representative miRNAs with maximal and minimal AUC values. The barplots show the read counts of the two miRNAs hsa-miR-590-3p (AUC = 1) that is up-regulated in lung cancer samples and hsa-miR-140-3p (AUC = 0.03) that is down-regulated in lung cancer samples. Lung cancer patients are indicated as blue bars and controls are indicated as red bars. The horizontal solid black lines denote the respective group medians and the horizontal dashed black lines denote the optimized separation threshold. Based on this threshold for hsa-miR-590-3p no sample is wrongly classified, for hsa-miR-140-3p one cancer sample highlighted with a blue circle is considered to be normal (false negative, FN) while one control sample highlighted with a red circle is considered to be a cancer sample (false positive, FP).

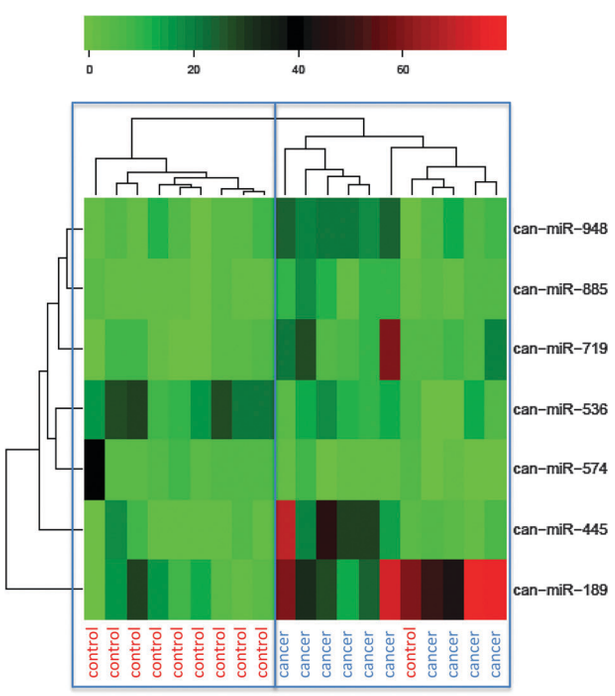


Fig. 6 Hierarchical clustering heatmap. The heatmap with dendrogram at the top and on the left side shows the clustering of the 7 novel miRNAs that were significantly deregulated in lung cancer samples compared to the control samples. The cluster on the left contains nine of ten controls, the cluster on the right contains all ten lung cancer samples and one control sample.

separating lung cancer from controls. Again, only one control clustered together with the lung cancer samples. For this clustering we achieved a highly significant *p*-value of 0.0001 using Fisher's exact test. We repeated the clustering 100 times with randomly distributed class labels by carrying out 100 non-parametric permutation tests but did not find any result with a likewise significance. To check the stability and validity specifically of the novel miRNAs we performed further permutation tests. Here, the class labels of all samples have been randomly shuffled at the beginning of the statistical analysis before WMW tests have been carried out. Again, we performed 100 non-parametric permutation tests and again we did not find a single one with a similarly high number of significant miRNAs.

Table 5 provides detailed information on all 39 significantly deregulated miRNAs.

Identification of novel mature miRNAs derived from known precursors

Considering the distribution of mature miRNA reads across known and novel miRNA precursors, we usually detected two clear peaks, representing the two mature forms of the respective miRNA precursor (Fig. 7). For hsa-miR-339, we found 3-fold up-regulation of hsa-miR-339-3p and 2-fold up-regulation of hsa-miR-339-5p in lung cancer, matching exactly the two known mature forms. Likewise, hsa-miR-98 showed two clear peaks. While the major form of hsa-miR-98 that was up-regulated in blood of controls was deposited in miRBase, a minor miRNA has not yet been annotated. qRT-PCR with the respective primer confirmed that this new minor miRNA, denoted as miR-98_new, was detectable and significantly over-expressed (2.3-fold) in blood of lung cancer patients (Fig. 8).

To identify further novel minor forms of known miRNA precursors, we aligned all reads against known precursor sequences and searched for hits that are not already known major miRNAs. Altogether, we detected 41 novel forms of validated miRNA precursors (Table 6). Out of those, 25 (61%) were even more abundant than the already known form considering all reads combined from the 20 blood samples. Comparing the blood samples of lung cancer patients and controls, we found that the abundance of 30 of the 41 newly identified forms (73%) was at least as high in blood of lung cancer patients as in blood of normal controls.

Table 5 Differentially expressed miRNAs in peripheral blood

miRNA	Median read counts in controls	Median read counts in patients	Fold change	WMW rawp	WMW adjp	AUC	HMDD	Micro-array[4] (fold change)	Concor- dance microarray and NGS
hsa-miR-140-3p	**1421.0**	**3593.2**	**0.4**	**0.0004**	**0.0161**	**0.03**	Down	—	—
hsa-miR-130b*	**181.0**	**313.8**	**0.6**	**0.0004**	**0.0161**	**0.03**	—	Up (0.05)	YES
hsa-miR-181a*	**10.1**	**40.0**	**0.3**	**0.0001**	**0.0150**	**0.03**	—	Up (0.8)	YES
hsa-miR-25	**9798.5**	**19948.3**	**0.5**	**0.0007**	**0.0172**	**0.05**	—	Up (0.85)	YES
hsa-miR-551a	**2.8**	**9.6**	**0.3**	**0.0005**	**0.0161**	**0.07**	—	Up (0.77)	YES
hsa-miR-22	**785.5**	**1917.5**	**0.4**	**0.0005**	**0.0161**	**0.07**	—	Up (0.92)	YES
hsa-miR-326	**27.9**	**53.4**	**0.5**	**0.0005**	**0.0161**	**0.07**	—	Up (0.82)	YES
hsa-miR-151-3p	**54.5**	**96.5**	**0.6**	**0.0005**	**0.0161**	**0.07**	—	Up (0.85)	YES
hsa-miR-501-5p	**53.1**	**72.5**	**0.7**	**0.0013**	**0.0222**	**0.07**	—	Up (0.7)	YES
hsa-miR-186	**2181.4**	**4476.8**	**0.5**	**0.0014**	**0.0231**	**0.08**	—	Up (0.34)	YES
hsa-miR-93*	**649.3**	**2242.3**	**0.3**	**0.0007**	**0.0172**	**0.08**	—	Up (0.67)	YES
hsa-can-miR-948	**4.8**	**15.6**	**0.3**	**0.0007**	**0.0172**	**0.08**	—	—	—
hsa-miR-1248	**3.8**	**15.4**	**0.2**	**0.0017**	**0.0246**	**0.08**	—	Up (0.92)	YES
hsa-miR-188-3p	**2.8**	**7.8**	**0.4**	**0.0017**	**0.0246**	**0.08**	—	—	—
hsa-miR-21*	**3.5**	**11.9**	**0.3**	**0.0017**	**0.0246**	**0.08**	—	Down (1.3)	NO
hsa-miR-339-5p	**1839.4**	**5461.0**	**0.3**	**0.0019**	**0.0253**	**0.09**	—	Up (0.78)	YES
hsa-miR-362-3p	**63.8**	**149.0**	**0.4**	**0.0022**	**0.0256**	**0.09**	—	Up (0.63)	YES
hsa-miR-145	**376.5**	**1287.2**	**0.3**	**0.0028**	**0.0316**	**0.10**	Tumor suppressor	Up (0.49)	YES
hsa-can-miR-445	**3.5**	**16.8**	**0.2**	**0.0032**	**0.0332**	**0.11**	—	—	—
hsa-can-miR-885	**2.5**	**6.6**	**0.4**	**0.0032**	**0.0332**	**0.11**	—	—	—
hsa-can-miR-189	**13.4**	**40.2**	**0.3**	**0.0021**	**0.0253**	**0.11**	—	—	—
hsa-can-miR-719	**5.1**	**11.1**	**0.5**	**0.0045**	**0.0405**	**0.12**	—	—	—
hsa-miR-378*	**342.9**	**708.8**	**0.5**	**0.0046**	**0.0405**	**0.12**	—	Up (0.42)	YES
hsa-miR-26b*	**119.6**	**232.9**	**0.5**	**0.0046**	**0.0405**	**0.12**	—	—	—
hsa-miR-505	**32.6**	**42.4**	**0.8**	**0.0046**	**0.0405**	**0.12**	—	Up (0.62)	YES
hsa-miR-339-3p	**103.5**	**214.3**	**0.5**	**0.0039**	**0.0372**	**0.13**	—	Up (0.98)	YES
hsa-miR-425	**20779.8**	**40304.7**	**0.5**	**0.0053**	**0.0454**	**0.13**	—	Up (0.92)	YES
hsa-miR-30e	**234.8**	**414.0**	**0.6**	**0.0052**	**0.0447**	**0.14**	Down	Up (0.83)	YES
hsa-can-miR-536	21.4	11.9	1.8	0.0028	0.0316	0.90	—	—	—
hsa-let-7d	1570.8	650.4	2.4	0.0017	0.0246	0.92	Down	Down (1.11)	YES
hsa-can-miR-574	5.9	1.4	4.2	0.0010	0.0186	0.94	—	—	—
hsa-miR-574-3p	3806.2	1344.6	2.8	0.0010	0.0186	0.94	—	Up (0.71)	NO
hsa-let-7g	1433.1	367.2	3.9	0.0010	0.0186	0.94	Down	Up (0.73)	NO
hsa-miR-98	48.8	11.3	4.3	0.0001	0.0161	0.96	Up NSCLC *vs.* SCLC	—	—
hsa-miR-144*	2643.0	788.8	3.4	0.0004	0.0161	0.97	—	Down (1.6)	YES
hsa-miR-3200-3p	30.1	1.2	24.6	0.0003	0.0161	0.98	—	—	—
hsa-miR-126*	574.1	155.0	3.7	0.0003	0.0161	0.98	—	—	—
hsa-miR-99b	38.4	1.1	36.5	0.0002	0.0161	1.00	—	Up (0.87)	NO
hsa-miR-590-3p	17.0	1.9	9.2	0.0002	0.0161	1.00	—	—	—

WMW = Wilcoxon Mann–Whitney test, rawp = raw *p*-value, adjp = adjusted *p*-value, AUC = area under the receiver operator characteristics curve. HMDD = human miRNA and disease database. Up-regulation in lung cancer patients is indicated in bold type and down-regulation in lung cancer patients is indicated in normal type.

Discussion

miRNAs are believed to change future diagnostics of many human diseases. In this study we identified miRNA profiles with diagnostic information for lung cancers by next-generation sequencing. We identified 32 known miRNAs and seven novel miRNAs that were significantly altered in cancer patients, providing a tool to detect manifest lung cancer of different histological grading in peripheral blood.

While most miRNA sequencing studies have been performed on cell lines and solid tissues, only a minority was done on human blood. Since miRNA profiles are known to be tissue specific,[32,33] it can be expected that blood also contains a specific profile with so-far unknown miRNAs. We were able to identify completely novel miRNAs and previously unknown mature miRNAs of already known miRNA precursors. These results together with previously published studies suggest that miRNA profiling from blood bears high potential to serve as a novel biomarker class for human diseases.

As for any biomarker approach, standardization is essential to make miRNA profiles derived from human blood comparable between different clinical centers, studies, cohorts, and disease entities. Hence, we employed a standardized workflow, starting from blood collection to molecular barcoding and miRNA profiling using SOLiD next-generation sequencing technology and ending with a sophisticated bioinformatic evaluation. We found by this approach an extremely high dynamic range of quantification of specific miRNAs. Over 8 million reads have been sequenced for hsa-miR-223 that was the most abundant miRNA. Recently, Fehniger *et al.* showed that this miRNA is present in resting NK cells where it may contribute to control Granzyme B translation.[34] An other recent sequencing study on PBMCs also reported a wide range of expression levels spanning five orders of magnitude.[13] In this study the let-7 family accounted for almost 80% of all reads. In agreement with these results, we found 0.5 million reads for all let-7 family members. Hence, although only few NGS studies on miRNAs have been published so far, the methodology seems

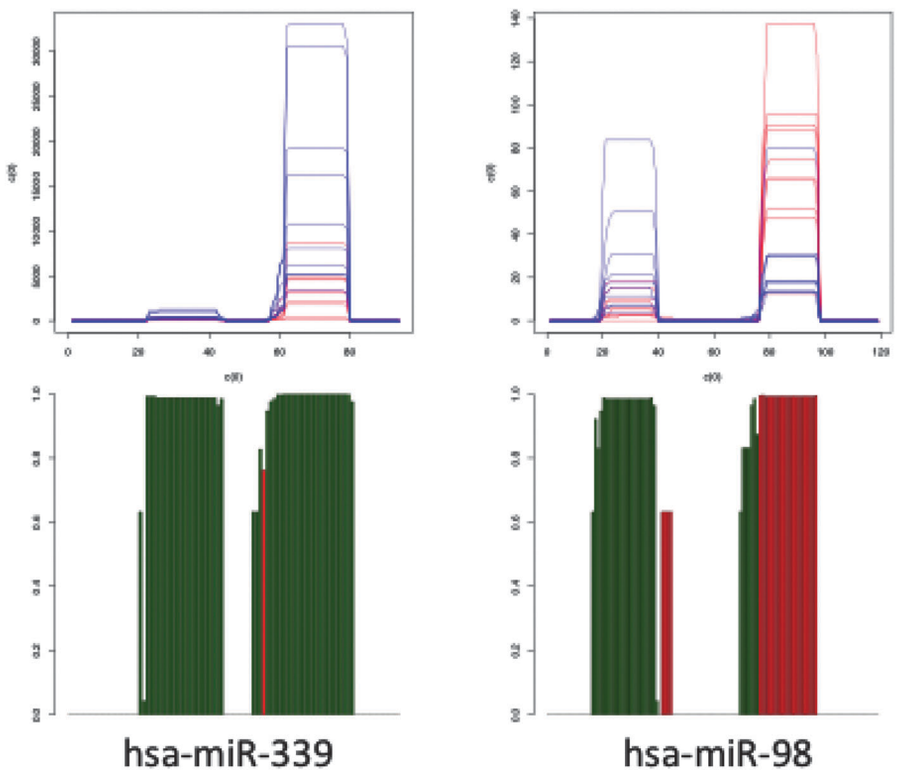


Fig. 7 Distribution of reads across miRNA precursor sequences. Red samples correspond to controls and blue samples correspond to lung cancer patients. For both miRNAs, two peaks can be detected, representing two mature forms. The barplots show the outcome of the WMW test at each position. The bar height corresponds to 1 minus the respective *p*-value. Green bars mean a significant up-regulation while red bars represent significantly down-regulated positions.

to be already very powerful.[7–13] To further underline this hypothesis, we related our NGS data to findings that we previously obtained by microarray-based technologies.[4] For this former study we also analyzed the miRNA expression profiles of blood from lung cancer patients. The blood samples were collected under the same conditions and the RNA isolation was performed using the same protocol as for the present study. Of the 32 known miRNAs deregulated in the present study, 25 miRNAs were also differentially expressed in our former study, including 21 miRNAs (84%) that were regulated in the same direction, *e.g.*, they were either up- or down-regulated in both studies. These findings confirm the high reproducibility of both approaches even on different biological replicates and across different platforms, and demonstrate the feasibility to perform biomarker discovery by these techniques.

In another recent study, we investigated 13 different human pathologies for deregulated miRNAs in patients' blood using the same microarray-based technology as mentioned above and the same protocol for the collection of blood and the isolation of RNA.[6] We found a high overlap between the 32 known miRNAs significantly deregulated in the present NGS study and the miRNAs significantly deregulated in the former microarray study. Nearly all of the 32 miRNAs were deregulated in at least one of the 13 diseases (see Table S2, ESI†). Only the two miRNAs hsa-miR-98 and hsa-miR-181a* were not deregulated in our former study and hsa-miR-3200-3p was not included in our former study as it was based on older miRBase versions (v12–14). The highest overlap was found for melanoma (15 miRNAs of 32 miRNAs, 46.86%), multiple sclerosis (12 miRNAs of 32 miRNAs, 37.5%), sarcoidosis (18 miRNAs of 32 miRNAs, 56.25%) and acute myocardial infarction (12 miRNAs of 32 miRNAs, 37.5%).

The comparison of our present study and a former sequencing study from Chen and co-workers from 2008 revealed less overlap.[35] In this study, that was based on miRBase v10, the miRNome of serum and PBMCs of lung cancer patients and healthy individuals was analyzed by SOLEXA sequencing. They only detected 12 of the 32 miRNAs in PMBCs of lung cancer patients and/or healthy individuals. But interestingly, eight of those 12 miRNAs were deregulated in the same direction. As one example, hsa-miR-25 that was twice as much expressed in blood of lung cancer patients compared to controls in our present study was also higher expressed in lung cancer PBMCs and identified as lung cancer specific serum miRNA in the study of Chen and co-workers.

Biomarker studies are often confounded by variables as gender, age or therapy status. While the lung cancer patients included in our study did not get any radio- or chemotherapy before blood drawing or cancer resection and there was no significant difference in the gender distribution, the age between control and case cohort varies. The study by Hooten *et al.*[36] addressed aging related changes of miRNAs. From their study, we extracted 165 miRNAs that may be influenced by the age. Of these, only three (1.8%) overlap with the 32 known miRNAs that were significantly deregulated in lung

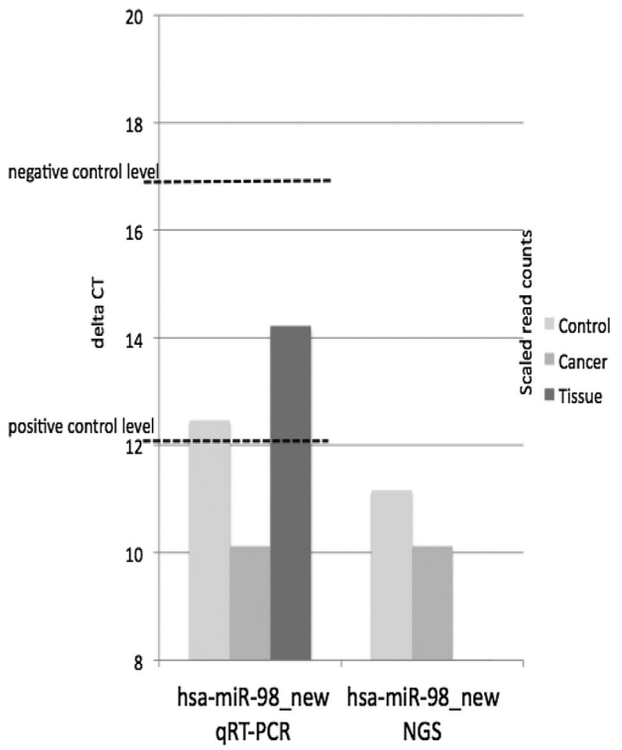


Fig. 8 qRT-PCR validation of the novel mature form of hsa-miR-98. The dashed lines denote the ΔC_t-values of the positive and negative controls as shown in Fig. 1. The ΔC_t-values of hsa-miR-98 are given for blood of controls, blood of patients and lung cancer tissues on the left side. The NGS results are shown on the right side; the *Y*-axis of the NGS data shows the read counts and is scaled to make the NGS data comparable to the qRT-PCR data, *i.e.*, the low NGS read count value of blood of tumor patients indicates a high expression compared to blood of controls.

cancer samples in our study. These include hsa-miR-181a*, hsa-miR-26b*, and hsa-let-7d. This analysis considers just the significance, without considering the direction of regulation. While all three miRNAs were down-regulated in older patients in the study of Hooten *et al.*, hsa-miR-181a* and hsa-miR-26b* were up-regulated in the lung cancer samples, representing in our case the older cohort as compared to the controls. In summary we can conclude that the aging effect seems to play a minor role as compared to the pathogenic processes in lung cancer.

Another putative confounding variable could be the smoking status of individuals. To check whether our miRNAs are related to smoking induced changes we extracted the miRNAs with known smoking association from a review by Tomankova *et al.*[37] We did not found a single miRNA out of the 32 significantly deregulated miRNAs to be related to smoking induced changes. Likewise, we extracted 24 miRNAs from the study of Schembri *et al.*[38] and again did not found any overlapping miRNAs.

We also evaluated the biological relevance of the 32 significantly differentially expressed miRNAs by carrying out a statistical pathway analysis. Using the miRanda algorithm[24] we predicted targets for different miRNA sets. Using GeneTrail[25,26] we computed significantly enriched biological categories, *i.e.*, categories with more target genes of miRNAs in a given set as compared to a reference set containing all miRNAs detected in our sequencing study. For example, we found the JNK and stress associated pathways being significantly enriched in the set of the 32 differentially expressed miRNAs. We performed the same analysis for the novel miRNAs and identified 59 KEGG pathways being significantly enriched for targets of our miRNA candidates with "Metabolic pathways" (p-value of 7.88×10^{-9}) and "Pathways in cancer" (p-value of 1.1×10^{-6}) being already described as target pathways of known miRNAs.[31] It is however, important to realize that the predictions of functional annotations

Table 6 41 novel mature forms (according to miRBase v16) of known miRNA precursors identified in blood: all reads from the 20 blood samples combined identified 25 novel miRNAs that were higher abundant than the known mature form (indicated in bold). Bold italic and italic values represent at least two-fold up- and down-regulated miRNAs, respectively

Known miRNA precursor	Total reads for known pre-miRNAs	Median for known pre-miRNAs in controls	Median for known pre-miRNAs in patients	Novel mature form miRBase v16 (known forms now included in miRBase release v17)	Total reads novel miRNAs	Median for novel miRNAs in controls	Median for novel miRNAs in patients
hsa-miR-1306	12	0.5	0.5	**CACCUCCCCUGCAAACGUCCAG**	16 507	427.5	490
hsa-miR-3194	6	0	0	**GCUCUGCUGCUCACUGGCA** (hsa-miR-3194-3p)	28	0.5	1
hsa-miR-597	7	0	0	**GUGGUUCUCUUGUGGCUCA**	35	1.5	1.5
hsa-miR-1303	28	1	1	**GGGCAACAUAGCGAGACC**	51	1.5	1.5
hsa-miR-3173	12	0.5	0	**GCCCUGCCUGUUUUCUCCUUUGU** (hsa-miR-3173-5p)	1042	45.5	34
hsa-miR-1273c	1721	*35.5*	*72*	AGAGUCUCGUUCUGUUGCCCAA	417	*9*	*34*
hsa-miR-1273d	9	0	0	**CUGCACUUCAGCCUGGGUGA**	39	1.5	2
hsa-miR-939	199	11.5	6	CCUGGGCCUCUGCUCCCCAGU	63	2.5	3
hsa-miR-153-1	—	—	—	**UCAUUUUUGUGAUCUGCAGCU**	27	0.5	1
hsa-miR-3153	1	0	0	**GUCCCUGUCCCCUUCCCCC**	25	1	0.5
hsa-miR-1307	862	41	33.5	CGACCGGACCUCGACCGGCU	410	11.5	17.5
hsa-miR-3155	1	0	0	**CUCCCACUGCAGAGCCUGG**	74	3	1.5
hsa-miR-107	4562	221	136	GCUUCUUUACAGUGUUGCCUUG	403	*13*	*26.5*
hsa-miR-579	278	16.5	12	CGCGGUUUGUGCCAGAUG	22	0	1
hsa-miR-2110	340	12	10	**CACCGCGGUCUUUUCCUCCCACU**	899	32.5	42.5
hsa-miR-1255b-2	—	—	—	**CACUUUCUUUGCUCAUCCA**	26	0.5	1.5
hsa-miR-3138	8	0	0	**CUUCCCCCACCUCACUGCC**	64	3.5	3
hsa-miR-1278	—	—	—	**AUGAUAUGCAUAGUACUCCCA**	26	1	1
hsa-miR-874	588	23.5	26	GGCCCCACGCACCAGGGUAAG	56	1	3.5

Table 6 (*continued*)

Known miRNA precursor	Total reads for known pre-miRNAs	Median for known pre-miRNAs in controls	Median for known pre-miRNAs in patients	Novel mature form miRBase v16 (known forms now included in miRBase release v17)	Total reads novel miRNAs	Median for novel miRNAs in controls	Median for novel miRNAs in patients
hsa-miR-610	—	—	—	**CCCAGCACACAUUUAGCUCAC**	27	1.5	1
hsa-miR-584	702	40.5	30.5	CAGUUCCAGGCCAACCAGGCU	448	13	19.5
hsa-miR-196a-1	—	—	—	**CAACAACAUUAAACCACCCGAU**	588	1	1
hsa-miR-3162	—	—	—	**CCCUACCCCUCCACUCCCCA** (hsa-miR-3162-3p)	89	4	2.5
hsa-miR-301a	1012	***64.5***	***26.5***	CUCUGACUUUAUUGCACUAC	68	1	4
hsa-miR-660	4154	179.5	206	CCUCCUGUGUGCAUGGAUUA	649	16	29.5
hsa-miR-3140	24	0.5	1	**CCUGAAUUACCAAAAGCUUU (hsa-miR-3140-5p)**	76	2.5	4.5
hsa-miR-98	544	***47.5***	***9.5***	UAUACAACUUACUACUUUCC	172	*3*	*8.5*
hsa-miR-382	31	1	1	**AUCAUUCACGGACAACACUUU**	40	1	1.5
hsa-miR-3143	223	12	12	AACUCUUUACAAUGUUUCU	29	1	1
hsa-miR-627	267	9	11	**CCUCUUUUCUUUGAGACUCACU**	319	*6.5*	*22.5*
hsa-miR-210	16 511	*330.5*	*975.5*	GCCCCUGCCCACCGCACACUGC	696	20.5	26.5
hsa-miR-421	7963	135.5	150.5	CCUCAUUAAAUGUUUGU	23	1	0.5
hsa-miR-3127	5	0	0	**CCCCUUCUGCAGGCCUGCU** (hsa-miR-3127-3p)	75	1.5	4
hsa-miR-1294	52	2	2	**CAACAGUGCCAACCUCACAGGA**	1101	42.5	58.5
hsa-miR-3944	—	—	—	**GUGCAGCAGGCCAACCGAGA** (hsa-miR-3944-5p)	50	2	2
hsa-miR-3922	—	—	—	**CAAGGCCAGAGGUCCCACA** (hsa-miR-3922-5p)	30	0.5	2
hsa-miR-1289-1	—	—	—	**AGACUCUUGGUUUCCACCCCCA**	48	1.5	0.5
hsa-miR-942	1862	57.5	85	ACAUGGCCGAAACAGAGAAGU	77	4	2.5
hsa-miR-190	478	***29.5***	***9***	CUAUAUAUCAAACAUAUUCCU	100	5.5	1.5
hsa-miR-1538	12	0	1	**AACAGCAGCAACAUGGGCCUCG**	146	5	6
hsa-miR-3676	19	0	1	**GAUCCUGGGUUCGAAUCCCA**	2551	89	151.5

are based on *in silico* approaches and each target awaits experimental confirmation.

The Sanger miRBase shows a rapidly increasing content, mainly driven by increased sequencing capacity at significantly decreased cost. Since the first release in 2008, a total of 32 versions have been released. Most recently, a new release (v17) was announced. Mapping the miRNAs detected in this study to this latest release we found an overlap of about 9% between our newly identified miRNAs and the miRNAs recently included in the new miRBase release v17, representing an independent validation. These miRNAs include hsa-can-miR-243, hsa-can-miR-277, hsa-can-miR-929, hsa-can-miR-586, hsa-can-miR-637, hsa-can-miR-912, hsa-can-miR-9, hsa-can-miR-288, hsa-can-miR-674, hsa-can-miR-49, hsa-can-miR-180, hsa-can-miR-1003, hsa-can-miR-74, hsa-can-miR-430, hsa-can-miR-782, hsa-can-miR-865, hsa-can-miR-670, hsa-can-miR-1065, and hsa-can-miR-814.

In summary, our study shows for the first time the potential of NGS to identify and quantify in a single step known and completely novel miRNAs with diagnostic potential for lung cancer. In the foreseeable future as many as 100 samples can be sequenced per run, making NGS of blood borne miRNAs an attractive alternative to other approaches. In addition, a standardized NGS approach as applied in this study will help to reveal specific expression patterns of miRNAs for a larger variety of diseases and patient cohorts.

Funding

This work was supported by funding of the German Ministry of Research Education (BMBF) under contract 01EX0806, the Hedwig-Stalter foundation, HOMFOR, and Deutsche Forschungsgemeinschaft (DFG).

Acknowledgements

The authors would like to thank all blood donors who participated in this study.

References

1 B. Meder, *et al.*, MicroRNA signatures in total peripheral blood as novel biomarkers for acute myocardial infarction, *Basic Res. Cardiol.*, 2011, **106**(1), 13–23.
2 P. Leidinger, *et al.*, High-throughput miRNA profiling of human melanoma blood samples, *BMC Cancer*, 2010, **10**, 262.
3 A. Keller, *et al.*, Multiple sclerosis: microRNA expression profiles accurately differentiate patients with relapsing–remitting disease from healthy controls, *PLoS One*, 2009, **4**(10), e7440.
4 A. Keller, *et al.*, miRNAs in lung cancer—studying complex fingerprints in patient's blood cells by microarray experiments, *BMC Cancer*, 2009, **9**, 353.
5 S. F. Hausler, *et al.*, Whole blood-derived miRNA profiles as potential new tools for ovarian cancer screening, *Br. J. Cancer*, **103**(5), 693–700.
6 A. Keller, *et al.*, Toward the blood-borne miRNome of human diseases, *Nat. Methods*, 2011, **8**(10), 841–843.
7 Z. Hu, *et al.*, Serum microRNA signatures identified in a genome-wide serum microRNA expression profiling predict survival of non-small-cell lung cancer, *J. Clin. Oncol.*, 2010, **28**(10), 1721–1726.
8 C. Zhang, *et al.*, Expression profile of MicroRNAs in serum: A fingerprint for esophageal squamous cell carcinoma, *Clin. Chem.*, 2010, **56**(12), 1871–1879.
9 R. Liu, *et al.*, A five-microRNA signature identified from genome-wide serum microRNA expression profiling serves as a fingerprint for gastric cancer diagnosis, *Eur. J. Cancer*, 2011, **47**(5), 784–791.

10 L. M. Li, *et al.*, Serum microRNA profiles serve as novel biomarkers for HBV infection and diagnosis of HBV-positive hepatocarcinoma, *Cancer Res.*, 2010, **70**(23), 9798–9807.
11 Q. Ge, *et al.*, Sequencing circulating miRNA in maternal plasma with modified library preparation, *Clin. Chim. Acta*, 2011, **412**(21-22), 1989–1994.
12 S. S. Luo, *et al.*, Human villous trophoblasts express and secrete placenta-specific microRNAs into maternal circulation via exosomes, *Biol. Reprod.*, 2009, **81**(4), 717–729.
13 C. Vaz, *et al.*, Analysis of microRNA transcriptome by deep sequencing of small RNA libraries of peripheral blood, *BMC Genomics*, 2010, **11**, 288.
14 S. Griffiths-Jones, miRBase: the microRNA sequence database, *Methods Mol. Biol.*, 2006, **342**, 129–138.
15 S. Griffiths-Jones, *et al.*, miRBase: tools for microRNA genomics, *Nucleic Acids Res.*, 2008, **36**(database issue), D154–D158.
16 S. Griffiths-Jones, *et al.*, miRBase: microRNA sequences, targets and gene nomenclature, *Nucleic Acids Res.*, 2006, **34**(database issue), D140–D144.
17 M. R. Friedlander, *et al.*, Discovering microRNAs from deep sequencing data using miRDeep, *Nat. Biotechnol.*, 2008, **26**(4), 407–415.
18 L. A. Goff, *et al.*, Ago2 immunoprecipitation identifies predicted microRNAs in human embryonic stem cells and neural precursors, *PLoS One*, 2009, **4**(9), e7192.
19 E. Bonnet, *et al.*, Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences, *Bioinformatics*, 2004, **20**(17), 2911–2917.
20 S. F. Altschul, *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 1997, **25**(17), 3389–3402.
21 L. Lestrade and M. J. Weber, *et al.*, snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs, *Nucleic Acids Res.*, 2006, **34**(database issue), D158–D162.
22 C. Liu, *et al.*, NONCODE: an integrated knowledge database of non-coding RNAs, *Nucleic Acids Res.*, 2005, **33**(database issue), D112–D115.
23 R. Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, 2008.
24 A. J. Enright, *et al.*, MicroRNA targets in Drosophila, *GenomeBiology*, 2003, **5**(1), R1.
25 A. Keller, *et al.*, GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments, *BMC Bioinformatics*, 2008, **9**, 552.
26 C. Backes, *et al.*, GeneTrail—advanced gene set enrichment analysis, *Nucleic Acids Res.*, 2007, **35**, W186–W192.
27 B. John, *et al.*, Human MicroRNA targets, *PLoS Biol.*, 2004, **2**(11), e363.
28 D. Karolchik, *et al.*, The UCSC Table Browser data retrieval tool, *Nucleic Acids Res.*, 2004, **32**(database issue), D493–D496.
29 P. Roth, *et al.*, A specific miRNA signature in the peripheral blood of glioblastoma patients, *J. Neurochem.*, 2011, **118**(3), 449–457.
30 M. Kanehisa and S. Goto, KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, 2000, **28**(1), 27–30.
31 C. Backes, *et al.*, A dictionary on microRNAs and their putative target pathways, *Nucleic Acids Res.*, 2010, **38**(13), 4476–4486.
32 T. Babak, *et al.*, Probing microRNAs with microarrays: tissue specificity and functional inference, *RNA*, 2004, **10**(11), 1813–1819.
33 J. Lu, *et al.*, MicroRNA expression profiles classify human cancers, *Nature*, 2005, **435**(7043), 834–838.
34 T. A. Fehniger, *et al.*, Next-generation sequencing identifies the natural killer cell microRNA transcriptome, *Genome Res.*, **20**(11), 1590–1604.
35 X. Chen, *et al.*, Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases, *Cell Res.*, 2008, **18**(10), 997–1006.
36 N. Hooten, *et al.*, microRNA expression patterns reveal differential expression of target genes with age, *PLoS One*, **5**(5), e10724.
37 T. Tomankova, M. Petrek and E. Kriegova, Involvement of microRNAs in physiological and pathological processes in the lung, *Respir. Res.*, **11**, 159.
38 F. Schembri, *et al.*, MicroRNAs as modulators of smoking-induced gene expression changes in human airway epithelium, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**(7), 2319–2324.

Laczny et al. BMC Bioinformatics 2012, 13:36
http://www.biomedcentral.com/1471-2105/13/36

BMC Bioinformatics

SOFTWARE Open Access

miRTrail - a comprehensive webserver for analyzing gene and miRNA patterns to enhance the understanding of regulatory mechanisms in diseases

Cedric Laczny[1], Petra Leidinger[2], Jan Haas[4], Nicole Ludwig[2], Christina Backes[2], Andreas Gerasch[3], Michael Kaufmann[3], Britta Vogel[4], Hugo A Katus[4], Benjamin Meder[4], Cord Stähler[5], Eckart Meese[2], Hans-Peter Lenhof[1] and Andreas Keller[2*]

Abstract

Background: Expression profiling provides new insights into regulatory and metabolic processes and in particular into pathogenic mechanisms associated with diseases. Besides genes, non-coding transcripts as microRNAs (miRNAs) gained increasing relevance in the last decade. To understand the regulatory processes of miRNAs on genes, integrative computer-aided approaches are essential, especially in the light of complex human diseases as cancer.

Results: Here, we present miRTrail, an integrative tool that allows for performing comprehensive analyses of interactions of genes and miRNAs based on expression profiles. The integrated analysis of mRNA and miRNA data should generate more robust and reliable results on deregulated pathogenic processes and may also offer novel insights into the regulatory interactions between miRNAs and genes. Our web-server excels in carrying out gene sets analysis, analysis of miRNA sets as well as the combination of both in a systems biology approach. To this end, miRTrail integrates information on 20.000 genes, almost 1.000 miRNAs, and roughly 280.000 putative interactions, for Homo sapiens and accordingly for Mus musculus and Danio rerio. The well-established, classical Chi-squared test is one of the central techniques of our tool for the joint consideration of miRNAs and their targets. For interactively visualizing obtained results, it relies on the network analyzers and viewers BiNA or Cytoscape-web, also enabling direct access to relevant literature. We demonstrated the potential of miRTrail by applying our tool to mRNA and miRNA data of malignant melanoma. MiRTrail identified several deregulated miRNAs that target deregulated mRNAs including miRNAs hsa-miR-23b and hsa-miR-223, which target the highest numbers of deregulated mRNAs and regulate the pathway "basal cell carcinoma". In addition, both miRNAs target genes like PTCH1 and RASA1 that are involved in many oncogenic processes.

Conclusions: The application on melanoma samples demonstrates that the miRTrail platform may open avenues for investigating the regulatory interactions between genes and miRNAs for a wide range of human diseases. Moreover, miRTrail cannot only be applied to microarray based expression profiles, but also to NGS-based transcriptomic data. The program is freely available as web-server at mirtrail.bioinf.uni-sb.de.

* Correspondence: ack@bioinf.uni-sb.de
[2]Department of Human Genetics, Saarland University, 66421 Homburg/Saar, Germany
Full list of author information is available at the end of the article

Background

Gene expression profiles have gained increasing relevance over the last three decades and have become essential in modern biomedical sciences. About two decades ago, a further class of RNAs has been discovered: these non-coding oligonucleotides are indeed transcribed from the human genome, but no proteins are assembled according to their blueprints. MicroRNAs are a subgroup of these non-coding RNAs, currently attracting more and more attention. They have first been reported in a work by Ruvkun [1] and their first appearance in experiments has been associated with Lee et al [2].

MicroRNAs usually consist of 17 to 23 nucleotides and are detectable in the majority of human tissues and almost all bodily fluids [3-5]. It is known today that microRNAs influence the expression of target genes by binding to the corresponding mRNA, leading to its inactivation. Over 50% of all human coding genes seem to be targets of these short non-coding RNAs. MicroRNAs hereby help to control and fine-tune physiological cellular processes like differentiation, proliferation, or apoptosis. Nowadays, it also became apparent that microRNAs have a strong impact on pathological processes as well: Various microRNAs show altered expression patterns in human disorders including malignant [6-10], neurological [11], cardiovascular [12,13], or rheumatic diseases [14,15]. In order to get new insights into the molecular mechanisms leading to a specific disease, increasing attention is paid to the interaction of microRNAs and mRNAs of target genes.

The technologies that are most commonly applied to measure miRNA expression profiles are closely related to the methods for measuring gene expression profiles, namely quantitative real-time polymerase chain reaction (qRT-PCR) [16,17], oligonucleotide microarrays [18,19], and high-throughput sequencing [20,21]. These three technologies allow measuring the expression of sets of miRNA very efficiently. While qRT-PCR is mostly applied to rather small subsets of miRNAs, microarrays enable to profile the whole human miRNome and high-throughput sequencing is additionally applied to detect novel mature forms of miRNAs. Remarkably, with the still growing number of miRNAs, and the likewise growing number of biological experiments carried out with the above-mentioned high-throughput methods, and the manifold of possible interactions between miRNAs and mRNAs, computer aided analyses are essential to grasp the information hidden in the large data sets. Therefore, much ongoing work focuses on the combined analysis of miRNAs and their targets. Two classes of bioinformatics approaches related to this topic are 1) tools that aim at discovering the targets of miRNAs and 2) tools that aim at an integrative analysis of miRNA and mRNA sets. Algorithms belonging to the first class usually rely on sequence-complementarity and often also include thermodynamical aspects [22], machine learning [23-25], or experimental validation steps [26]. An overview of respective programs, including a comparison, can be found in [27]. Additionally, approaches primarily based on experiments are becoming prominent in recent years [28,29]. Naturally, these approaches are more likely to reveal significant miRNA - mRNA interaction pairs than computational approaches. However, they usually require not unimportant amounts of time and resources and, e.g. by design, might also miss relevant interactions. While not strictly being a tool for the discovery of targets or for an integrated analysis of miRNA and mRNA sets, TAM [30] offers enrichment analyses on miRNA sets, thus potentially paving the way to link common functions with related miRNAs. Tools for the second purpose, an integrative analysis of genes and their miRNA regulators, include MMIA [31], DIANA-mirExTra [32], or miRGator [33]. MMIA, allows to combine expression profiles of miRNA and mRNA experiments and then performs a pathway analysis on the intersection of the predicted target mRNAs and the according inversely correlated mRNAs. Additional analyses include Transcription Factor Binding Sites enrichment and diseases that are found to be associated with the inversely deregulated miRNAs. DIANA-mirExTra web-server integrates the potentially novel prediction of miRNAs having one or more of the submitted genes as their targets. This, likewise, allows shedding light on the function of the miRNAs. In detail, the algorithm investigates the 3' UTR sequences of deregulated genes and searches for over-represented six nucleotide long motifs, thus, enabling the identification of matching miRNAs. Finally, miRGator uses public expression data to analyze expression correlation between miRNA and target mRNA/proteins. The miRNA - target interactions are based on miRanda [22], PicTar [34], and TargetScanS [35] and the function of miRNAs is inferred from the related target mRNAs. To this end, a statistical enrichment analysis is performed for the established GO-terms, pathways, and also disease associations. Moreover, it integrates a first approach towards a manual inspection of the underlying network, offering vertex- or edge-filtering but, to-date, no ways to further cope with this information. Here, we present miRTrail (freely available to non-commercial users at mirtrail.bioinf.uni-sb.de), a knowledge-based tool for integrative network analysis that allows for studying the interactions between microRNAs and their target genes, and especially in the case of diseases, the implications of expression changes on pathogenic processes. Our tool excels by its broad

functionality, as (1) it can be applied to a single disease or a group of diseases, (2) it covers a wide variety of biochemical categories, and it can be used to evaluate (3) qRT-PCR, microarray, as well as NGS-based transcriptome data. In its current stage, the organisms of Homo sapiens, Mus musculus, and Danio rerio are supported and further extension is continuing. While many solutions exist that provide either analyses of miRNAs, or mRNAs, or a combination of both, miRTrail allows for the simultaneous, combined statistical analysis of all of these three. A schematic description of its workflow is presented in Figure 1, depicting the integration of the provided data about miRNA and mRNA deregulation and the offered statistical analyses intended to facilitate the work with such complex information, especially when used in combination. As such, our tool is able to not only give initial but also thorough insights, even for a very detailed inspection of the given input based on the network analysis.

One of the original goals of our research was to improve the understanding on the molecular level of melanoma. Thus, as a first application, we investigated miRNA and mRNA expression profiles of this cancer entity, integrating information from 1) the gene expression omnibus GEO [36], 2) the PhenomiR 2.0 human miRNA and diseases database [37], 3) target prediction algorithms [38], and 4) biochemical pathway information of different resources integrated via the BNDB and GeneTrail [39]. By applying miRTrail to these data, we found highly significant coherences between dysregulated miRNAs and matching dysregulated targets of these miRNAs. An additional network analysis highlighted the potential implications of eight miRNAs via their target mRNAs on pathogenic processes in melanomas.

Implementation

In this section, we start by describing the general idea behind miRTrail, followed by the data, techniques and tools that are used to provide the rich functionality, as well as information on the exemplary input data. Here, the input is originating from publicly available services like NCBI GEO and PhenomiR. Our tool is not restricted to these services, as they are intended for demonstration purposes. Especially, all of miRTrail's functionality is available for the organisms of Homo sapiens, Mus musculus, and Danio rerio, and can easily be extended to support other organisms in the future.

Methodoloy - Multipartite graph

Our webservice miRTrail allows for the joint/integrated analysis of miRNA and mRNA entities, in respect to given diseases - since the latter protein-coding RNAs are targets of the former non-coding RNAs. We decided to realize the integration of data by constructing a graph

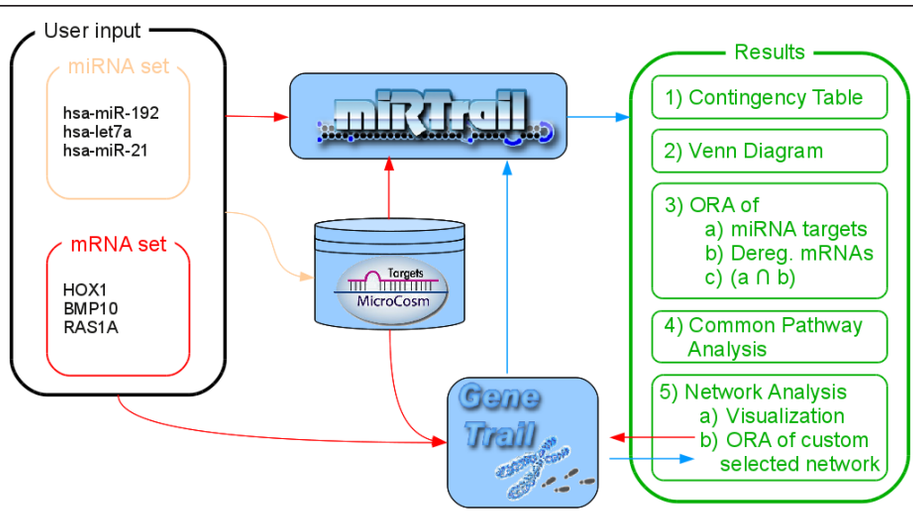


Figure 1 Workflow. Workflow of miRTrail. User submits two RNA sets (one is the set of deregulated miRNAs, the other is the set of deregulated mRNAs, both for the same disease). Orange color represents information flow of miRNA-related information: For each provided miRNA, the target mRNAs are determined (based on microCosm predictions or custom, uploaded interactions). This information then is used by miRTrail, indicated by the red arrow. In general, red color represents flow of mRNA-related information: The uploaded mRNA set as well as the miRNA targets are used in GeneTrail to perform ORAs as described in the "Methods"-section. Blue color represents information flow of results-related information, e.g. for the overlap of pathway sets. Finally in the results, the network analysis allows for targeted inspection of the provided information, based e.g. on miRNA families-related subnetworks. The modular design of miRTrail becomes visible here, also allowing for convenient extension of future analyses and usage for a diversity of different organisms.

or network. To be exact, an r-partite or multipartite graph G:

$$G = (V, E)$$

with:

- $V = V_{\text{miR}} \cup V_{\text{mR}} \cup V_{\text{di}}$
- $V_X \cap V_Y = \emptyset$ for $X \neq Y$ and $X, Y \in \{miR, mR, di\}$
- $E = E_{\text{mR-di}} \cup E_{\text{miR-di}} \cup E_{\text{miR-mR}}$
- $E_{X-Y} = \{(u, v) | \ u \in V_X, v \in V_Y, X \neq Y\}$ for $X, Y \in \{miR, mR, di\}$

where the vertex-set V is the union of disjoint vertex-subsets, the edge-set E is the union of disjoint edge-subsets, and an edge only connects vertices of different vertex-subsets.

For each user and the according uploads, individual networks are created. An efficient and open-source interface for the creation of graphs is offered by the C++ Boost Graph Library and the associated `adjacency_list`-construct.

MicroRNA - Target mRNA Interactions

The actually known 20,000 genes and 1,000 miRNAs allow for 20 million possible interaction pairs, where a miRNA may regulate a gene. To find the most reliable candidates, prediction algorithms have been developed. One of the most prominent algorithms for miRNA - target mRNA interactions is the miRanda algorithm and the respective web-resource microCosm [38]. The miRanda algorithm is sequence-complementarity based and includes a thermodynamic analysis of the miRNA - target mRNA complex. The results are then post-processed by a filtering on conservation of the target site. MicroCosm offers miRNA - target mRNA interactions in combination with a p-value threshold. In the beginning, we decided to perform analyses for three thresholds (0.01, 0.001, 0.0001), and extracted all interactions having a value smaller than the respective alpha level, yielding 279,225, 85,050, and 26,984 interactions, respectively. Because of the heterogenity of the expression data, we finally chose to use a threshold of 0.01, in turn leading to approximately 400 target mRNAs per miRNA in human. The appropriate predictions for the other supported organisms are automatically selected by miRTrail according to the organism in the identifiers of the uploaded miRNA deregulation information.

Alternatively, custom pairwise miRNA - target mRNA interactions can be uploaded in a tab-delimited format, thus allowing e.g. for the use of experimentally validated interactions. Details on the exact format for this input can be found on the homepage of miRTrail, especially regarding gene and miRNA identifiers.

Analysis of independance

Based on the pairwise miRNA - target mRNA interactions (for a custom prediction-threshold (default of 0.01) or from a custom list provided as upload), the miRNA and (target) mRNA of each pair is compared to the input in order to see if it is up-, down- or not deregulated. This information is tabulated in a contingency table to provide an overview to the user. An according p-value is calculated, based on a Chisq-distribution with 4 (6 - 2) degrees of freedom. Given that the uploaded information about dysregulated genes/mRNAs only contains entries with the same direction of deregulation, the computation of an according p-value is not allowed by miRTrail and no p-value will be displayed, but instead a note for the user. However, the table will nevertheless be displayed as an overview.

To help the researcher get an impression about the influence of the used miRNA - target mRNA interactions in this step, especially when using data from prediction algorithms, we offer an option to randomize upon the provided data of miRNA and mRNA deregulation. The deregulation pattern (genes/miRs being up- or downregulated) is kept as-is while the identifiers are sampled at random. This functionality is available via the “Randomize"-button.

MicroRNAs from PhenomiR

For integrating dysregulated miRNAs, we used PhenomiR 2.0 (last update: 2011-02-15). This service offers manually curated data about differential regulation for a variety of diseases.

Specifically, we used the data for entry/ID: 639, concluding the results of a published melanoma study based on microRNA low density arrays, including 666 microRNAs. Selection of the statistically significantly dysregulated miRNAs in the miRNA extracts of adult melanoma patients and benign nevi controls was done with univariate Two-sample T-test and a significance level of 0.05. The size of the patient samples is 10 and 4 for the control, respectively. An overview of the numbers of up- and downregulated miRNAs can be found in Table 1.

MessengerRNAs from NCBI's GEO

The NCBI Gene Expression Omnibus (GEO) is a public repository providing data from microarray experiments,

Table 1 Summary of deregulated genes and miRNAs

	# up-reg	# down-reg	sum of up and down
genes	2550	2218	4768
miRNAs	16	17	33

Deregulated genes (from NCBI GEO) ($\alpha = 0.05$) and deregulated miRNAs (Phenomir 2.0).

next-generation sequencing, and other high-throughput functional genomic data. The microarray experiments, in particular, must comply with MIAME guidelines in order to be accepted by NCBI's GEO.

We extracted the microarray expression profiles from data set GDS1375 (series published: 2005-08-25), including 63 arrays for 45 melanoma and 18 benign nevi samples. Due to possible variations between the experiments, we carried out a quantile-normalization of the expression values for all genes present on the respective data set. Selection of differentially expressed genes was performed on the normalized data using the univariate Two-sample T-test and a significance level of 0.05. An overview of the numbers of up- and downregulated genes can be found in Table 1.

GeneTrail

The gene set analysis tool GeneTrail has been developed to help in the analysis of readily available or newly created high-throughput data. It allows for a comprehensive and efficient statistical evaluation of large genomic or proteomic datasets and covers a plethora of biological categories and pathways, e.g. KEGG, TRANSPATH, TRANSFAC, and GO. Analyses can be either performed via an 'Over-Representation Analysis' (ORA) comparing a reference set of genes to a test set or a 'Gene Set Enrichment Analysis' (GSEA) based on a sorted list of genes. While the calculation of ORA p-values relies on Hypergeometric distribution, many existing tools offer the calculation of GSEA p-values based on permutation tests, usually limited to a fixed number of permutations for performance reasons. GeneTrail integrates an exact calculation [40] corresponding to a commonly used non-parametric unweighted permutation test. This calculation is based on dynamic programming and thus allows, especially for large sets, a higher accuracy than by using a fixed number of permutations.

Recently, GeneTrail has been extended to directly allow the analysis of expression data originating from the NCBI GEO, resulting in GeneTrail Express [41]. This integration greatly facilitates the selection of differentially regulated genes and allows for a fast evaluation of the expression profiles in respect to biological categories and pathways.

Visualization: BiNA and Cytoscape-web

While computational approaches are very important in contemporary research, manual inspection is often desireable to support the automatic analyses or to identify new aspects. To this end, we decided to include the visualization of the resulting interaction network of miRNAs and their (putative) targets. Due to the large amount of integrated data, efficient means for focusing are crucial. Therefore, we provide respective subnetworks, depending either on the choice of individual miRNAs or on members of miRNA-families contained in the input. Furthermore, only deregulated miRNAs are respected that are connected to deregulated target mRNAs, either by the prediction algorithm or the provided custom interactions, as we envision these entities and relations as the most relevant. For the actual visualisation, two selections are available for the user: BiNA and Cytoscape-web.

BiNA is a visualization and analysis tool for various biological networks. We developed a plug-in for the Java Webstart version of BiNA, which takes the miRTrail results and uses the visualization capabilities of BiNA for presenting the network. The user can choose between different graph layouts (organic, hierarchic, and orthogonal) and can modify the visualization in many ways. By default, the target-mRNA nodes are sized according to their degree for easier retrieval of high-degree nodes. It is also possible to save the network in different file formats for reusing the data in other tools or BiNA again. For larger graphs, this visualization-option is probably beneficial.

Cytoscape-web is modeled after the Cytoscape Java network visualization and analysis software [42]. Its JavaScript API allows for an integration into HTML-pages and convenient display of networks. We offer the user a choice of three different graph layouts (Circular, Radial, Tree) and the possibility to select the first neighbors of a selected node. Zoom and pan functionality is available and target-nodes are also sized according to their degree. As the graph is directly displayed in the browser-window, this visualization is especially suitable for a quick inspection of the network. Finally, we implemented context-menu items that greatly facilitate the search for related publications by performing NCBI PubMed queries ("inclusive" or "exclusive") for a custom selection of miRNA and mRNA nodes, given a disease was specified in the input.

Results and Discussion

In the following, we will describe the range of different functions offered by miRTrail. Subsequently, an analysis of cutaneous malignant melanoma versus benign nevi is performed to illustrate the potential of our tool.

Functionality of miRTrail

The miRTrail webserver recieves two dysregulation sets in separate text-files as input, one is the set of dysregulated miRNAs and the other is the set of dysregulated genes. For each uploaded identifier (for miRNAs, the standard annotation of miRBase is used, for genes the HGNC GeneSymbol annotation, respectively), the information whether the respective gene/miRNA is upregulated ('1') or down-regulated ('–1') has to be provided in

the files by the user. Optionally, the disease of interest can be specified to allow for convenient NCBI PubMed queries for related information.

As the next step, the user can either choose a target-prediction threshold for microCosm targets predictions or can provide a list of custom pairwise miRNA - target mRNA interactions, potentially originating from proprietary experiments or other prediction algorithms. The default threshold for microCosm targets predictions is 0.01, amounting to around 280,000 miRNA - target mRNA interactions. Here, the user can also opt-in for a thorough GeneTrail analysis. Based on this information, the analyses are then carried out and, finally, the user is directed to the results. These will be stored uniquely for each analysis performed and can be shared with others by simply providing them with the link of the results page. The results presented herein can be reproduced using the example files provided by miRTrail.

The first provided analysis computes a contingency table relating the dysregulation of miRNAs and the dysregulation of target mRNAs and calculates the according p-value, based on a χ^2 distribution. This analysis allows for estimating whether there is an independance in the deregulation of the miRNAs and the target mRNAs.

Second, a Venn diagram is computed, providing the dysregulated genes that are targets of dysregulated miRNAs (overlap of the diagram), the not-dysregulated targets of dysregulated miRNAs (left part of the diagram) and the dysregulated genes that are not targets of the dysregulated miRNAs (right part of the diagram). For this Venn diagram, a p-value using the Hypergeometric distribution is calculated to show whether there exists a significant overlap between dysregulated genes and targets of dysregulated miRNAs. Third, gene set enrichment analyses for three gene sets are carried out using the comprehensive functionality of GeneTrail. Independently of each other, a so-called Over-Representation Analysis (ORA) - based on the Hypergeometric distribution - is carried out for the dysregulated genes, targets of dysregulated miRNAs and dysregulated targets of dysregulated miRNAs. In all cases, the gene sets are tested for significant enrichments/depletions in KEGG pathways. If the user previously decided to perform all GeneTrail analyses, the results will also include information about GO terms, TransPath pathways, transcription factors from Transfac, SNPs, and chromosomal location, among many others. By clicking on the 'details' button, the complete list of results is provided. Moreover, an overview showing the biological categories being significant in at least two of the three analyses is created. The "code" represents in which of the pairwise overlaps the respective category was found, similar to the file-permission scheme in Linux. So, a code of "2" e.g. shows that a category was found in the enrichment analysis of the dysregulated targets of dysregulated miRNAs as well as in the results of the dysregulated genes. A code of "3" would hence mean that this category was additionally found in the results of the targets of dysregulated miRNAs. Accordingly in Table 2, e.g. the "DNA replication" pathway was found to be enriched for dysregulated targets of dysregulated miRNAs as well as for the dysregulated genes/mRNAs.

Finally, we carry out an integrative network-analysis approach on the comprehensive network containing dysregulated genes, dysregulated miRNAs, and the target interactions between them. A subset of interesting miRNAs and their according targets is selected as well as a custom degree constraint. The subset can be constructed either by selecting individual miRNAs, s. Figure 2, or miRNA families based on an Over-Representation Analysis of miRNA-family data from miRBase (miFam.dat) [43], s. Figure 3. The custom degree constraint allows the selection of the genes being the target of at least as many miRNAs as specified by the parameter. Based on this selection, using the Java Webstart-based viewer BiNA [44] or the web-based viewer Cytoscape-web [45], we show the resulting network, allowing for a manual inspection of the inherent interactions. In the network visualizations, nodes with rectangular shapes belong to miRNAs, nodes with round shapes to genes, red color means up-regulation, green color means down-regulation, and genes and miRNAs are connected by edges if a putative miRNA - target interaction exists. Additionally, a more fine-grained ORA is available, being performed only on the genes that are contained in the custom selection, which is also separately available as a list. These genes are assumed to be the most disease-relevant as they are found to be deregulated and simultaneously putative targets of deregulated miRNAs while, at the same time, being central to the network, according to their degree. This list is analysed for enrichments/depletions in KEGG pathways, Gene

Table 2 Overlapping pathways

Pathway	Related mRNA set	Code
Olfactory transduction 04740	a, b, c	7
DNA replication 03030	b, c	2
Lysosome 04142	b, c	2
Prostate cancer 05215	b, c	2
Small cell lung cancer 05222	b, c	2
Systemic lupus erythematosus 05322	b, c	2

a: Pathways related to targets of dereg. miRNAs
b: Pathways related to dereg. targets of dereg. miRNAs
c: Pathways related to dereg. mRNAs
Code: Arithmetic sum of the following:
1: Found for a and b
2: Found for b and c
4: Found for a and c

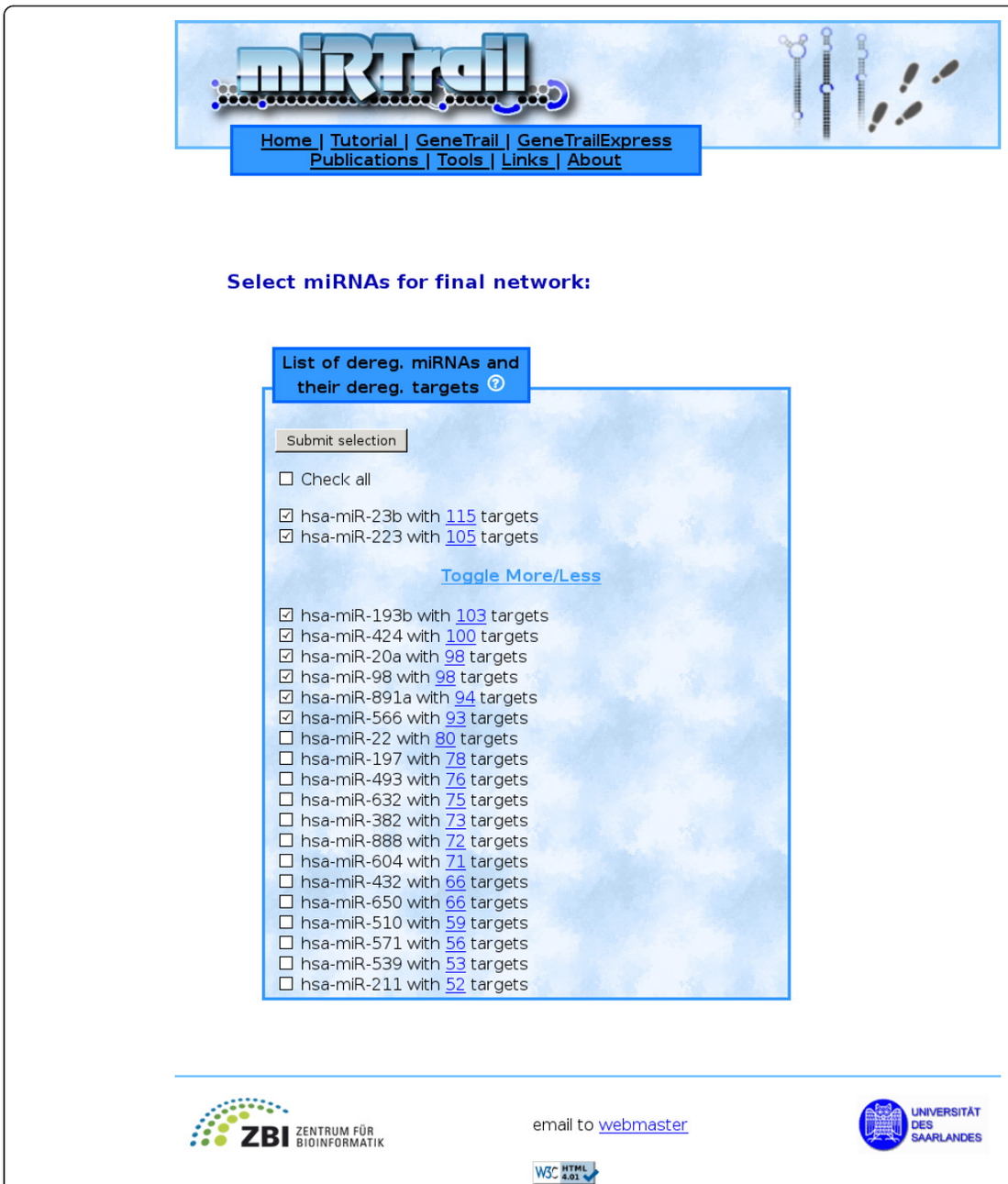


Figure 2 MiRNA selection (individual). Demonstrates the selection of miRNAs of interest. Link next to the each miRNA shows the respective dysregulated target mRNAs.

Ontology terms, OMIM disease relations, and NIA human disease gene sets. A thorough GeneTrail analysis can also be chosen here.

Melanoma case study

We compared the expression profiles of cutaneous malignant melanoma to those of benign skin nevi samples from adult patients. While the proportion of melanoma cases among skin cancer patients is rather low (4%), it accounts for almost 75% of all skin cancer-related deaths. Even more, the prognosis for advanced melanoma is very poor (5-year survival-rate is only 5%) [46]. Hence, we decided to validate our tool based on melanoma data and to identify new aspects of this disease, potentially helping in the creation of promising new therapies for advanced melanoma patients.

An illustration of the the results page for the miRTrail analysis on the melanoma miRNA and mRNA samples,

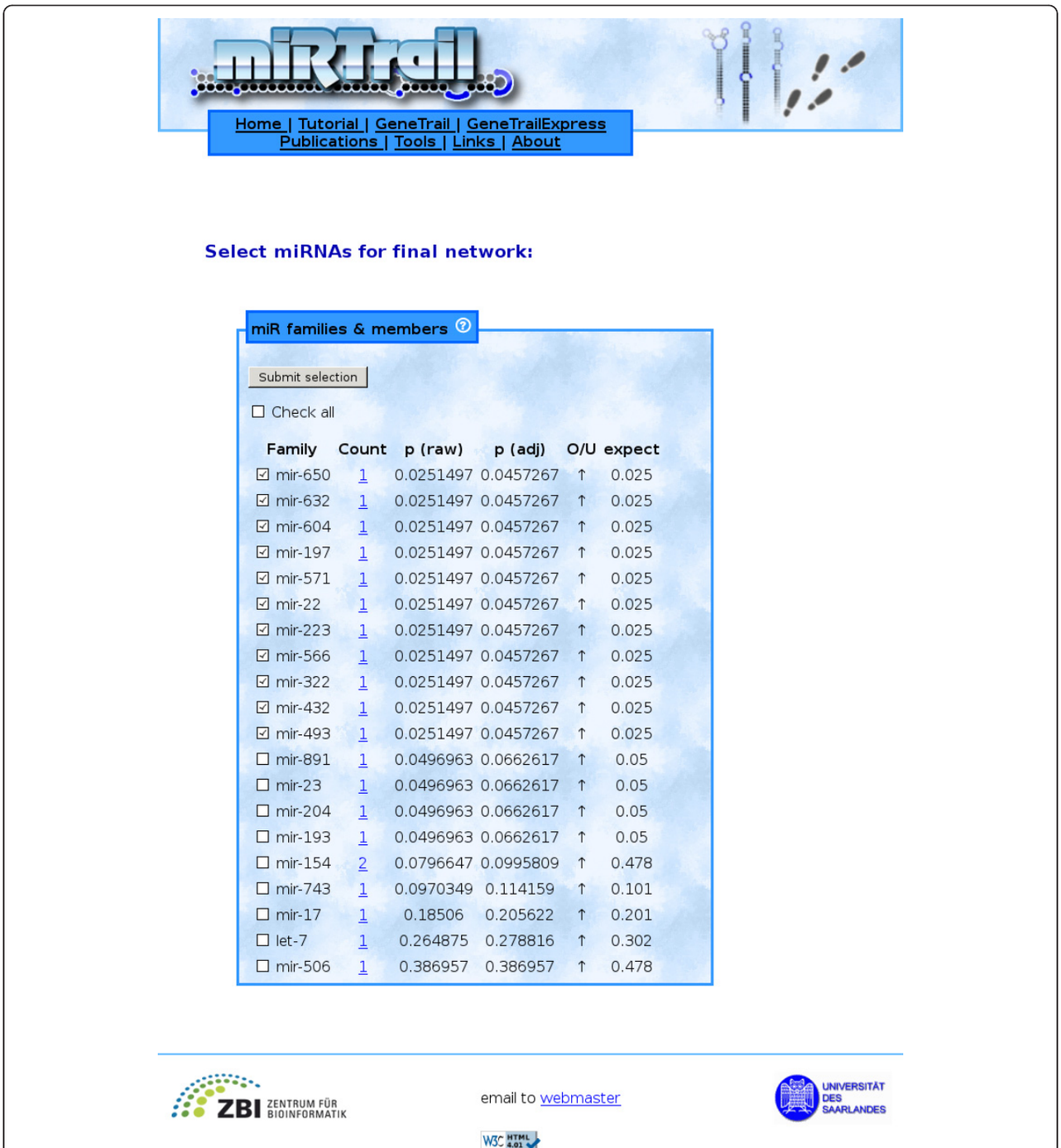


Figure 3 MiRNA selection (families). Demonstrates the selection of miRNAs based on enriched miRNA families. All significant families (p(adj) <0.05) are preselected.

as mentioned in the previous section, can be seen in Figure 4.

Analysis of independance

Using our tool, for the melanoma samples and a prediction threshold of 0.01 for the human miRNA - target mRNA interactions from MicroCosm targets, we were able to find statistical evidence about the dependance of deregulation of miRNAs and target mRNAs. The contingency table yielded a p-value of 0.025 (α = 0.05). Interestingly, independant of the miRNA being up- or downregulated, similar amounts of interactions were found for targets, then, being up (450 and 480), down

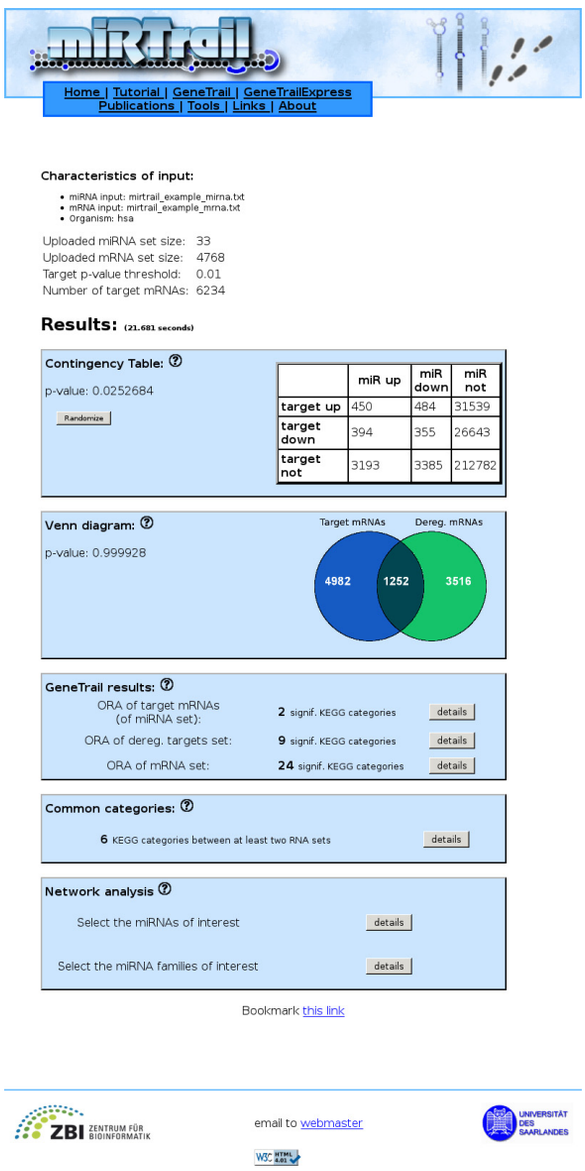


Figure 4 Results Page. This illustrates the results page of a miRTrail analysis. The network visualization and subsequent analyses are available in the subpanel at the bottom.

(394 and 355), or not deregulated (3193 and 3385), respectively. In turn, more targets were found to be up-regulated for dysregulated miRNAs as well as for not dysregulated miRNAs. Surprisingly, 805 (450 + 355) interactions were found with both miRNA and predicted target, being deregulated in the same direction. Finally, as expected, a large number of interactions was found were both, the miRNA and the predicted target, were not deregulated (212782).

ORA of the three mRNA sets

Inspecting the ORA results, showed one KEGG pathway (GPI-anchor biosynthesis) to be significantly enriched for the set of targets of dysregulated miRNAs. In turn, the over-representation analysis of dysregulated targets of dysregulated miRNAs revealed nine significant pathways, including cancer related categories, like Non-small cell lung cancer, Prostate cancer, Small cell lung cancer, Endometrial cancer, and Glioma as well as enrichments in the Lysosome pathway and DNA replication. The analysis of the dysregulated mRNAs revealed the highest number of statistically signifi cant KEGG categories with a total of 24. Here again, several cancer-related pathways were found to be enriched as well as pathways like Cell cycle, Focal adhesion, or even signaling pathways (e.g. TGF-beta signaling pathway).

The result of the pairwise overlaps of the resulting pathway sets is described in Table 2. Among those, the DNA replication pathway [47] as well as the Lysosome pathway [48,49] have already been attributed to melanoma.

Network analysis

From the 33 input miRNAs, 21 were found to have dysregulated targets for a prediction threshold of 0.01. The miRNA with the least dysregulated targets was hsa-miR-211 (52 targets) while hsa-miR-23b was the miRNA with the most dysregulated targets (115). For this analysis, we decided to use the eight miRNAs having more than 80 dysregulated targets (miR-23b [50], miR-223, miR-193b [51], miR-424, miR-20a [52], miR-98, miR-891a, and miR-566), see Figure 2. We left the custom degree constraint at the default of 1 for the subsequent ORA. The resulting mRNA set comprises the dysregulated mRNAs that were predicted targets of at least one of the eight earlier miRNAs. Specifying a higher constraint would lead to a smaller network with only the mRNA nodes being targets of at least as many miRNAs as specified by this parameter and the according miRNA nodes, respectively.

ORA of subnetwork The analysis of KEGG pathways showed significant enrichments for the three cancer-related categories: Prostate cancer, Non-small cell lung cancer, and Endometrial cancer. These categories were found to be enriched for genes that were deregulated while being targets of deregulated miRNAs, hence, the genes that are assumed to be the most disease-relevant due to their joint deregulation.

A total of 274 GO terms were found to be enriched or depleted for all of the three GO-trees, with enrichments in anti-apoptosis, cell proliferation, cell cycle, transcript initiation, RNA elongation, and regulation of translational initiation among others in the biological subtree.

Furthermore, an enrichment (RASA1 and PTCH1) for "Susceptibility to basal cell carcinoma" was found in the OMIM categories.

Visualization For this step, we decided to focus on smaller miRNA and mRNA sets to increase the visibility. However, also large selections can be efficiently handled and used for detailed manual inspections. Exemplary visualization can be found in Figure 5 and 6, for BiNA and for Cytoscape-web, respectively.

Conclusions

The constantly increasing availability of data from different origins and of different nature allows for more complex and comprehensive analyses. To this end, we developed miRTrail to integrate information about RNA deregulation in diseases and putative interactions of miRNAs and mRNAs. Our tool provides a large

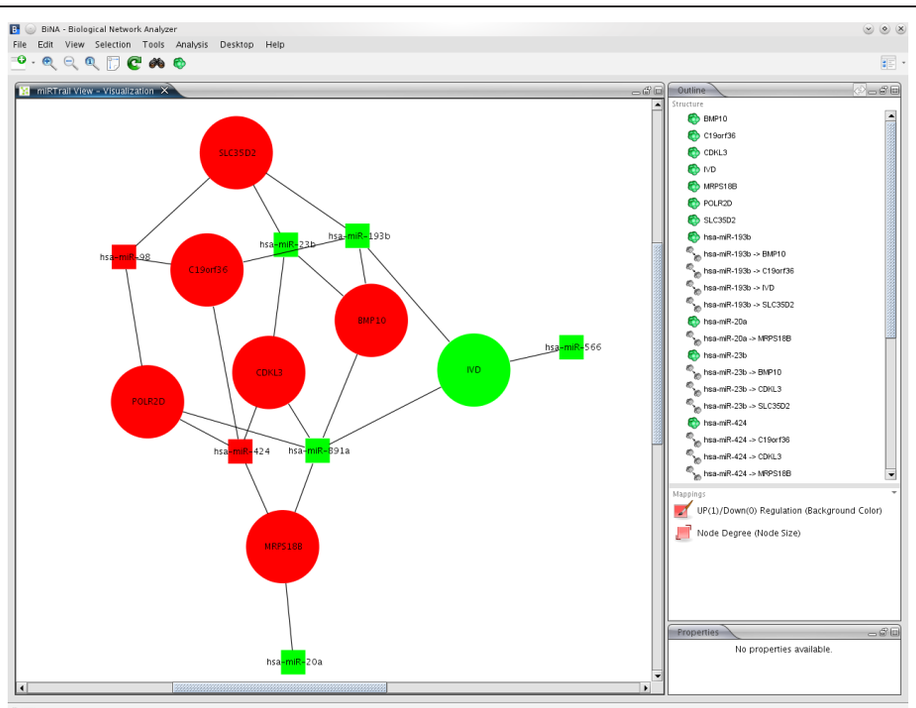


Figure 5 Visualization: BiNA. Subnetwork of top-8 miRNAs and a degree constraint for the target mRNAs of 3, thus, only seven miRNA nodes are displayed. Round shape represents mRNAs, rectangular miRNAs, respectively. Red color indicates upregulation, green color downregulation, respectively. Size of the mRNAs is according to their degree.

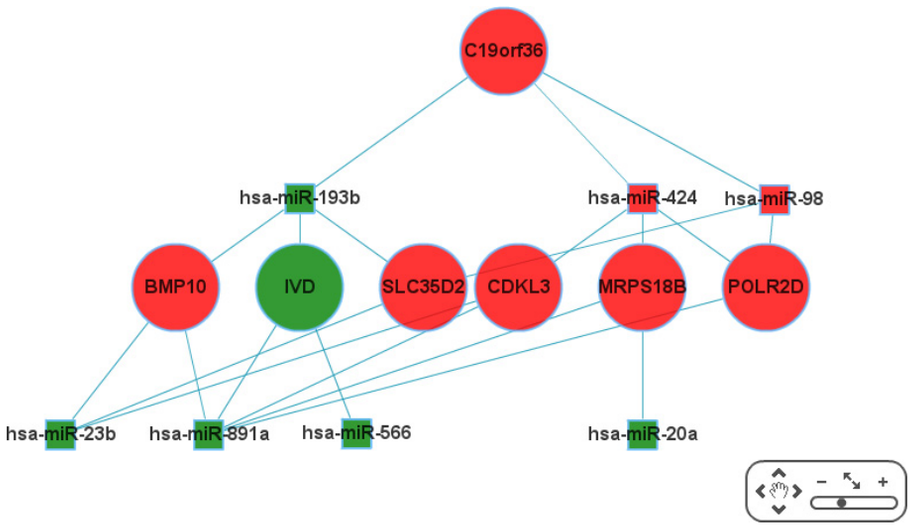


Figure 6 Visualization: Cytoscape-web. Subnetwork of top-8 miRNAs and a degree constraint for the target mRNAs of 3, thus, only seven miRNA nodes are displayed. Round shape represents mRNAs, rectangular miRNAs, respectively. Red color indicates upregulation, green color downregulation, respectively. Size of the mRNAs is according to their degree.

collection of different analyses and performs in a transparent way, requiring only minor activity by the user, while offering versatile results. This greatly facilitates the adaption of this tool as it does not require complicated initial learning. Via the visualisation component, miRTrail enables the user to easily inspect the interactions and, thus, also to further process upon the selection.

MiRTrail - results will also be of great help in any scheme that aims in experimental confirmation of miRNA-targets. The final proof, here, requires extended experiments including the identification of the specific targeted region of a gene by in vitro binding and the analysis of in vivo effects by altered miRNA expression. The melanoma case study shows that we were able to detect highly significant results, despite the fact that we did not use autologous samples. This sets the ground for specific experimental assays that focus on significant miRNA - mRNA interactions in this tumor type. Hence, miRTrail is of great interest for the life sciences community as it can use data from next-generation sequencing, qRT-PCR, or microarray experiments.

Availability and requirements

Project name: miRTrail

Project home page: http://mirtrail.bioinf.uni-sb.de

Operating system(s): Platform independent

Programming languages: C++, php

Other requirements: JavaWS version 1.6 or higher

Acknowledgements

This work has been funded by DFG Priority Program SPP 1335: LE 952/3- 1, DFG LE 952/5-1, and by DFG Me 917/20-1.

Author details

[1]Center for Bioinformatics, Saarland University, Campus E2 1, 66041 Saarbrücken, Germany. [2]Department of Human Genetics, Saarland University, 66421 Homburg/Saar, Germany. [3]Department of Computer Sciences, University of Tuebingen, Sand 13, 72076 Tübingen, Germany. [4]Department of Internal Medicine III, University of Heidelberg, Im Neuenheimer Feld 350, 69120 Heidelberg, Germany. [5]Siemens Healthcare, Hartmannstr. 16, 91052 Erlangen, Germany.

Authors' contributions

CL implemented the underlying framework as well as the web-interface. PL and NL performed the melanoma data analysis. AG and MK developed the BiNA visualization component. JH, BV, and BM tested and benchmarked miRTrail and took part in the writing of this manuscript. HAK collaborated in the project and the overall study and contributed to this manuscript. Support in the implementation of GeneTrail-related components came from CB. CS contributed the conceptual design of this study. EM, HPL, and AK developed the overall study and contributed to this manuscript. AK further supported the analysis and implementation of the herein presented solution. EM, HPL, and AK are the senior authors. All authors read and approved the final manuscript.

Received: 1 October 2011 Accepted: 22 February 2012
Published: 22 February 2012

References

1. Ruvkun G: **Molecular biology. Glimpses of a tiny RNA world.** *Science* 2001, **294(5543)**:797-9.
2. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** *Cell* 1993, **75(5)**:843-54.
3. Zen K, Zhang CY: **Circulating MicroRNAs: a novel class of biomarkers to diagnose and monitor human cancers.** *Med Res Rev* 2010.
4. Iguchi H, Kosaka N, Ochiya T: **Secretory microRNAs as a versatile communication tool.** *Commun Integr Biol* 2010, **3(5)**:478-81.
5. Weber JA, Baxter DH, Zhang S, Huang DY, Huang KH, Lee MJ, Galas DJ, Wang K: **The microRNA spectrum in 12 body fluids.** *Clin Chem* 2010, **56(11)**:1733-41.
6. Medina PP, Slack FJ: **MicroRNAs and cancer: an overview.** *Cell Cycle* 2008, **7**:2485-2492.
7. Zhang B, Pan X, Cobb GP, Anderson TA: **MicroRNAs as oncogenes and tumor suppressors.** *Dev Biol* 2007, **302**:1-12.

8. Roth P, Wischhusen J, Happold C, Chandran A, Hofer S, Eisele G, Weller M, Keller A: **A specific miRNA signature in the peripheral blood of glioblastoma patients.** *J Neurochem* 2011.
9. Keller A, Leidinger P, Borries A, Wendschlag A, Wucherpfennig F, Scheffler M, Huwer H, Lenhof HP, Meese E: **miRNAs in lung cancer: Studying complex fingngerprints in patient's blood cells by microarray experiments.** *BMC Cancer* 2009.
10. Häusler S, Keller A, Chandran A, Ziegler K, Zipp K, Heuer S, Krockenberger M, Engel J, Hönig A, Scheffler M, Dietl J, Wischhusen J: **Whole blood-derived miRNA profiles as potential new tools for ovarian cancer screening.** *Br J Cancer* 2010.
11. Saugstad JA: **MicroRNAs as effectors of brain function with roles in ischemia and injury, neuroprotection, and neurodegeneration.** *J Cereb Blood Flow Metab* 2010, **30(9)**:1564-76.
12. Frost RJ, van Rooij E: **miRNAs as therapeutic targets in ischemic heart disease.** *J Cardiovasc Transl Res* 2010, **3(3)**:280-9.
13. Meder B, Keller A, Vogel B, Sedaghat F, Kayvanpour E, Haas J, Just S, Borries A, Rudloff J, Leidinger P, Meese E, Katus H, Rottbauer W: **MicroRNA Signatures as Novel Biomarkers for Acute Myocardial Infarction.** *Basic Res Cardiol* 2010.
14. Alevizos I, Illei GG: **MicroRNAs as biomarkers in rheumatic diseases.** *Nat Rev Rheumatol* 2010, **6(7)**:391-8.
15. Keller A, Leidinger P, Lange A, Borries J, Schroers H, Scheffler M, Lenhof HP, Ruprecht K, Meese E: **Multiple sclerosis: MicroRNA expression profiles accurately differentiate patients with relapsing-remitting disease from healthy controls.** *PLoS One* 2009.
16. Fiedler SD, Carletti MZ, Christenson LK: **Quantitative RT-PCR Methods for Mature microRNA Expression Analysis.** *Methods Mol Biol* 2010, **630**:49-64.
17. Chen C, Ridzon DA, Broomer A, Zhou Z, Lee DH, Nguyen JT, Barbisin M, Xu NL, Mahuvakar VR, Andersen MR, Lao KQ, Livak KJ, Guegler KJ: **Real-time quantification of microRNAs by stem-loop RT-PCR.** *Nucleic Acids Res* 2005, **33(20)**:e179.
18. Thomson JM, Parker J, Perou CM, Hammond SM: **A custom microarray platform for analysis of microRNA gene expression.** *Nat Methods* 2004, **1**:47-53.
19. Miska1 EA, Alvarez-Saavedra E, Townsend M, Yoshii A, Šestan N, Rakic P, Constantine-Paton M, Horvitz HR: **Microarray analysis of microRNA expression in the developing mammalian brain.** *Genome Biol* 2004, **5**:R68.
20. Motameny S, Wolters S, NÄurnberg P, Schumacher B: **Next Generation Sequencing of miRNAs - Strategies, Resources and Methods.** *Genes* 2010, **1(1)**:70-84.
21. Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M, Eaves CJ, Marra MA: **Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells.** *Genome Res* 2008, **18**:610-621.
22. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS: **Human MicroRNA targets.** *PLoS Biol* 2004, **2**:e363.
23. Wang X, El Naqa IM: **Prediction of both conserved and nonconserved microRNA targets in animals.** *Bioinformatics* 2008, **24(3)**:325-332.
24. Wang X: **miRDB: a microRNA target prediction and functional annotation database with a wiki interface.** *RNA* 2008, **14(6)**:1012-1017.
25. Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, Hughes TR, Blencowe BJ, Frey BJ, Morris QD: **Using expression profiling data to identify human microRNA targets.** *Nat Methods* 2007, **4**:1045-1049.
26. Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, Hatzigeorgiou AG: **The database of experimentally supported targets: a functional update of TarBase.** *Nucleic Acids Res* 2009, , **37 Database:** D155-8.
27. Thomas M, Lieberman J, Lal A: **Desperately seeking microRNA targets.** *Nat Struct Mol Biol* 2010, 1169-74.
28. Su WL, Kleinhanz RR, Schadt EE: **Characterizing the role of miRNAs within gene regulatory networks using integrative genomics techniques.** *Mol Syst Biol* 2011, **7**:490.
29. Jayaswal V, Lutherborrow M, Ma DD, Yang YH: **Identification of microRNA-mRNA modules using microarray data.** *BMC Genomics* 2011, **12**:138.
30. Lu M, Shi B, Wang J, Cao Q, Cui Q: **TAM: A method for enrichment and depletion analysis of a microRNA category in a list of microRNAs.** *BMC Bioinformatics* 2010, **11**:419.
31. Nam S, Li M, Choi K, Balch C, Kim S, Nephew KP: **MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression.** *Nucleic Acids Res* 2009, , **37 Web Server:** W356-W362.
32. Alexiou P, Maragkakis M, Papadopoulos GL, Simmosis VA, Zhang L, Hatzigeorgiou AG: **The DIANA-mirExTra web server: from gene expression data to microRNA function.** *PLoS ONE* 2010, **5(2)**.
33. Nam S, Kim B, Shin S, Lee S: **miRGator: an integrated system for functional annotation of microRNAs.** *Nucleic Acids Res* 2008, , **36 Database:** D159-64.
34. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade1 I, Gunsalus KC, Stoffel M, Rajewsky N: **Combinatorial microRNA target predictions.** *Nat Genet* 2005, **37**:495-500.
35. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.** *Cell* 2005, **120(1)**:15-20.
36. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R: **NCBI GEO: archive for high-throughput functional genomic data.** *Nucleic Acids Res* 2009, , **37 Database:** D5-15.
37. Ruepp A, Kowarsch A, Schmidl D, Buggenthin F, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Theis FJ: **PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes.** *Genome Biol* 2010, **11**:R6.
38. **Computational Prediction Protocol of EMBL-EBI microCosm targets.** [http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/info.html].
39. Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, Müller R, Meese E, Lenhof HP: **GeneTrail - advanced gene set enrichment analysis.** *Nucleic Acids Res* 2007, **Web Server.**
40. Keller A, Backes C, Lenhof HP: **Computation of significance scores of unweighted Gene Set Enrichment Analyses.** *BMC Bioinformatics* 2007, **8.**
41. Keller A, Backes C, Al-Awadhi M, Gerasch A, Kuentzer J, Kohlbacher O, Kaufmann M, Lenhof HP: **GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments.** *BMC Bioinformatics* 2008.
42. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, 2498-504.
43. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics.** *Nucleic Acids Res* 2008, , **36 Database:** D154-D158.
44. Kuentzer J, Blum T, Gerasch A, Backes C, Hildebrandt A, Kaufmann M, Kohlbacher O, Lenhof HP: **BN++ - A Biological Information System.** *J Integr Bioinform* 2006, **3(2)**:34.
45. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD: **Cytoscape Web: an interactive web-based network browser.** *Bioinformatics* 2010.
46. Mueller DW, Bosserhoff AK: **Role of miRNAs in the progression of malignant melanoma.** *Br J Cancer* 2009, **101(4)**:551-6.
47. Kauffmann A, Rosselli F, Lazar V, Winnepenninckx V, Mansuet-Lupo A, Dessen P, van den Oord JJ, Spatz A, Sarasin A: **High expression of DNA repair pathways is associated with metastasis in melanoma patients.** *Oncogene* 2008, **27(5)**:565-73.
48. Ma XH, Piao S, Wang D, McAfee QW, Nathanson KL, Lum J, Li LZ, Amaravadi RK: **Measurements of tumor cell autophagy predict invasiveness, resistance to chemotherapy, and survival in melanoma.** *Clin Cancer Res* 2011, **17(10)**:3478-89.
49. Tormo D, Checińska A, Alonso-Curbelo D, Pérez-Guijarro E, Cañón E, Riveiro-Falkenbach E, Calvo TG, Larribere L, Megías D, Mulero F, Piris MA, Dash R, Barral PM, Rodríguez-Peralto JL, Ortiz-Romero P, Tüting T, Fisher PB, Soengas MS: **Targeted activation of innate immunity for therapeutic induction of autophagy and apoptosis in melanoma cells.** *Cancer Cell* 2009, **16(2)**:103-14.
50. Philippidou D, Schmitt M, Moser D, Margue C, Nazarov PV, Muller A, Vallar L, Nashan D, Behrmann I, Kreis S: **Signatures of microRNAs and selected microRNA target genes in human melanoma.** *Cancer Res* 2010, **70(10)**:4163-73.
51. Caramuta S, Egyházi S, Rodolfo M, Witten D, Hansson J, Larsson C, Lui WO: **MicroRNA expression profiles associated with mutational status and survival in malignant melanoma.** *J Invest Dermatol* 2010, **130(8)**:2062-70.

52. Zhang L, Huang J, Yang N, Greshock J, Megraw MS, Giannakakis A, Liang S, Naylor TL, Barchetti A, Ward MR, Yao G, Medina A, O'brien-Jenkins A, Katsaros D, Hatzigeorgiou A, Gimotty PA, Weber BL, Coukos G: **MicroRNAs exhibit high frequency genomic alterations in human cancer.** *Proc Natl Acad Sci USA* 2006, **103(24)**:9136-41.

doi:10.1186/1471-2105-13-36
Cite this article as: Laczny *et al.*: miRTrail - a comprehensive webserver for analyzing gene and miRNA patterns to enhance the understanding of regulatory mechanisms in diseases. *BMC Bioinformatics* 2012 **13**:36.

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Genomics Proteomics Bioinformatics 10 (2012) 285–294

GENOMICS PROTEOMICS & BIOINFORMATICS

www.elsevier.com/locate/gpb

Original Research

Whole miRNome-wide Differential Co-expression of MicroRNAs

Cord F. Stäehler [1,#], Andreas Keller [1,2,#,*], Petra Leidinger [2], Christina Backes [2], Anoop Chandran [3], Jöerg Wischhusen [3], Benjamin Meder [4], Eckart Meese [2]

[1] *Siemens Healthcare, Strategy, 91052 Erlangen, Germany*
[2] *Department of Human Genetics, Saarland University, 66421 Homburg, Germany*
[3] *Interdisciplinary Center for Clinical Research, University of Wuerzburg, 97070 Wuerzburg, Germany*
[4] *Department of Internal Medicine III, Heidelberg University, 69115 Heidelberg, Germany*

Received 3 May 2012; revised 4 June 2012; accepted 10 June 2012
Available online 23 August 2012

Abstract

Co-regulation of genes has been extensively analyzed, however, rather limited knowledge is available on co-regulations within the miRNome. We investigated differential co-expression of microRNAs (miRNAs) based on miRNome profiles of whole blood from 540 individuals. These include patients suffering from different cancer and non-cancer diseases, and unaffected controls. Using hierarchical clustering, we found 9 significant clusters of co-expressed miRNAs containing 2–36 individual miRNAs. Through analyzing multiple sequencing alignments in the clusters, we found that co-expression of miRNAs is associated with both sequence similarity and genomic co-localization. We calculated correlations for all 371,953 pairs of miRNAs for all 540 individuals and identified 184 pairs of miRNAs with high correlation values. Out of these 184 pairs of miRNAs, 16 pairs (8.7%) were differentially co-expressed in unaffected controls, cancer patients and patients with non-cancer diseases. By computing correlated and anti-correlated miRNA pairs, we constructed a network with 184 putative co-regulations as edges and 100 miRNAs as nodes. Thereby, we detected specific clusters of miRNAs with high and low correlation values. Our approach represents the most comprehensive co-regulation analysis based on whole miRNome-wide expression profiling. Our findings further decrypt the interactions of miRNAs in normal and human pathological processes.

Keywords: Co-expression; Microarray; MicroRNA; Network analysis

Introduction

Microarray experiments have been applied for almost three decades in the detection of disease-relevant changes in gene expression patterns. While in early ages genes have mostly been considered independently from each other, cluster [1,2] and classification [3–5] technologies have more recently been applied to find patterns of differentially expressed genes. Finally, gene set analysis approaches [6,7] and methods integrating pathway topology [8,9] have been developed to understand the interplay of genes. These approaches have been successfully applied to studies in small non-coding RNAs, *e.g.*, microRNAs (miRNAs).

With the increasing availability of expression profiles for various diseases, differential co-expression of genes moved into the focus of attention. The term "differential co-expression" was firstly coined by Bennets in 1986 [10] studying the co-expression of alpha-actins within the human heart. In 1992, Swiderski reported differential co-expression of long and short form type IX collagen transcripts during avian limb chondrogenesis [11]. Co-expression analysis of genes using microarray technology has also been applied to other human pathologies, including cancer [12]. In 2009, Mo and co-workers presented a stochastic model to identify co-expression patterns of differential gene pairs in prostate cancer progression [13]. Comprehensive methods to detect differential co-

# Equal contribution.
* Corresponding author.
E-mail: ack@bioinf.uni-sb.de (Keller A).

http://dx.doi.org/10.1016/j.gpb.2012.08.003

expression have been developed by Lai who reported an efficient pattern recognition algorithm [14]. This algorithm used Expected Conditional F-statistic that incorporates statistical information of location and correlation or other scores as proposed by Koska and Spang [15]. Subsequently, several tools and software packages with respective functionality have been developed including CoXpress [16], DiffCoEx [17], dCoxS [18] and differential co-expression framework [19].

Only a few studies have been reported for analysis of differential co-expression for miRNAs. An example is the construction of an miRNA–miRNA synergistic network via co-regulating functional modules and disease miRNA topological features [20]. One reason for the lack of miRNA co-expression studies is certainly the paucity of miRNA expression profiling data. Gene expression profiles have been measured for almost three decades in numerous microarray experiments, of which hundreds of thousands are currently available through the Gene Expression Omnibus [21,22], however, only a fraction of array data sets are available for miRNAs. The most frequently applied microarray platform is the Agilent miRNA microarray 2.0. Another technology which is frequently applied is the MPEA assay (Febit Biomed, Heidelberg) that has been used to measure several hundred blood-based miRNA profiles which are the source for our meta-analysis.

Previously, Riveros and co-workers reported a comprehensive study for differential co-expression of miRNA that was derived from whole blood of patients with multiple sclerosis [23], providing evidence that differential co-expression from body fluids can be accessed. miRNA expression patterns from human blood cells are increasingly discussed for their potential as a minimal invasive diagnostic tool. Most recently, we reported blood-based miRNA expression patterns for 14 different human pathologies [24] including lung cancer [25], COPD [26], multiple sclerosis [27], ovarian cancer [28], glioblastoma [29], and acute myocardial infarction [30]. Since the various cohorts are relatively small as compared to the large number of potential pair-wise co-expressions, we combined the different data sets into a meta-analysis. Here, we investigate the differential co-expression patterns using the data of a total of 540 blood-based miRNA expression profiles.

Results and discussion

Co-localization and co-expression of miRNAs

As a first approach towards understanding the interplay of miRNAs, we applied hierarchical clustering to the data set containing 540 samples measured for the expression of 863 miRNAs. To reduce the noise, we excluded miRNAs with low expression values (detailed in Material and methods). An average linkage bottom up clustering detected a total of nine significant clusters ($P < 0.05$). These clusters each contain 2–36 miRNAs (**Figure 1**). Notably, many clusters contained miRNAs with similar sequences. Good example for co-expression related to similar sequences is Cluster 8 that contains hsa-miR-23a and hsa-miR-23b or Cluster 5 that contains hsa-miR-19a and hsa-miR-19b. The biological mechanism underlying co-expression of miRNAs with similar sequence remains to be elucidated. It is possible that

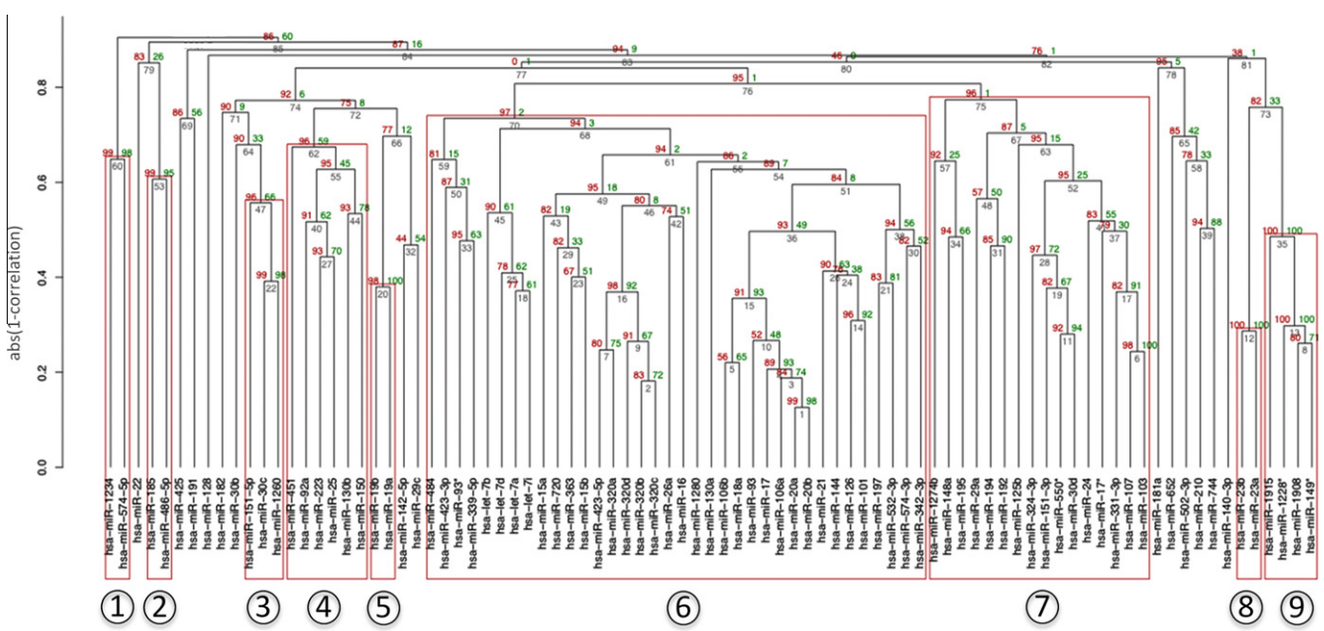


Figure 1 Cluster dendrogram of miRNAs

Red boxes denote significant clusters as computed by bootstrap re-sampling. The red values were calculated by bootstrap re-sampling and those >95%, corresponding to significance level of 0.05, are considered as significant. Values in green and gray indicate bootstrap probability (BP) and the edge number in the dendrogram, respectively. The significant clusters with approximately unbiased (AU) value greater than 95% ($P < 0.05$) are labeled with numbers in circle in increasing order from left to right.

```
hsa-miR-23b     GGTAATCCCTGGCAATGTGAT 21
hsa-miR-23a     GGAAATCCCTGGCAATGTGAT 21
                ** ******************

hsa-miR-19b     TCAGTTTTGCATGGATTTGCACA 23
hsa-miR-19a     TCAGTTTTGCATAGATTTGCACA 23
                *********** **********

hsa-miR-30c     --GCTGAGAGTGTAGGATGTTTACA 23
hsa-miR-1260    TGGTGGCAGAGGTGGGAT------- 18
                  *  *      ** ****
```

Figure 2 Alignments of co-expression miRNA clusters with similar or different sequences

Pairwise sequence alignment indicated that hsa-miR-23a and hsa-miR-23b in Cluster 8 (upper panel) and hsa-miR-19a and hsa-miR-19b in Cluster 5 (middle panel) show high sequence similarity, while there is lower sequence similarity for hsa-miR-1260 and hsa-miR-30c in Cluster 3 (lower panel).

Table 1 Co-localization of correlated miRNAs

miRNA	Position	Strand	Correlation
hsa-miR-20a	13:90801320	+	0.79
hsa-miR-17	13:90800860	+	
hsa-miR-20b	X:133131505	−	0.79
hsa-miR-106a	X:133131894	−	
hsa-miR-18a	13:90801006	+	0.74
hsa-miR-20a	13:90801320	+	
hsa-miR-423–5p	17:25468223	+	−0.56
hsa-miR144	17:24212677	−	
hsa-miR-423–5p	17:25468223	+	−0.51
hsa-miR-21	17:55273409	+	

co-expressed miRNAs of similar sequence share similar targets. Other than that, reduced specificity of hybridization-based approaches could partially explain this co-expression. On the other hand, as expected, we also found many miRNAs that clustered together but had different sequences, such as hsa-miR-1260 and hsa-miR-30c in Cluster 3. Respective pair-wise sequence alignments (hsa-miR-19a/hsa-miR-19b, hsa-miR23a/hsa-miR23b, hsa-miR-1260/hsa-miR-30c) are shown in **Figure 2**.

To test the hypothesis that miRNAs belonging to the same polycistronic miRNA cluster or the same miRNA family are co-expressed, we additionally performed enrichment analyses. For each significant set containing more than 5 miRNAs (Clusters 4, 6 and 7 in Figure 1), we performed the enrichment analysis separately to see whether the selected miRNA clusters or families are over-represented. In line with our expectations, the let-7a, miR-106a, miR-106b, miR-15a and miR-17 clusters were significantly enriched (all $P \leqslant 0.005$) in our Cluster 6, whereas members of the miR-192 polycistronic miRNA cluster were mostly found in Cluster 7 ($P = 0.001$). Likewise, we also found a strong enrichment of miRNA families in our clusters, such as the let-7 family ($P = 0.002$), the miR-15 family ($P = 0.001$), the miR-320 family ($P = 0.00002$) and the miR-17 family ($P = 3E\text{-}8$) in Cluster 6 and the miR-103 family ($P = 0.001$) in Cluster 7. Interestingly, no significant enrichment for a known miRNA cluster or family was found in Cluster 4, indicating that our clustering approach groups not only polycistronic (and thus co-transcribed) miRNA clusters or known miRNA families, but also miRNAs that are co-expressed for different reasons. In addition, divergent behavior of individual

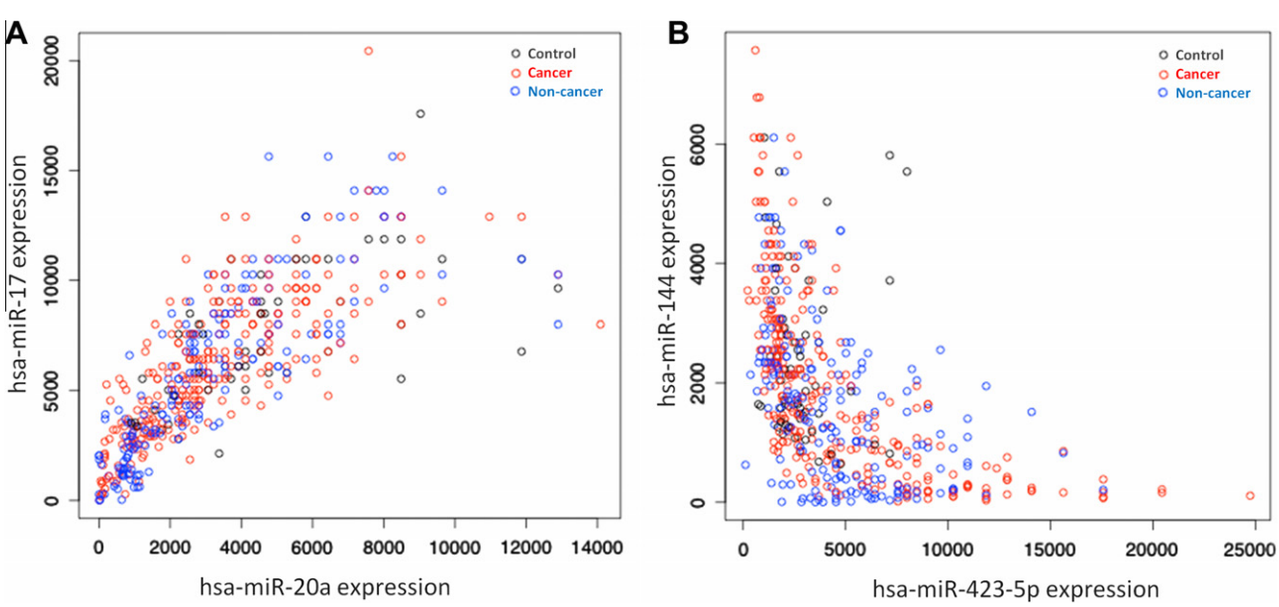


Figure 3 Representative expression profiles of correlated miRNA pairs

A. Positive correlation. Expression of two positively-correlated miRNAs, hsa-miR-17 and hsa-miR-20a, was measured for 540 individuals including controls ($n = 72$, black circle), cancer patients ($n = 276$, red circle) and non-cancer patients ($n = 192$, blue circle). **B.** Negative correlation. Expression of two negatively-correlated miRNAs, hsa-miR-423-5p and hsa-miR-144, was measured for 540 individuals including controls (black circle), cancer patients (red circle) and non-cancer patients (blue circle). A complete list of the disease types and the respective numbers of patients is shown in Table 4.

Table 2 Differential co-expression in diseases

miRNA 1	miRNA 2	Overall correlation	Control correlation	Cancer correlation	Non cancer correlation	Variance
hsa-miR-377*	hsa-miR-548l	0.756	−0.129	0.522	0.842	0.245
hsa-miR-196b	hsa-miR-548l	0.73	−0.028	0.416	0.864	0.199
hsa-miR-548l	hsa-miR-135b	0.72	−0.023	0.373	0.863	0.197
hsa-miR-423-5p	hsa-miR-144	−0.556	0.052	−0.651	−0.462	0.132
hsa-miR-595	hsa-miR-574-5p	0.743	0.121	0.78	0.62	0.118
hsa-miR-363	hsa-miR-320d	−0.507	0.019	−0.481	−0.624	0.114
hsa-miR-320c	hsa-miR-363	−0.519	−0.022	−0.539	−0.56	0.093
hsa-miR-320b	hsa-miR-144	−0.575	−0.078	−0.619	−0.582	0.091
hsa-miR-106a	hsa-miR-720	−0.577	−0.128	−0.626	−0.582	0.076
hsa-miR-106a	hsa-miR-320c	−0.504	−0.068	−0.499	−0.568	0.073
hsa-miR-144	hsa-miR-320a	−0.571	−0.143	−0.642	−0.542	0.07
hsa-miR-126	hsa-miR-720	−0.513	−0.079	−0.537	−0.534	0.069
hsa-miR-144	hsa-miR-720	−0.552	−0.144	−0.584	−0.56	0.061
hsa-miR-720	hsa-miR-20b	−0.58	−0.187	−0.599	−0.602	0.057
hsa-miR-23b	hsa-miR-23a	0.708	0.323	0.751	0.709	0.056
hsa-miR-17	hsa-miR-151-3p	−0.512	−0.205	−0.378	−0.657	0.052

miRNAs belonging to the same polycistronic cluster or family provides evidence for a significant post-transcriptional component in miRNA expression.

An additional reason for putative co-regulation of miRNAs might be their co-localization in the genome. To this end, we searched for miRNAs that have been clustered together based on the expression data and are located on the same chromosome. Subsequently candidate pairs were mapped to the exact chromosomal position. We found five pairs of miRNAs that showed a high absolute correlation ($\leqslant -0.5$ or $\geqslant 0.5$) and are located on the same chromosome, as presented in **Table 1**. Three of those five miRNA pairs

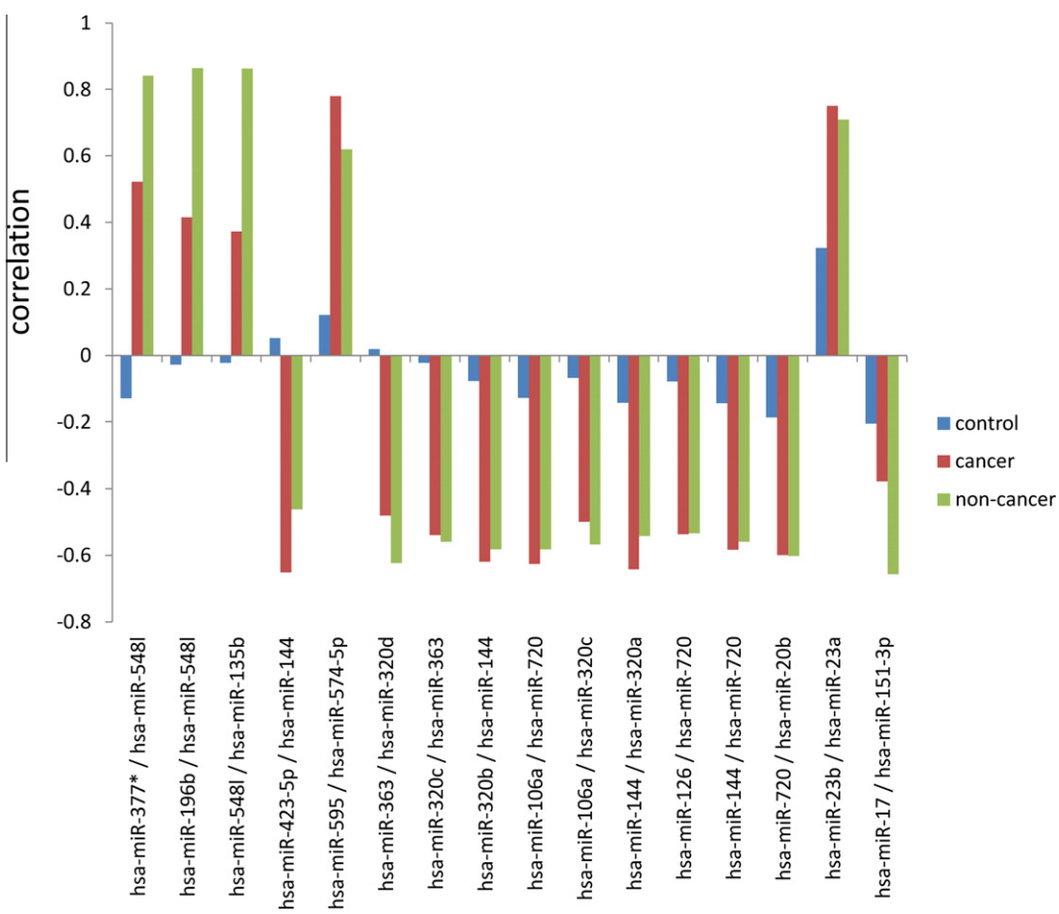


Figure 4 Correlations of 16 miRNA pairs with variance > 0.05
Differential co-expression of these miRNA pairs is shown separately for cancer patients (red), non-cancer patients (green) and healthy controls (blue). Co-expression of the miRNA pairs was more frequently detected in cancer and non-cancer patients than in healthy controls.

Table 3 Non-differential co-expression in diseases

miRNA 1	miRNA 2	Overall correlation	Control correlation	Cancer correlation	Non-cancer correlation	Variance
hsa-miR-593*	hsa-miR-646	0.791	0.813	0.803	0.77	<0.001
hsa-miR-93	hsa-miR-20b	0.737	0.758	0.751	0.721	<0.001
hsa-miR-593*	hsa-miR-214	0.834	0.815	0.824	0.852	<0.001
hsa-miR-330-3p	hsa-miR-621	0.858	0.887	0.859	0.851	<0.001
hsa-miR-593*	hsa-miR-331-3p	−0.503	−0.5	−0.488	−0.523	<0.001
hsa-miR-374b	hsa-miR-374a	0.722	0.729	0.734	0.702	<0.001
hsa-miR-621	hsa-miR-593*	0.838	0.852	0.849	0.822	<0.001
hsa-miR-330-3p	hsa-miR-214	0.801	0.825	0.793	0.815	<0.001
hsa-miR-452*	hsa-miR-593*	0.748	0.754	0.756	0.727	<0.001
hsa-miR-500	hsa-miR-195	−0.532	−0.509	−0.54	−0.52	<0.001
hsa-miR-1228*	hsa-miR-149*	0.719	0.755	0.731	0.73	<0.001
hsa-miR-107	hsa-miR-331-3p	−0.67	−0.685	−0.658	−0.675	<0.001
hsa-miR-330-3p	hsa-miR-452*	0.793	0.809	0.793	0.792	<0.001
hsa-miR-509-5p	hsa-miR-933	0.842	0.854	0.841	0.848	<0.001
hsa-miR-584	hsa-miR-362-5p	0.713	0.709	0.703	0.715	<0.001
hsa-miR-1184	hsa-let-7i*	0.792	0.786	0.788	0.792	<0.001

showed positive correlation while the remaining two pairs showed negative correlation. The three pairs with positive correlation are located within a distance of 500 base pairs of each other and were each on the same strand. On the other hand, larger genomic distances were found for the two negatively-correlated miRNA pairs. For example, the distance between hsa-miR-423-5p and hsa-miR-144 was about 10 million base pairs (Mb). Moreover, hsa-miR-423-5p was located on the plus strand whereas hsa-miR-144 was located on the minus strand. **Figure 3** shows expression values of one pair of positively-correlated miRNAs, namely hsa-miR-20a/hsa-miR-17 (Figure 3A) and one pair of negatively-correlated miRNAs, namely hsa-miR-423-5p/hsa-miR-144 (Figure 3B) for 540 analyzed

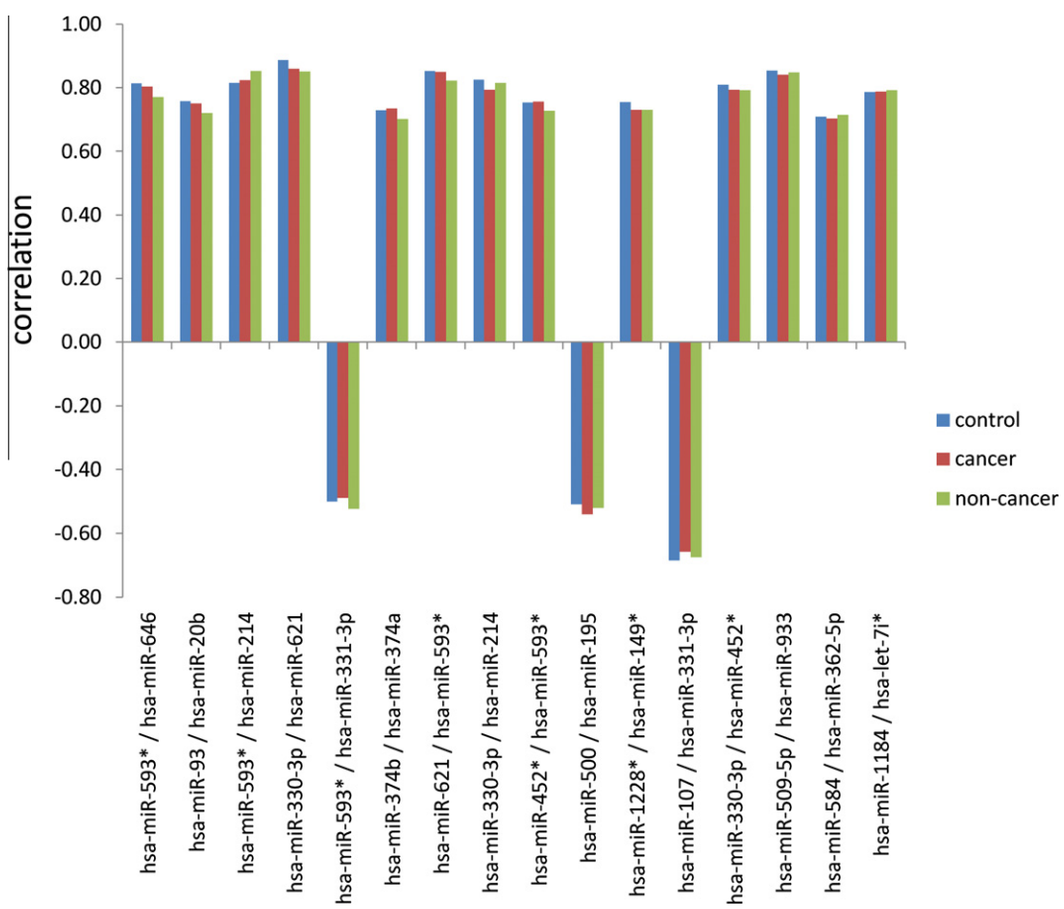


Figure 5 Correlations of 16 miRNA pairs with variance ⩽ 0.00053
Differential co-expression of these miRNA pairs is shown separately for cancer patients (red), non-cancer patients (green) and healthy controls (blue). Co-expression of the miRNA pairs was comparable in the three groups.

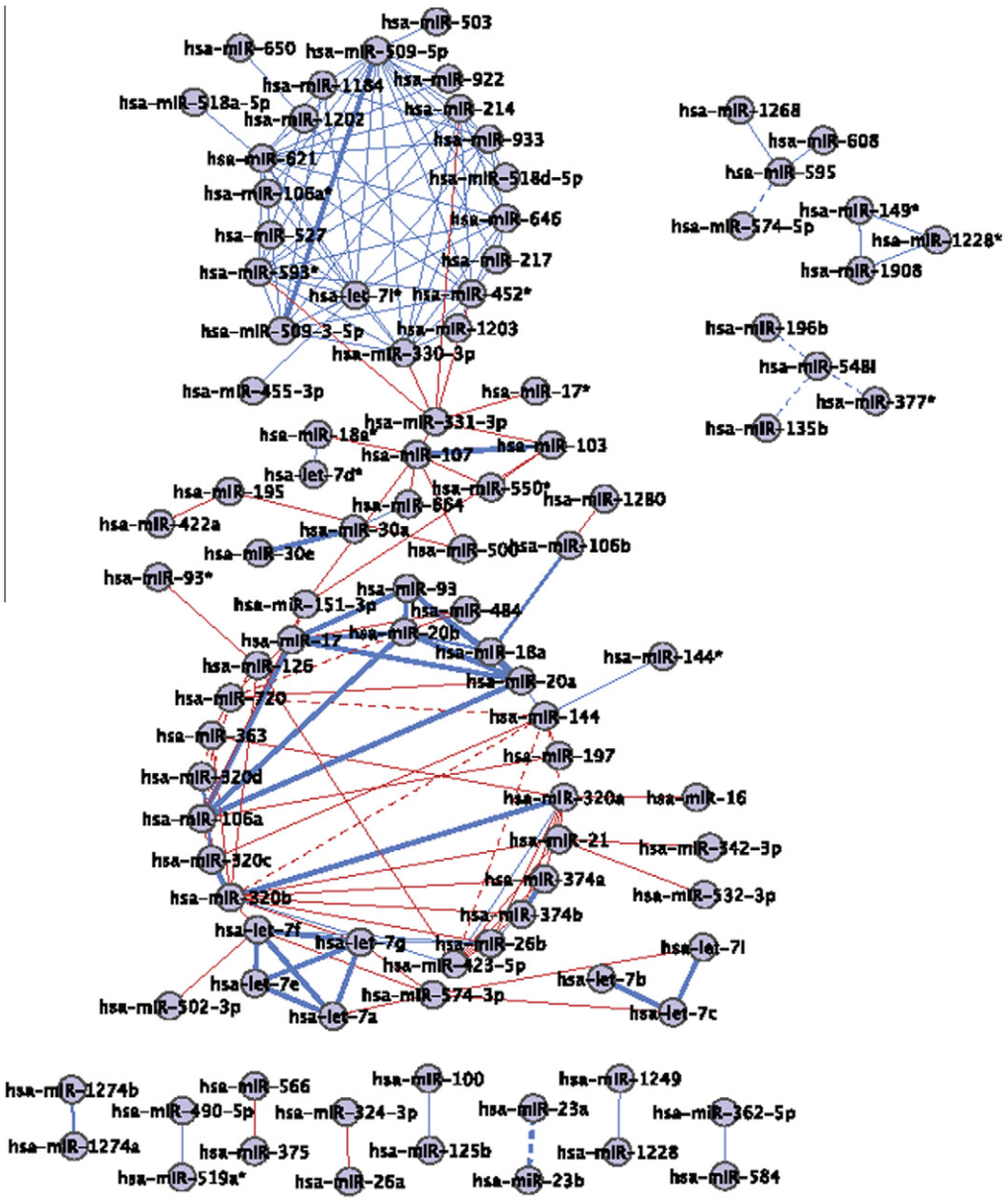


Figure 6 Correlation network

Network was constructed with all positive correlations between pairs of miRNAs with a correlation of at least 0.7 (blue edges) and all negative correlations of at least −0.5 (red edges). The thickness of the edges corresponds to the alignment score. The pair-wise sequence alignment was computed using edit distance. Dashed edges indicate correlations that were different in controls and in patients.

blood samples. The results showed that the cohorts behaved similarly for each of the pairs.

Differential co-expression of miRNAs

The 540 individuals participating in this study can be grouped in three different cohorts, including unaffected healthy individuals (control), cancer patients (cancer) and non-cancer patients (non cancer). For these three cohorts we asked whether the correlation is equally high in all three groups or whether certain cohorts deviate from the others. To this end, we computed for each pair of miRNAs the correlation values for the three cohorts separately. As a result of the calculation for all $\binom{863}{2} = \frac{863 * 862}{2} = 371,953$ pairs, the values of correlation range from −0.67 to 0.89 with average correlation of 0.013. As the slight positive average correlation already indicates, we obtained slightly more positive correlations than negative ones. Thus, we applied different thresholds for positive and negative correlations to acknowledge this non-symmetric distribution. We only considered positively-correlated miRNA pairs with correlation values higher than 0.7 and negatively-correlated miRNA pairs with values lower than −0.5. Using these thresholds we obtained 184 miRNA pairs out of 371,953 pairs in total

(0.05%). Of these 184 miRNA pairs, 118 were positively correlated and 66 were negatively correlated. To estimate the extent of differential expression in the 3 cohorts, we computed the variance of the correlation values, ranging from 10^{-5} to 0.24 with an average of 0.02. The 16 miRNA pairs with the highest variance, corresponding to the most differentially-regulated miRNAs (variance >0.05), are summarized in **Table 2** and **Figure 4** and the 16 miRNA pairs with the lowest variance (variance ⩽0.00053) are indicated in **Table 3** and **Figure 5**.

By examining the differential co-expression of the 16 miRNA pairs with variance >0.05, we found that both the cancer patients and the non-cancer patients deviate from the healthy controls. As compared to the healthy controls, co-expression of these 16 miRNA pairs was detected significantly more frequently in both cancer and non-cancer disease groups. Overall the correlation between cancer and non-cancer diseases was 0.95 while decreased correlation was revealed between control and cancer and between control and non-cancer diseases, which is 0.59 and 0.49, respectively. Further analysis identified five miRNA pairs that were positively correlated in patients but not in healthy controls. For example, the pair hsa-miR-23a/hsa-miR-23b showed correlation of 0.71 in non-cancer patients ($P < 10^{-10}$) and 0.75 in cancer patients ($P < 10^{-10}$) but only 0.32 in healthy controls ($P < 0.01$) with the respective 95% confidence intervals as 0.69–0.80, 0.63–0.79 and 0.07–0.54. Moreover, we found 11 miRNAs that were highly anti-correlated, *i.e.*, negatively correlated both in cancer and in non-cancer patients but again not correlated in healthy controls.

These results indicate that the observed overall high variance for the 16 miRNA pairs is mostly due to the healthy controls. While the 16 pairs were only weakly correlated in healthy controls they were correlated or anti-correlated in the patients. The results that may be biased due to slightly different cohort sizes may give first and certainly only preliminary evidence that miRNA expression may be more homogenously coordinated in patients, possibly indicating a change in expression regulation that is common to different types of diseases.

Interestingly, all miRNAs of the miRNA pairs that are negatively correlated in cancer or non-cancer patients but not in healthy controls have been previously related to human diseases according to the Human miRNA and Disease Database (HMDD) [31]. For examples, the known disease-associated miRNAs that were identified as negatively correlated in our study included hsa-miR-17 that was according to the HMDD associated with 33 different diseases, hsa-miR-128 with 18 diseases, hsa-miR-20b with 9 diseases, hsa-miR-423 with 4 diseases and hsa-miR-363 with 3 diseases. This finding is even more profound, since only about one third of all known miRNAs in the HMDD are related to one or more diseases [31]. Obviously, the analysis of co-expression can contribute to the identification of disease-associated miRNAs.

Anti-correlation of expression and co-localization of miRNAs

Besides pairs of miRNAs that were correlated in patients and that were co-localized, we also identified co-localization of miRNA pairs for miRNAs which are anti-correlated in patients. For example, hsa-miR-423-5p and hsa-miR-144 are co-localized on chromosome 17 and are negatively correlated (−0.56). Specifically, the correlation value for this miRNA pair was −0.65 in non-cancer patients and −0.46 in cancer patients, respectively. However, we did not find a negative correlation for this miRNA pair in healthy controls (correlation of 0.05) (Tables 1 and 2).

Putative co-regulation network

Based on the analysis of co-regulated miRNAs, we constructed a network with 184 correlations (correlation value >0.7 or <−0.5). As shown in **Figure 6**, the derived network

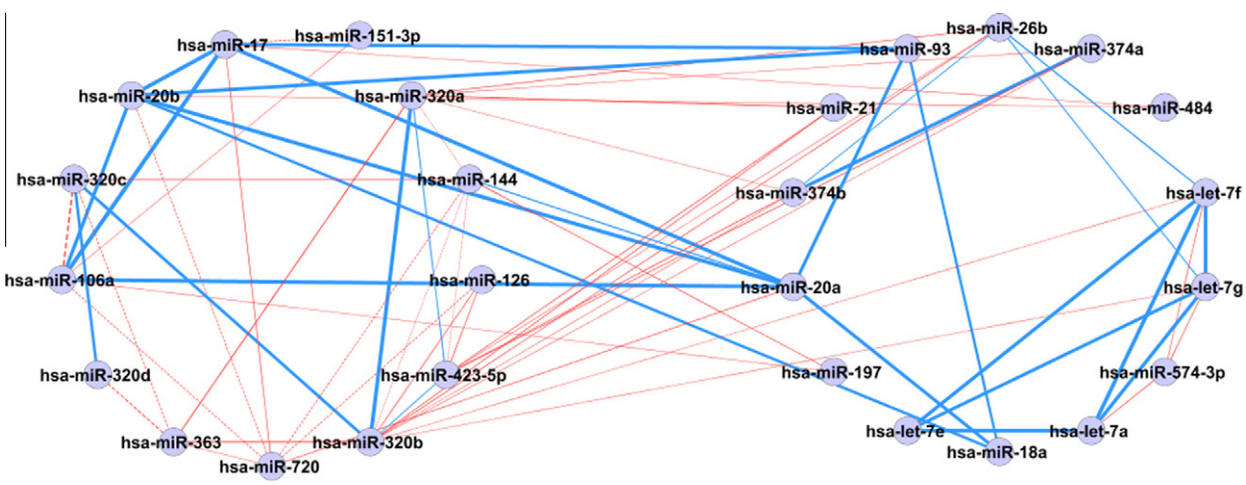


Figure 7 Differential co-expression sub-network
Sub-network of miRNAs was constructed with correlations that were different between controls and patients (left side) and correlations that were similar between controls and patients (right side). The indication of edges was same as that used in Fig. 6.

Table 4 Cohort characteristics

Class	Disease	No. of samples
Control	Healthy	72
Cancer	Lung tumor	35
	Ductal adenocarcinoma	45
	Melanoma	35
	Ovarian cancer	15
	Prostate carcinoma	35
	Wilms tumor	50
	Other pancreatic tumors	48
	Tumor of stomach	13
Non cancer	Multiple sclerosis	23
	Sarcoidosis	45
	Periodontitis	18
	COPD	27
	Myocardial infarction	20
	Pancreatitis	37
	Benign prostate hyperplasia	22
Sum		540

Note: COPD, chronic obstructive pulmonary disease.

encompasses 100 different miRNAs. Roughly, the network can be divided into one large connected component and several small components consisting of 2 to 4 miRNAs each. The large connected component can again be subdivided into two clusters. The upper cluster shown in Figure 6 contains mostly positively correlated miRNA pairs as indicated by blue edges while the cluster shown at the bottom of Figure 6 contains both positive correlations and negative correlations (indicated by red edges). While the miRNAs in the upper cluster do not show obvious sequence similarity as indicated by thin edges, positively correlated miRNAs of the lower cluster show high sequence similarity as indicated by thick edges. The small components with two to four miRNAs are mostly positively correlated. For most of these pairs, the positive correlation is associated with sequence similarities, as for example for the pair of hsa-miR-23a and hsa-miR-23b and the pair of hsa-miR-1247a and hsa-miR-1247b.

Additionally, many of the positive and negative correlations shown in the bottom cluster are different between healthy controls and patients. The differences are visualized as a sub-network in **Figure 7**. The sub-network separates miRNAs with different correlation between healthy controls and patients (on the left) and miRNAs with similar correlations in healthy controls and patients (on the right). Among the miRNAs with similar correlations are four miRNAs of the let-7 family that have previously been associated with many human malignancies [31]. In addition, each of the remaining miRNAs of the sub-network has previously been associated with at least one human disease according to the HMDD [31].

Conclusion

Over almost three decades it has been shown that co-expression and specifically differential co-expression of genes play an important role in human pathogenic processes. However, differential co-expression has not been thoroughly analyzed for the miRNome. This is in part due to the lack of respective high-throughput data sets allowing the analysis of miRNA-miRNA interactions. We enlarged a recently published set of 454 whole miRNome profiles [24] to a total of 540 profiles. Analysis of the miRNA co-expression from these profiles provides supporting evidence that genomic localization and sequence similarity are associated with co-expression. In addition, we report a significantly enriched clustering for miRNAs that belong to the same miRNA families or polycistronic miRNA clusters. Moreover, our findings also support that the co-expression may be more pronounced in patients, compared to the healthy controls. Network based analysis allows us to detect specific clusters of miRNAs with high and low correlation. Interestingly, many of the identified differentially co-expressed miRNAs have previously been associated with human pathogenic processes. Notably, the reported data have not been measured from tissues but from blood cells. Since different tissues have specific miRNA profiles, a co-expression analysis would make sense only for one tissue type but not enabling a meta-analysis of different diseases. However, different blood cell compositions in different diseases might influence the overall result of our meta-analysis. Another limitation of our study is certainly the applied microarray technology. Novel approaches such as next-generation small RNA sequencing will likely improve the specificity of respective analyses in the future.

In summary, in human diseases, co-expression and differential co-expression of miRNAs seems to be of similar importance to co-expression of protein-coding genes.

Materials and methods

Patients

The screened cohort contains a total of 540 subjects including healthy controls ($n = 72$), cancer patients ($n = 276$) and patients with non-cancer diseases ($n = 192$). The detailed characteristics of the cohort are listed in **Table 4**. All blood donors participating in this study signed the informed consent form and the local ethics committee approved the analysis of miRNA expression in blood. Blood samples were collected using PAXgene Blood RNA tubes (BD, Franklin Lakes, New Jersey, USA).

miRNA extraction and microarray screening

Total RNA isolation was performed using miRNeasy Mini Kit (Qiagen) as described previously [25].

Microarray analyses were done on the Geniom RT Analyzer using Geniom miRNA Biochips (Febit Biomed GmbH). Each of the 863 human miRNAs (Sanger miRBase v12.0 to v15.0) was present in at least seven replicates

on each of the 540 arrays. The screening was done using micro-fluidics primer extension assay (MPEA) [32]. This assay differs from a standard hybridization assay in that it includes an additional primer extension step. The MPEA assay is verified to be very sequence-specific at the end of miRNAs and thus minimizes the cross-hybridization effect, which is of especially high importance for our cross-hybridization study.

Data processing and bioinformatics analysis

At the first preprocessing step, all arrays were locally background corrected. Then, the replicates of each miRNA on each microarray were merged by computing the median intensity value. To account for between-array effects, standard quantile normalization was performed [33], which showed superior performance as compared to normalization via spike-in or housekeeping miRNAs in previous experiments. All computations then were carried out on the expression matrix containing 863 rows representing 863 miRNAs and 540 columns representing the 540 individuals. The statistical analyses were carried out using R [34] if not mentioned otherwise. To reduce the noise, we excluded miRNAs with low expression values, *i.e.*, the median signal intensity of a miRNA must be great than 500 for a specific miRNA to be considered in our study.

For hierarchical clustering, the pvclust package has been used. The package computes *P* values for hierarchical clustering based on a multiscale bootstrap resampling, helping to interpret clusters. Specifically, clusters that are highly supported by the data will have low *P* values while weaker clusters end up with non-significant *P* values. Significant clusters are enclosed with red boxes in the respective dendrogram. We used *1-abs(cor(x,y))* as a distance measure for the clustering, where *cor(x,y)* corresponds to the Pearson correlation coefficient of all 540 observations for two miRNAs x and y. By using this distance function, we detect miRNAs that are highly correlated and anti-correlated. In more detail, an average linkage bottom up clustering was carried out.

To compute differential expression and differential co-expression, we again used the pair-wise Pearson correlation of all 863 miRNAs, for all 540 samples together but also for the different groups of controls, cancers and non-cancer diseases, separately. As a result of this analysis, we calculated for $\binom{863}{2} = \frac{863 * 862}{2} = 371,953$ miRNA pairs four different correlation values, the overall value and the single values for the three groups. To find the miRNA pairs with different behavior in different groups, we computed the variance of the 371,953 pairs as $\frac{1}{2}((cor_{control,i} - cor_i)^2 + (cor_{cancer,i} - cor_i)^2(cor_{non-cancer,i} - cor_i^2))$, where $cor_{control,i}$ corresponds to the correlation of control samples for a miRNA pair i, $cor_{cancer,i}$ corresponds to the correlation of cancer samples for a miRNA pair i, $cor_{non\text{-}cancer,i}$ corresponds to the correlation of control samples for a miRNA pair i, and cor_i corresponds to the average of the three correlation values for miRNA pair i.

Empirically, we considered miRNA pairs with correlations above 0.7 as highly co-expressed and with correlations below -0.5 as anti-correlated. Using Cytoscape [35] we visualized the network of these miRNA pairs and visualized the results using Cytoscapes mapping functionality with an organic layout.

Authors' contributions

CS contributed to the study design, data evaluation and wrote the manuscript; AK designed the study, wrote the manuscript and carried out computations; PL carried out the experiments; CB contributed to data analysis and revision of the manuscript; AC contributed to the analysis of miRNA clusters and families; JW contributed to study design and revision of the manuscript; BM assisted the data analysis and reviewed the manuscript; EM designed the study, wrote the manuscript. All authors read and approved the final manuscript.

Conflict of interest

CS and AK are affiliates of Siemens Healthcare, Erlangen, Germany.

Acknowledgements

We acknowledge all patients participating in this study. Financial support was granted by HOMFOR, Deutsche Forschungsgemeinschaft (DFG, LE 2783/1-1), and Hedwig-Stalter Foundation.

References

[1] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95:14863–8.
[2] De Bin R, Risso D. A novel approach to the clustering of microarray data via nonparametric density estimation. BMC Bioinformatics 2011;12:49.
[3] Zhang J, Liu B, Jiang X, Zhao H, Fan M, Fan Z, et al. A systems biology-based gene expression classifier of glioblastoma predicts survival with solid tumors. PLoS ONE 2009;4:e6274.
[4] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 286:531–7.
[5] Gill R, Datta S. A statistical framework for differential network analysis from microarray data. BMC Bioinformatics 2010;11:95.
[6] Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, et al. GeneTrail – advanced gene set enrichment analysis. Nucleic Acids Res 2007;35:W186–92.
[7] Al-Shahrour F, Minguez P, Vaquerizas JM, Conde L, Dopazo J. BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. Nucleic Acids Res 2005;33:W460–4.
[8] Rahnenfuhrer J, Domingues FS, Maydt J, Lengauer T. Calculating the statistical significance of changes in pathway activity from gene expression data. Stat Appl Genet Mol Biol 2004; 3:Article 16.

[9] Keller A, Backes C, Gerasch A, Kaufmann M, Kohlbacher O, Meese E, et al. A novel algorithm for detecting differentially regulated paths based on gene set enrichment analysis. Bioinformatics 2009;25:2787–94.
[10] Bennetts BH, Burnett L, dos Remedios CG. Differential co-expression of alpha-actin genes within the human heart. J Mol Cell Cardiol 1986;18:993–6.
[11] Swiderski RE, Solursh M. Differential co-expression of long and short form type IX collagen transcripts during avian limb chondrogenesis in ovo. Development 1992;115:169–79.
[12] Choi JK, Yu U, Yoo OJ, Kim S. Differential coexpression analysis using microarray data and its application to human cancer. Bioinformatics 2005;21:4348–55.
[13] Mo WJ, Fu XP, Han XT, Yang GY, Zhang JG, Guo FH, et al. A stochastic model for identifying differential gene pair co-expression patterns in prostate cancer progression. BMC Genomics 2009;10:340.
[14] Lai Y, Wu B, Chen L, Zhao H. A statistical method for identifying differential gene-gene co-expression patterns. Bioinformatics 2004;20:3146–55.
[15] Kostka D, Spang R. Finding disease specific alterations in the co-expression of genes. Bioinformatics 2004;20:i194–9.
[16] Watson M. CoXpress: differential co-expression in gene expression data. BMC Bioinformatics 2006;7:509.
[17] Tesson BM, Breitling R, Jansen RC. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. BMC Bioinformatics 2010;11:497.
[18] Cho SB, Kim J, Kim JH. Identifying set-wise differential co-expression in gene expression microarray data. BMC Bioinformatics 2009;10:109.
[19] Chia BK, Karuturi RK. Differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms. Algorithms Mol Biol 2010;5:23.
[20] Xu J, Li CX, Li YS, Lv JY, Ma Y, Shao TT, et al. MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features. Nucleic Acids Res 2011;39:825–36.
[21] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10.
[22] Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets – 10 years on. Nucleic Acids Res 2011;39:D1005–10.
[23] Riveros C, Mellor D, Gandhi KS, McKay FC, Cox MB, Berretta R, et al. A transcription factor map as revealed by a genome-wide gene expression analysis of whole-blood mRNA transcriptome in multiple sclerosis. PLoS ONE 2010;5:e14176.
[24] Keller A, Leidinger P, Bauer A, ElSharawy A, Haas J, Backes C, et al. Toward the blood-borne miRNome of human diseases. Nat Methods 2011;8:841–3.
[25] Keller A, Leidinger P, Borries A, Wendschlag A, Wucherpfennig F, Scheffler M, et al. MiRNAs in lung cancer – studying complex fingerprints in patient's blood cells by microarray experiments. BMC Cancer 2009;9:353.
[26] Leidinger P, Keller A, Borries A, Huwer H, Rohling M, Huebers J, et al. Specific peripheral miRNA profiles for distinguishing lung cancer from COPD. Lung Cancer 2011;74:41–7.
[27] Keller A, Leidinger P, Lange J, Borries A, Schroers H, Scheffler M, et al. Multiple sclerosis: microRNA expression profiles accurately differentiate patients with relapsing-remitting disease from healthy controls. PLoS ONE 2009;4:e7440.
[28] Hausler SF, Keller A, Chandran PA, Ziegler K, Zipp K, Heuer S, et al. Whole blood-derived miRNA profiles as potential new tools for ovarian cancer screening. Br J Cancer 2010;103:693–700.
[29] Roth P, Wischhusen J, Happold C, Chandran PA, Hofer S, Eisele G, et al. A specific miRNA signature in the peripheral blood of glioblastoma patients. J Neurochem 2011;118:449–57.
[30] Meder B, Keller A, Vogel B, Haas J, Sedaghat-Hamedani F, Kayvanpour E, et al. MicroRNA signatures in total peripheral blood as novel biomarkers for acute myocardial infarction. Basic Res Cardiol 2011;106:13–23.
[31] Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W, et al. An analysis of human microRNA and disease associations. PLoS ONE 2008;3:e3420.
[32] Vorwerk S, Ganter K, Cheng Y, Hoheisel J, Stahler PF, Beier M. Microfluidic-based enzymatic on-chip labeling of miRNAs. Nat Biotechnol 2008;25:142–9.
[33] Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19:185–93.
[34] R Development Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2008.
[35] Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD. Cytoscape Web: an interactive web-based network browser. Bioinformatics 2010;26:2347–8.

Multiple Sclerosis Journal
http://msj.sagepub.com/

Comprehensive analysis of microRNA profiles in multiple sclerosis including next-generation sequencing

Andreas Keller, Petra Leidinger, Florian Steinmeyer, Cord Stähler, Andre Franke, Georg Hemmrich-Stanisak, Andreas Kappel, Ian Wright, Jan Dörr, Friedemann Paul, Ricarda Diem, Beatrice Tocariu-Krick, Benjamin Meder, Christina Backes, Eckart Meese and Klemens Ruprecht

Mult Scler published online 8 July 2013
DOI: 10.1177/1352458513496343

The online version of this article can be found at:
http://msj.sagepub.com/content/early/2013/07/04/1352458513496343

Published by:
SAGE
http://www.sagepublications.com

Research Paper

MULTIPLE SCLEROSIS JOURNAL | MSJ

Multiple Sclerosis Journal
0(0) 1–9

DOI: 10.1177/1352458513496343
msj.sagepub.com
SAGE

Comprehensive analysis of microRNA profiles in multiple sclerosis including next-generation sequencing

Andreas Keller[1,2], Petra Leidinger[1], Florian Steinmeyer[3,4], Cord Stähler[2], Andre Franke[5], Georg Hemmrich-Stanisak[5], Andreas Kappel[6], Ian Wright[6], Jan Dörr[4,7], Friedemann Paul[4,7], Ricarda Diem[8], Beatrice Tocariu-Krick[9], Benjamin Meder[10], Christina Backes[1], Eckart Meese[1] and Klemens Ruprecht[3,4]

Abstract

Background: MicroRNAs (miRNAs) are short, noncoding RNAs with gene regulatory functions whose expression profiles may serve as disease biomarkers.

Objective: The objective of this study was to perform a comprehensive analysis of miRNA expression profiles in blood of patients with a clinically isolated syndrome (CIS) or relapsing–remitting multiple sclerosis (RRMS) including next-generation sequencing (NGS).

Methods: miRNA expression was analyzed in whole blood samples from treatment-naïve patients with CIS (n = 25) or RRMS (n = 25) and 50 healthy controls by NGS, microarray analysis, and quantitative real-time polymerase chain reaction (qRT-PCR).

Results: In patients with CIS/RRMS, NGS and microarray analysis identified 38 and eight significantly deregulated miRNAs, respectively. Three of these miRNAs were found to be significantly up- (hsa-miR-16-2-3p) or downregulated (hsa-miR-20a-5p, hsa-miR-7-1-3p) by both methods. Another five of the miRNAs significantly deregulated in the NGS screen showed the same direction of regulation in the microarray analysis. qRT-PCR confirmed the direction of regulation for all eight and was significant for three miRNAs.

Conclusions: This study identifies a set of miRNAs deregulated in CIS/RRMS and reconfirms the previously reported underexpression of hsa-miR-20a-5p in MS. hsa-miR-20a-5p and the other validated miRNAs may represent promising candidates for future evaluation as biomarkers for MS and could be of relevance in the pathophysiology of this disease.

Keywords

Multiple sclerosis, clinically isolated syndrome, microRNAs, biomarker, next-generation sequencing, microarray, real-time polymerase chain reaction

Date received: 9 December 2012; revised: 5 June 2013; accepted: 10 June 2013

Introduction

According to current diagnostic criteria, diagnosis of multiple sclerosis (MS) relies on a combination of clinical, radiological, and cerebrospinal fluid findings.[1-3] While establishing a diagnosis of MS is usually straightforward in

[1]Department of Human Genetics, Saarland University, Germany.
[2]Siemens Healthcare, Germany.
[3]Department of Neurology, Charité – Universitätsmedizin Berlin, Germany.
[4]Clinical and Experimental Multiple Sclerosis Research Center, Charité – Universitätsmedizin Berlin, Germany.
[5]Institute of Clinical Molecular Biology, Christian-Albrechts Universität Kiel, Germany.
[6]Siemens Healthcare Diagnostics, USA.
[7]NeuroCure Clinical Research Center, Charité – Universitätsmedizin Berlin, Germany.
[8]Department of Neurooncology, Heidelberg University, Germany.
[9]Department of Neurology, Saarland University, Germany.
[10]Department of Internal Medicine II, Heidelberg University, Germany.

Corresponding author:
Klemens Ruprecht, Klinik für Neurologie, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany.
Email: klemens.ruprecht@charite.de
A.K., P.L. and F.S. contributed equally as first authors. C.B., E.M. and K.R. contributed equally as senior authors.

patients with typical clinical and paraclinical presentations, it can be challenging in patients with atypical features.[4] Furthermore, differentiation of MS from alternative diagnoses, such as other inflammatory central nervous system (CNS) diseases, can be difficult, especially in patients with a clinically isolated syndrome (CIS). Identification of biomarkers, defined as parameters that can be objectively measured and evaluated as indicators of pathogenic processes,[5] therefore appears desirable to further facilitate the diagnosis of MS. In addition, biomarkers could aid in monitoring disease activity and in the evaluation of treatment responses.

MicroRNAs (miRNAs) are short (about 20–24 nucleotides in length), single-stranded regulatory RNAs that modulate gene expression at the posttranscriptional level by repressing translation or degradation of specific messenger RNA (mRNA) targets. About 1500 miRNAs have been described in humans so far, and more than one-third of all human genes may be controlled by miRNAs.[6,7] miRNAs thus represent an important gene regulatory mechanism, increasingly recognized to be involved in physiologic and pathologic processes both in the CNS and the immune system.[8,9] Of note, miRNAs are present in a stable form in human blood,[10] and previous studies performed by others and ourselves suggest that miRNA expression profiles determined in serum or whole blood samples hold promise as diagnostic biomarkers in various human diseases, including cancer and autoimmunity.[11,12] Others, and our group, have consequently investigated miRNA profiles in whole blood, peripheral blood mononuclear cells, purified leukocyte subsets, or plasma of patients with MS in comparison to healthy controls.[13-25] While all those studies identified some differences in the expression levels of certain miRNAs, they were limited by either the number of miRNAs studied, the number of patients included, or possible confounding effects of concomitant immunomodulatory therapy.[26] Moreover, while former studies were based on microarray technology or quantitative real-time polymerase chain reaction (qRT-PCR), next-generation sequencing (NGS) has meanwhile emerged as a novel, powerful, and unbiased methodological approach to miRNA expression profiling.[27,28]

Here, we performed a comprehensive analysis of miRNA expression patterns in whole blood samples from 50 treatment-naïve patients with a CIS or relapsing–remitting MS (RRMS) as well as 50 matched healthy controls using NGS, microarray analysis of 1205 human miRNAs, and qRT-PCR. Our analysis identified several miRNAs deregulated in patients with CIS/RRMS, which may represent promising candidates for future evaluation as biomarkers for MS and could provide insights into the pathophysiology of this disease.

Patients and methods

Sample collection

From November 2009 to February 2011 about 2.5 ml of blood was collected in PAXgene Blood RNA tubes (Becton Dickinson, Heidelberg, Germany) from 50 patients (36 female, 14 male) followed at the Department of Neurology and NeuroCure Clinical Research Center, Charité – Universitätsmedizin Berlin, with a diagnosis of a CIS (n = 25) or RRMS (n = 25) according to the McDonald 2005 criteria.[2] Fifty age- (± 4 years) and gender-matched healthy adults were included as controls. Patients were categorized into those with stable disease (no relapse within a period of at least two months before blood withdrawal, n = 31) and patients with active disease (relapse at the time of or within two months before blood withdrawal, n = 19). Data on the use of oral contraceptives were available from 19 of the 36 female patients included in the study. Seven of these 19 women took oral contraceptives. None of the patients took any long-term immunomodulatory or immunosuppressive therapy at the time of or prior to inclusion into the study. Patients had not been treated with glucocorticosteroids for at least two months before blood withdrawal. Pregnancy or intercurrent diseases at the time of blood withdrawal were exclusion criteria. The study was approved by the institutional review board of Charité – Universitätsmedizin Berlin (EA1/131/09) and all participants provided written informed consent. Coded samples were stored at –20°C and shipped on dry ice to the Department of Human Genetics, Saarland University, for further blinded processing.

RNA isolation

Total RNA including miRNA was isolated using the PAXgene Blood miRNA Kit (Qiagen) following the manufacturer's recommendations. Isolated RNA was stored at –80°C. RNA integrity was analyzed using Bioanalyzer 2100 (Agilent) and concentration and purity were measured using NanoDrop 2000 (Thermo Scientific). A total of four samples (three controls and one patient with RRMS) failed the quality criteria and were excluded from the study.

NGS

The total RNA concentration required for NGS was ≥ 1 µg per sample. A total of 37 of the 100 samples collected in our study met this requirement and were included in the NGS analysis. Isolated RNA was shipped on dry ice to the Institute of Clinical Molecular Biology (IKMB), Christian-Albrechts Universität, Kiel, Germany, where NGS was performed. Individual samples were tagged with molecular barcodes and then sequenced together in multiplexed pools. The TruSeq Small RNA sample preparation Kit (Illumina) was used to generate multiplexed sequencing libraries, which were afterwards sequenced on a HiSeq2000 System (Illumina) using the 50 bp fragment sequencing protocol. Resulting sequencing reads were demultiplexed using the CASAVA 1.8 software package (Illumina) and quality checked using FastQC tools (Babraham Institute). A primary mapping analysis using the miRDeep2-pipeline[29] was conducted to ensure that a significant proportion of miRNAs

Table 1. miRNAs deregulated in NGS and microarray analysis and validated using qRT-PCR. For the NGS analyses the average read counts are given, for the microarray analyses the mean signal intensity values are given, and for the qRT-PCR the mean ΔCT values are given. Bold font indicates upregulation of the respective miRNA in CIS/RRMS, normal font indicates downregulation in CIS/RRMS compared to controls. Note that higher ΔCT values indicate lower expression.

miRNA	NGS			Microarray			qRT-PCR		
	Control	CIS/RRMS	p value	Control	CIS/RRMS	p value	Control	CIS/RRMS	p value
hsa-miR-22-5p	**6**	**9.681**	*0.004*	**1034.25**	**1075.84**	*0.594*	**7.17**	**7.12**	*0.88*
hsa-miR-125b-5p	**4.806**	**14.444**	*0.018*	**22.979**	**42.036**	*0.156*	**5.02**	**3.66**	*0.0006*
hsa-miR-629-5p	**4.847**	**8.958**	*0.024*	**8.385**	**9.988**	*0.224*	**7.45**	**7.21**	*0.46*
hsa-miR-16-2-3p	**418.792**	**793.625**	*0.05*	**26.77**	**50.538**	*0.001*	**7.62**	**6.55**	*0.014*
hsa-miR-100-5p	**3.993**	**44.25**	*0.04*	**12.678**	**26.817**	*0.355*	**6.06**	**4.46**	*0.005*
hsa-miR-20a-5p	7.194	6.847	*0.049*	255.718	96.802	*0.018*	1.93	2.62	*0.2*
hsa-miR-151a-3p	580.403	455.056	*0.009*	51.266	43.992	*0.61*	2.23	2.33	*0.7*
hsa-miR-7-1-3p	2.681	0.563	*0.001*	3.106	0.1	*0.02*	6.02	6.46	*0.16*

miRNA: microRNA; NGS: next-generation sequencing; qRT-PCR: quantitative real-time polymerase chain reaction; CIS: clinically isolated syndrome; RRMS: relapsing–remitting multiple sclerosis.

were sequenced. In total, 37 samples from 16 patients (five RRMS, 11 CIS) and 21 controls were analyzed in two multiplexed pools. On average, 1.5 million–2 million high-quality sequencing reads per sample were obtained (at a total of 92.28 million reads), of which up to 70% contained miRNA information. The raw illumina reads were first preprocessed by cutting the 3' adapter sequence. This was performed by the program fastx_clipper from the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Reads shorter than 18 nucleotides after clipping were removed. The remaining reads were collapsed, i.e. after this step we had only unique reads and their frequency per sample. For the remaining steps, we used the miRDeep2 pipeline. These steps consist of mapping the reads against the genome (hg19), mapping the reads against miRNA precursor sequences from miRBase release v18 (http://www.mirbase.org/), and summarizing the counts for the samples.

Microarray measurement

Microarray analysis was performed as previously described using *SurePrint* 8x60K Human v16 *miRNA* microarrays (Agilent, CatNo G4870A) that contain 40 replicates of each of the 1205 miRNAs of miRBase v16 (http://www.mirbase.org/).[30] Except for the four samples that failed the quality control criteria, all remaining 96 samples were included in the microarray study. All samples were analyzed as individual samples and not pooled.

qRT-PCR

We composed a set of 40 age- and gender-matched patient and control samples that were also used for microarray and NGS analyses. Samples included in the qRT-PCR study were analyzed as individual and not as pooled samples. The group of patients included 10 CIS and 10 RRMS patients. qRT-PCR was performed at the Comprehensive Biomarker Center GmbH, Heidelberg, Germany, using the Taqman qRT-PCR system (Applied Biosystems). The small RNAs RNU6B and RNU48 were used as endogenous controls. However, as RNU6B yielded very high Ct values, we used only RNU48 for normalization with the deltaCT method.[31] The mean ± standard deviation Ct value of RNU48 of the 40 samples analyzed was 25.45 ± 0.82.

Bioinformatic analysis

The same analyses were performed for NGS as well as microarray results. Following quantile normalization, we computed for each miRNA the area under the receiver operator characteristic curve (AUC), the fold-change, and the significance value (p value) using t tests. Because of the exploratory nature of this study, no adjustments for multiple testing were made. P values < 0.05 were considered statistically significant. Based on this analysis, we computed a Venn diagram for the significant NGS and microarray results. Concordant candidate miRNAs were validated using qRT-PCR and further analyzed. For each concordant miRNA we extracted relevant disease interactions from the human microRNA disease database (HMDD, http://202.38.126.151/hmdd/mirna/md/).

Results

Demographics of patients with CIS/RRMS and healthy controls studied in this work are summarized in Supplemental Table 1. We applied three experimental approaches to comprehensively analyze miRNA profiles in patients with CIS/RRMS (Figure 1). Using NGS, we first carried out a screening in a cohort of 16 cases and 21 controls. Secondly, we performed a microarray analysis on an enlarged cohort encompassing 49 cases and 47 age- and

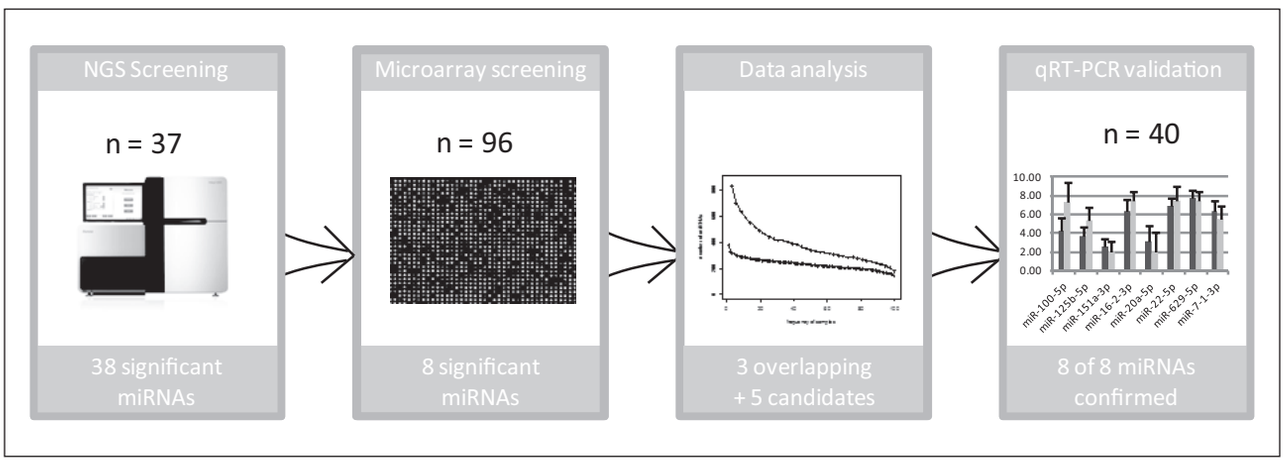


Figure 1. Project overview. Three experimental methods including NGS, microarray, and qRT-PCR were applied to comprehensively analyse miRNA expression profiles in patients with CIS/RRMS and healthy controls.
NGS: next-generation sequencing; qRT-PCR: quantitative real-time polymerase chain reaction; CIS: clinically isolated syndrome; RRMS: relapsing–remitting multiple sclerosis.

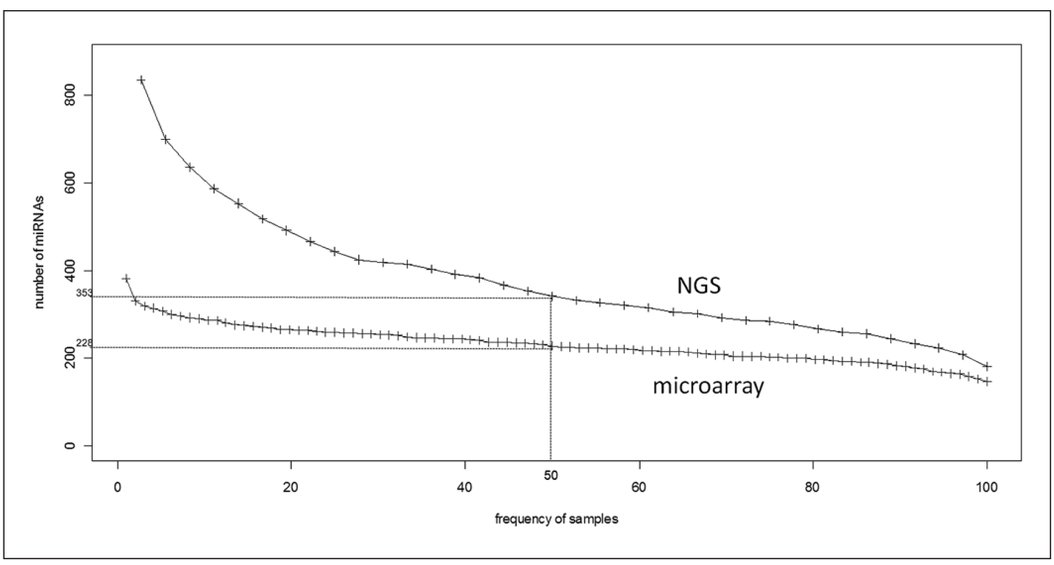


Figure 2. Number of miRNAs and frequency of samples in which these miRNAs were detected. The upper curve indicates the results of NGS, the lower curve indicates the results of microarray screening. By NGS 353 miRNAs were detectable in at least half of all investigated samples, by microarray analysis 228 miRNAs were detectable in at least half of all investigated samples (see dashed lines).
miRNA: microRNA; NGS: next-generation sequencing.

gender-matched controls. Both high-throughput analyses yielded eight miRNA candidates that were, thirdly, analyzed by qRT-PCR in 20 cases and 20 controls.

NGS screening

We initially performed a high-throughput screening in blood samples from 16 patients with CIS/RRMS and 21 controls. Altogether, we found a total of 835 miRNAs being expressed in at least one of the samples. Figure 2 shows the number of miRNAs and the frequency of samples in which these miRNAs were detected; 353 miRNAs were detectable in at least half of the investigated samples. Following normalization, *t* tests demonstrated that expression of a total of 38 miRNAs significantly differed between patients and controls. Out of the 38 deregulated miRNAs 16 were downregulated and 22 were upregulated in CIS/RRMS. The eight strongest deregulated miRNAs are shown in Figure 3. These eight miRNAs included five downregulated miRNAs, namely hsa-miR-361-5p, hsa-miR-7-1-3p, hsa-miR-548o-3p, hsa-miR-151a-3p, and hsa-miR-548am-3p and three upregulated miRNAs, namely hsa-miR-22-5p, hsa-miR-27a-5p, and hsa-miR-4677-3p.

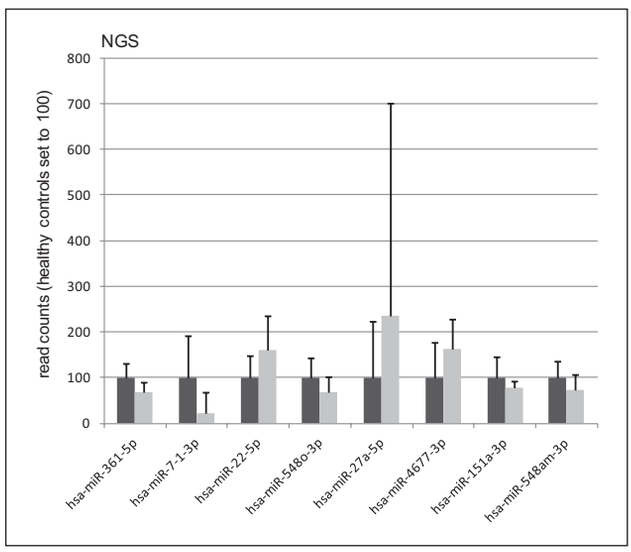


Figure 3. The eight most deregulated miRNAs identified by NGS. The median read counts of the control samples (dark gray) and the median read counts of the MS samples (light gray) of the eight most deregulated miRNAs identified by NGS are indicated together with the standard deviation. Expression in healthy controls is set to 100.
miRNA: microRNA; NGS: next-generation sequencing; MS: multiple sclerosis.

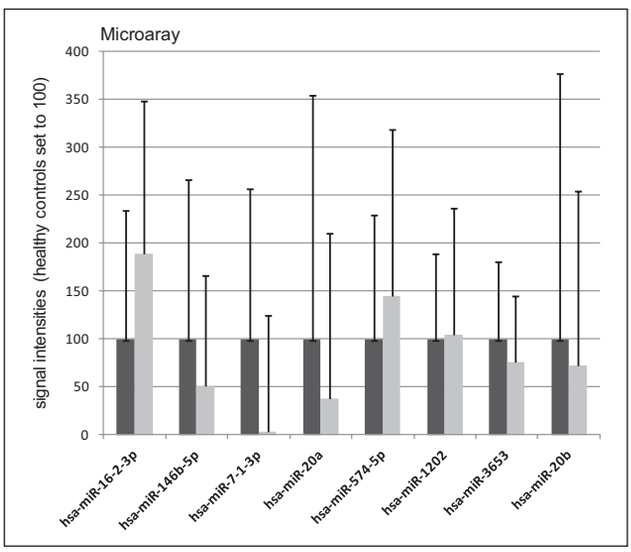


Figure 4. The eight most deregulated miRNAs identified by microarray analysis. The median signal intensities of the control samples (dark gray) and the median signal intensities of the MS samples (light gray) of the eight most deregulated miRNAs identified by microarray analysis are indicated together with the standard deviation. Values of healthy controls are set to 100.
miRNA: microRNA; MS: multiple sclerosis.

Microarray screening

We next screened an enlarged cohort of 49 patients with CIS/RRMS and 47 matched healthy controls by microarrays. We detected significantly fewer ($p = 6.5 \times 10^{-7}$) miRNAs in the microarray compared to the NGS study (Figure 2). In detail, microarray analysis detected 382 miRNAs that were expressed in at least one sample. These were only 46% of the 835 miRNAs that were detected by NGS in at least one of the samples. Furthermore, microarray analysis detected only 228 miRNAs that were expressed in at least 50% of all samples. In the microarray experiments we detected a total of eight significantly deregulated miRNAs (Figure 4). Out of the eight deregulated miRNAs, five miRNAs were downregulated (hsa-miR-146b-5p, hsa-miR-7-1-3p, hsa-miR-20a-5p, hsa-miR-3653, hsa-miR-20b) and three were upregulated (hsa-miR-16-2-3p, hsa-miR-574-5p, hsa-miR-1202) in patients with CIS/RRMS.

Overlap in significantly deregulated miRNAs between NGS and microarray analyses and correlation with clinical parameters

As described above, we detected 38 significantly deregulated miRNAs by NGS and eight significantly deregulated miRNAs by microarray analysis. These numbers correspond to 1.9% of all known miRNAs for the NGS experiment and 0.7% of the miRNAs on the biochip for microarray analysis, respectively. This makes a random overlap between the two data sets unlikely. However, three miRNAs, namely hsa-miR-16-2-3p, hsa-miR-20a-5p, and hsa-miR-7-1-3p, were identified by both NGS and microarray analysis (Figure 5). We performed one million permutation tests to confirm that this overlap is highly significant ($p = 0.004$). In addition, five of the 38 miRNAs identified by NGS (miRNAs hsa-miR-22-5p, hsa-miR-125b-5p, hsa-miR-629-5p, hsa-miR-100-5p, and hsa-miR-151a-3p) showed the same direction of regulation in the microarray analysis, i.e. each of these miRNAs was either up- or downregulated in both approaches, although the deregulation of these five miRNAs in the microarray experiments was not statistically significant. Table 1 summarizes the expression and significance values of the eight miRNAs identified as deregulated by both methods. We also compared the expression levels, as measured by microarray, of those eight miRNAs with the clinical disease activity (active vs stable disease) and diagnosis (CIS vs RRMS) of the patients included in this work. When assessed by unpaired *t* tests, none of the comparisons revealed significant differences, suggesting that within our patient group the analyzed miRNAs were not influenced by clinical disease activity or a diagnosis of CIS vs RRMS. Finally, as an estimate of the individual ability of each of the eight differentially expressed miRNAs to discriminate patients with CIS/RRMS and controls, we also calculated receiver operating characteristic (ROC) curves for each of these miRNAs using microarray and NGS data (Supplemental Figure 1).

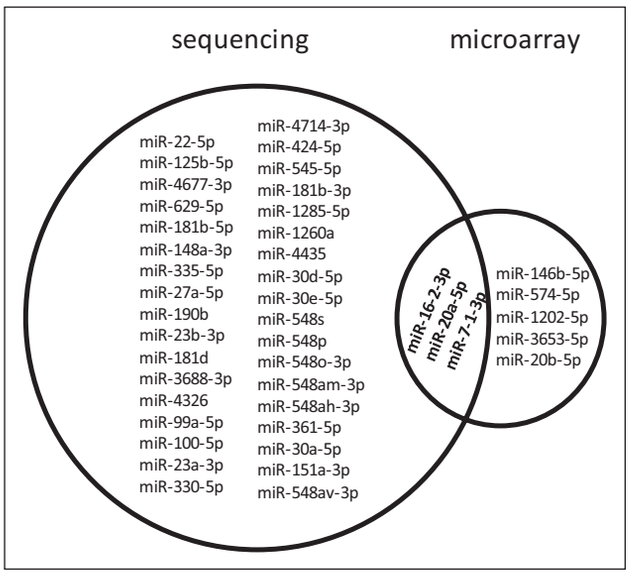


Figure 5. Venn diagram showing the significantly deregulated miRNAs identified by NGS and microarray screening. NGS identified 38 miRNAs and microarray analysis eight miRNAs significantly deregulated in CIS/RRMS. Three miRNAs were identified by both approaches.
miRNA: microRNA; NGS: next-generation sequencing; CIS: clinically isolated syndrome; RRMS: relapsing–remitting multiple sclerosis.

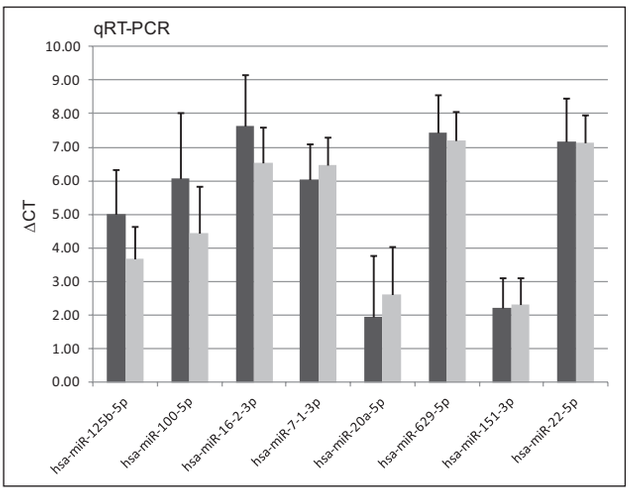


Figure 6. qRT-PCR validation of the eight miRNAs. The bar diagram shows the mean ΔCT values and standard deviations for the eight tested candidate MS markers. Note that higher ΔCT values indicate lower expression. Controls: dark gray bars; MS: light gray bars. qRT-PCR: quantitative real-time polymerase chain reaction; miRNA: microRNA; MS: multiple sclerosis.

qRT-PCR validation

The eight miRNAs listed in Table 1 were further analyzed using qRT-PCR in a set of 20 patients with CIS/RRMS and 20 healthy controls. All eight miRNAs showed the same direction of regulation in the qRT-PCR analysis as in the NGS or microarray experiments. Three miRNAs, including hsa-miR-125b-5p, hsa-miR-16-2-3p, and hsa-miR-100-5p, were significantly deregulated according to the qRT-PCR results (Table 1). Figure 6 shows the mean ΔCT values and standard deviations for the qRT-PCR validation. Figure 7 summarizes the comparison of the expression analysis of the eight miRNAs using NGS, microarray, and qRT-PCR.

Disease specificity of the identified miRNAs

We extracted the known disease associations for all human miRNAs deposited in the HMDD and calculated the number of miRNAs in relation to the number of disease associations (Figure 8). We then focused on the disease associations of the eight concordant miRNAs identified by our NGS and microarray analyses. On average, each human miRNA deposited in the HMDD is associated with eight diseases. Computing the number of disease interactions for each of the eight miRNAs identified in this work, we found that all but one (hsa-miR-16-2-3p) of the eight miRNAs have previously been associated with more than eight diseases, indicating that they have a higher than average number of disease interactions.

Discussion

The present study is the first to apply NGS as a novel methodological approach to miRNA profiling in patients with MS. Using NGS and subsequent verification by microarray analyses, we identified a set of eight miRNAs, including five miRNAs that were found to be upregulated and three miRNAs that were found to be downregulated by both methods in patients with CIS/RRMS as compared to controls. qRT-PCR experiments corroborated regulation of all of these miRNAs.

One advantage of NGS is that it permits the unbiased detection of theoretically all miRNAs in a given sample, regardless of whether they have previously been described.[27] Besides not being restricted to the annotated human miRNome, the sensitivity of NGS is also higher than that of microarray technologies. Thus, out of the 38 differentially expressed miRNAs identified by NGS, seven miRNAs (18.4%) were not included on the *SurePrint* 8x60K Human v16 *miRNA* microarray, which is restricted to the content of miRBase v16, and 14 miRNAs (36.8%) were included but not detected by the array approach. In line with the higher sensitivity of NGS, the maximum number of miRNAs detected in a single blood sample was more than two times higher (835 vs 382) in the NGS as compared to the microarray screen. Nevertheless, the overlap of three significantly deregulated miRNAs identified by NGS and microarray technology indicates that converging results can be obtained by these two approaches, in keeping with recent data from a study of lung cancer patients.[28]

Concerning possible functions of the identified miRNAs in MS, a potential role in the regulation of immune response

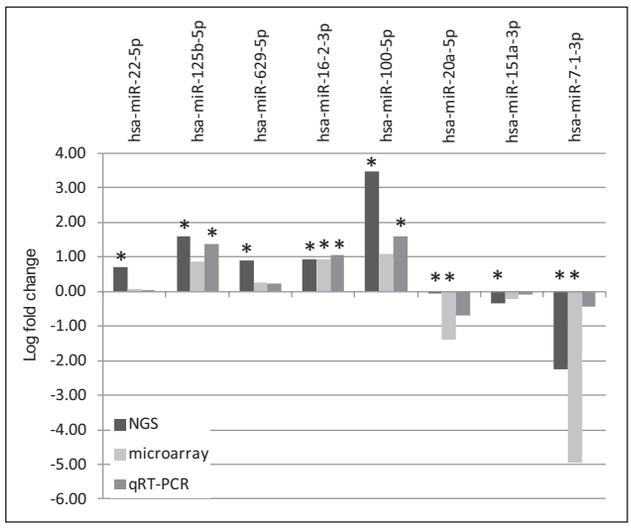


Figure 7. Comparison of the expression analysis of the eight miRNAs using NGS, microarray, and qRT-PCR. The height of the bars represents the logarithmized fold changes of each miRNA and each used analysis method (NGS: dark gray bars, microarray: light gray bars, qRT-PCR: middle gray bars). Significant differences as compared to controls ($p < 0.05$) are indicated by asterisks.
miRNA: microRNA; NGS: next-generation sequencing; qRT-PCR: quantitative real-time polymerase chain reaction.

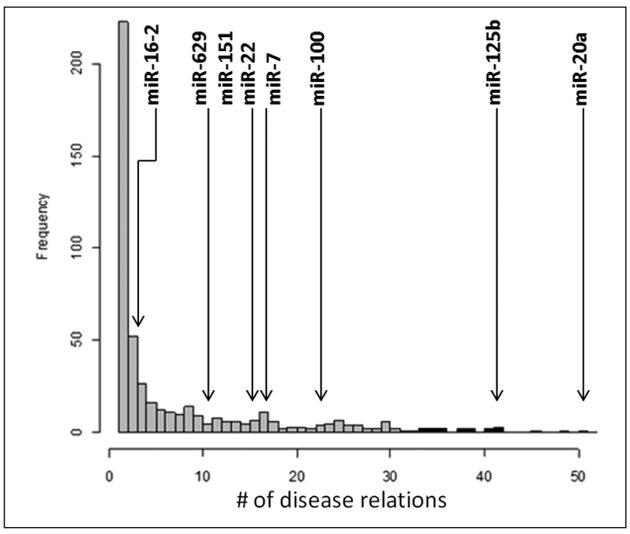


Figure 8. Frequency of disease associations. The figure shows the frequency of relations to diseases for all miRNAs according to the HMDD. The eight miRNAs identified by NGS and microarray analysis as deregulated in CIS/RRMS are indicated. Black bars represent the top 5% of all miRNAs with the most disease associations.
miRNA: microRNA; HMDD: human microRNA disease database; NGS: next-generation sequencing; CIS: clinically isolated syndrome; RRMS: relapsing–remitting multiple sclerosis.

pathways has been previously described for three of the eight miRNAs, namely hsa-miR-20a-5p, hsa-miR-100-5p, and hsa-miR-125b-5p.[19,32,33] Nevertheless, whether and how the identified miRNAs may play a pathogenically relevant role in MS await further clarification. A database search of human miRNA disease interactions showed that all but one of the eight identified miRNA were previously associated with at least 10 different human diseases. Although this suggests that each single miRNA is not highly specific for CIS/RRMS, future analyses should explore whether combinations of certain miRNAs may display an increased specificity.

Factors influencing miRNA expression profiles in blood under physiological conditions have not been studied in detail. As a possible limitation of this study, we cannot exclude that, for instance, hormonal changes during the menstrual cycle or use of oral contraceptives might influence miRNA expression levels in blood. However, since patients and controls were very well matched for gender and age, and assuming that a similar percentage of female patients and controls took oral contraceptives, we consider it unlikely that hormonal changes during the menstrual cycle or use of oral contraceptives might have severely biased our results.

The overall number of miRNAs identified as significantly deregulated tended to be lower in the present as compared to previous miRNA expression studies in MS.[13-25] This is most likely explained by the more stringent experimental strategy applied in the present work, consisting of an initial screen with two independent methods (NGS and microarray) and further concentration on those miRNAs that were found to be deregulated by both methods. Interestingly, miRNAs identified herein partially overlapped with miRNAs formerly shown to be deregulated in MS. hsa-miR-22-5p, one of the five miRNA found to be upregulated in the present screen, has previously been reported as upregulated in plasma,[13] active brain lesions,[34] and $CD4^+CD25^+$ regulatory T cells[18] of patients with MS. Furthermore, in accordance with our present findings, hsa-miR-20a-5p was found to be underexpressed in patients with MS by microarray analysis (Illumina Sentrix Array Matrix) and qRT-PCR.[19] Comparing the present results with our own initial study,[17] we also detected significant overlaps. In detail, we previously identified hsa-miR-629-5p ($p = 0.0009$) and hsa-miR-100-5p ($p = 0.04$) as significantly upregulated in MS, while we found hsa-miR-20a-5p to be downregulated ($p = 0.0009$). Likewise, hsa-miR-125b-5p was upregulated in our former work, although barely missing the significance threshold ($p = 0.06$). Importantly, together with the present study, hsa-miR-20a-5p has now been shown to be downregulated in whole blood of patients with MS in three independent cohorts of patients with MS and controls by various methodological approaches (different microarray platforms, qRT-PCR, NGS). Facing the rapidly growing number of miRNAs being associated with MS, reproduction of results in independent cohorts appears essential for identification of meaningful candidates, and hsa-miR-20a-5p may be

one of those. Indeed, in a recent in silico analysis of miRNA-mRNA interaction networks in MS hsa-miR-20a-5p emerged as one of the central hubs, regulating about 500 genes, as identified by miRNA-mRNA predictions algorithms.[35] Furthermore, many of the 19 currently known experimentally verified genes being targeted by hsa-miR-20a-5p are involved in the regulation of T cells.[35] For instance, the hsa-miR-20a-5p target gene CDKN1A (coding for cyclin kinase inhibitor p21) plays a role in T cell activation and has been associated with systemic autoimmunity.[36]

Altogether, we herein show that application of NGS to miRNA profiling in MS is feasible and can identify novel as well as previously described miRNAs that are deregulated in patients with MS as compared to healthy controls. The identified miRNAs may be regarded as a set of interesting candidates for future evaluation as biomarkers for MS. Further experimental analyses of functional aspects of those miRNAs may help to improve our understanding of the pathophysiology of this multifactorial disease.

Conflict of interest statement

AK, CS, AK, and IW are employees of Siemens Healthcare.
FP has received research support, travel grants, and speaking fees from Bayer HealthCare, Teva, Sanofi-Aventis, Biogen Idec, Novartis, and Merck Serono as well as research support from the Arthur Arnstein Foundation, Berlin, Germany, and travel reimbursement from the Guthy Jackson Charitable Foundation. FP is supported by the German Research Foundation (DFG Exc 257) and is a member of the steering committee of the OCTIMS study.
KR has received research support from Novartis as well as speaking fees and travel grants from Bayer HealthCare, Biogen Idec, Merck Serono, Sanofi Aventis, Teva, and Novartis.

Funding

This work was supported by Deutsche Forschungsgemeinschaft [DFG Exc 257]; Hedwig-Stalter-Stiftung; University Research Fund, Charité – Universitätsmedizin Berlin; and Siemens Healthcare.

References

1. Polman CH, Reingold SC, Banwell B, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Ann Neurol* 2011; 69: 292–302.
2. Polman CH, Reingold SC, Edan G, et al. Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria". *Ann Neurol* 2005; 58: 840–846.
3. McDonald WI, Compston A, Edan G, et al. Recommended diagnostic criteria for multiple sclerosis: Guidelines from the International Panel on the diagnosis of multiple sclerosis. *Ann Neurol* 2001; 50: 121–127.
4. Miller DH, Weinshenker BG, Filippi M, et al. Differential diagnosis of suspected multiple sclerosis: A consensus approach. *Mult Scler* 2008; 14: 1157–1174.
5. Bielekova B and Martin R. Development of biomarkers in multiple sclerosis. *Brain* 2004; 127: 1463–1478.
6. Lewis BP, Burge CB and Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 2005; 120: 15–20.
7. Kozomara A and Griffiths-Jones S. miRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* 2011; 39: D152–D157.
8. Xiao C and Rajewsky K. MicroRNA control in the immune system: Basic principles. *Cell* 2009; 136: 26–36.
9. Kosik KS. The neuronal microRNA system. *Nat Rev Neurosci* 2006; 7: 911–920.
10. Chen X, Ba Y, Ma L, et al. Characterization of microRNAs in serum: A novel class of biomarkers for diagnosis of cancer and other diseases. *Cell Res* 2008; 18: 997–1006.
11. Keller A, Leidinger P, Bauer A, et al. Toward the blood-borne miRNome of human diseases. *Nat Methods* 2011; 8: 841–843.
12. Mitchell PS, Parkin RK, Kroh EM, et al. Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci U S A* 2008; 105: 10513–10518.
13. Siegel SR, Mackenzie J, Chaplin G, et al. Circulating microRNAs involved in multiple sclerosis. *Mol Biol Rep* 2012; 39: 6219–6225.
14. Du C, Liu C, Kang J, et al. MicroRNA miR-326 regulates TH-17 differentiation and is associated with the pathogenesis of multiple sclerosis. *Nat Immunol* 2009; 10: 1252–1259.
15. Lindberg RL, Hoffmann F, Mehling M, et al. Altered expression of miR-17–5p in CD4+ lymphocytes of relapsing–remitting multiple sclerosis patients. *Eur J Immunol* 2010; 40: 888–898.
16. Otaegui D, Baranzini SE, Armañanzas R, et al. Differential micro RNA expression in PBMC from multiple sclerosis patients. *PLoS One* 2009; 4: e6309.
17. Keller A, Leidinger P, Lange J, et al. Multiple sclerosis: MicroRNA expression profiles accurately differentiate patients with relapsing–remitting disease from healthy controls. *PLoS One* 2009; 4: e7440.
18. De Santis G, Ferracin M, Biondani A, et al. Altered miRNA expression in T regulatory cells in course of multiple sclerosis. *J Neuroimmunol* 2010; 226: 165–171.
19. Cox MB, Cairns MJ, Gandhi KS, et al. MicroRNAs miR-17 and miR-20a inhibit T cell activation genes and are underexpressed in MS whole blood. *PLoS One* 2010; 5: e12132.
20. Guerau-de-Arellano M, Smith KM, Godlewski J, et al. Micro-RNA dysregulation in multiple sclerosis favours pro-inflammatory T-cell-mediated autoimmunity. *Brain* 2011; 134: 3578–3589.
21. Fenoglio C, Cantoni C, De Riz M, et al. Expression and genetic analysis of miRNAs involved in CD4+ cell activation in patients with multiple sclerosis. *Neurosci Lett* 2011; 504: 9–12.
22. Lorenzi JC, Brum DG, Zanette DL, et al. miR-15a and 16-1 are downregulated in CD4(+) T cells of multiple sclerosis relapsing patients. *Int J Neurosci* 2012; 122: 466–471.
23. Paraboschi EM, Solda G, Gemmati D, et al. Genetic association and altered gene expression of mir-155 in multiple sclerosis patients. *Int J Mol Sci* 2011; 12: 8695–8712.
24. Martinelli-Boneschi F, Fenoglio C, Brambilla P, et al. MicroRNA and mRNA expression profile screening in multiple sclerosis patients to unravel novel pathogenic steps and identify potential biomarkers. *Neurosci Lett* 2012; 508: 4–8.

25. Waschbisch A, Atiya M, Linker RA, et al. Glatiramer acetate treatment normalizes deregulated microRNA expression in relapsing remitting multiple sclerosis. *PLoS One* 2011; 6: e24604.
26. Junker A, Hohlfeld R and Meinl E. The emerging role of microRNAs in multiple sclerosis. *Nat Rev Neurol* 2010; 7: 56–59.
27. Guerau-de-Arellano M, Alder H, Ozer HG, et al. miRNA profiling for biomarker discovery in multiple sclerosis: From microarray to deep sequencing. *J Neuroimmunol* 2012; 248: 32–39.
28. Keller A, Backes C, Leidinger P, et al. Next-generation sequencing identifies novel microRNAs in peripheral blood of lung cancer patients. *Mol Biosyst* 2011; 7: 3187–3199.
29. Friedlander MR, Mackowiak SD, Li N, et al. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* 2012; 40: 37–52.
30. Leidinger P, Keller A, Backes C, et al. MicroRNA expression changes after lung cancer resection: A follow-up study. *RNA Biol* 2012; 9: 900–910.
31. Livak KJ and Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods* 2001; 25: 402–408.
32. Leonhardt F, Grundmann S, Behe M, et al. Inflammatory neovascularization during graft-versus-host disease is regulated by alphav integrin and miR-100. *Blood* 2013; 121: 3307–3318.
33. Sun YM, Lin KY and Chen YQ. Diverse functions of miR-125 family in different cell contexts. *J Hematol Oncol* 2013; 6: 6.
34. Junker A, Krumbholz M, Eisele S, et al. MicroRNA profiling of multiple sclerosis lesions identifies modulators of the regulatory protein CD47. *Brain* 2009; 132: 3342–3352.
35. Angerstein C, Hecker M, Paap BK, et al. Integration of microRNA databases to study microRNAs associated with multiple sclerosis. *Mol Neurobiol* 2012; 45: 520–535.
36. Santiago-Raber ML, Lawson BR, Dummer W, et al. Role of cyclin kinase inhibitor p21 in systemic autoimmunity. *J Immunol* 2001; 167: 4067–4074.

Leidinger et al. Genome Biology 2013, 14:R78
http://genomebiology.com/2013/14/7/R78

Genome Biology

RESEARCH Open Access

A blood based 12-miRNA signature of Alzheimer disease patients

Petra Leidinger[1†], Christina Backes[1†], Stephanie Deutscher[1], Katja Schmitt[1], Sabine C Mueller[1], Karen Frese[2], Jan Haas[2], Klemens Ruprecht[3], Friedemann Paul[3,4], Cord Stähler[5], Christoph JG Lang[6], Benjamin Meder[2], Tamas Bartfai[7], Eckart Meese[1†] and Andreas Keller[1,5*†]

Abstract

Background: Alzheimer disease (AD) is the most common form of dementia but the identification of reliable, early and non-invasive biomarkers remains a major challenge. We present a novel miRNA-based signature for detecting AD from blood samples.

Results: We apply next-generation sequencing to miRNAs from blood samples of 48 AD patients and 22 unaffected controls, yielding a total of 140 unique mature miRNAs with significantly changed expression levels. Of these, 82 have higher and 58 have lower abundance in AD patient samples. We selected a panel of 12 miRNAs for an RT-qPCR analysis on a larger cohort of 202 samples, comprising not only AD patients and healthy controls but also patients with other CNS illnesses. These included mild cognitive impairment, which is assumed to represent a transitional period before the development of AD, as well as multiple sclerosis, Parkinson disease, major depression, bipolar disorder and schizophrenia. miRNA target enrichment analysis of the selected 12 miRNAs indicates an involvement of miRNAs in nervous system development, neuron projection, neuron projection development and neuron projection morphogenesis. Using this 12-miRNA signature, we differentiate between AD and controls with an accuracy of 93%, a specificity of 95% and a sensitivity of 92%. The differentiation of AD from other neurological diseases is possible with accuracies between 74% and 78%. The differentiation of the other CNS disorders from controls yields even higher accuracies.

Conclusions: The data indicate that deregulated miRNAs in blood might be used as biomarkers in the diagnosis of AD or other neurological diseases.

Keywords: Alzheimer disease, miRNA, biomarker, next-generation sequencing, quantitative Real Time PCR

Background

Alzheimer disease(AD) is the most common form of neurodegenerative illness leading to dementia which is predicted to affect as much as 1 in 85 people globally by 2050 [1]. While early-onset (familiar) AD has been reported in younger people, the majority of (sporadic) AD cases is diagnosed in people aged over 65 years [2]. As of today, final diagnosis of AD can only be achieved by autopsy making the identification of reliable, early, and non-invasive biomarkers a major challenge. Finding such non-invasive, reliable diagnostic tools is of paramount importance as it appears that early intervention in the prodromal stage of AD or the identification and therapy of those patients with mild cognitive impairment who will transform to AD rapidly might be a possibility to delay the onset of AD substantially [3].

A prominent example of recently developed AD biomarker assays is the combinatorial analysis of the concentration of peptides and proteins: beta-amyloid-1-42 (Aß 42), tau, and/or p-tau in the cerebrospinal fluid (CSF). According to the S3 guidelines, an increased level of tau protein together with a decreased level of beta-amyloid-1-42 provides strong evidence for the presence of AD [4]. The combinatorial analysis of all three factors yields even higher diagnostic accuracy than the combination of only two of the above-mentioned proteins [5].

* Correspondence: keller.andreas@siemens.com
† Contributed equally
[1]Department of Human Genetics, Saarland University, Kirrbergerstraße, Building 60, 66421 Homburg, Germany
Full list of author information is available at the end of the article

Furthermore, combinatorial analysis of Aß levels and tau levels can discriminate between patients with stable mild cognitive impairment (MCI) and patients with progressive MCI into AD or other types of dementia with a sufficient diagnostic accuracy [6]. Nevertheless, according to the S3 guidelines, the analysis of CSF biomarker is only indicated to confirm the diagnosis if other clinical symptoms give evidence for the presence of neurodegenerative dementia or for the differential diagnostics of other forms of diseases that can cause symptoms like dementia (encephalitis, neuroborreliosis, multiple sclerosis, Lues, brain abscess, metastases).

The use of peripheral markers, like Aß and tau in easily accessible peripheral cells (in particular platelets and skin fibroblasts), as a diagnostic tool has been under investigation for more than 10 years [7,8]. Molecular genetics analyses of common single nucleotide polymorphisms (SNPs) in genes such as presenilin or ApoE4 did not significantly improve risk estimation for the susceptibility of AD [9]. Likewise, there is no consistent evidence for an association between AD and genetic variation of mitochondrial DNA (mtDNA) [10].

There is increasing effort to develop molecular diagnostic markers that meet requirements like easy accessibility, for example, from blood, sufficiently high specificity and sensitivity, low costs and applicability by laboratories with standard equipment. Several blood, plasma, or serum born AD biomarkers have been proposed to meet these criteria. Doecke et al. recently presented a panel of protein biomarkers to reliably detect AD with an accuracy of 85% [11]. Moreover, Tan et al. provided evidence that the proteins p53 and p21 can be used to detect AD using blood samples. A receiver operating characteristic curve analysis revealed a specificity of 76% and a sensitivity of 84% for p53, 88% and 82% for p53(ser15), 80% and 75% for p21, and 84% and 68% for p21(thr145) [12].

Besides proteins microRNAs (miRNAs) have also demonstrated their potential as non-invasive biomarkers from blood and serum for a wide variety of human pathologies [13]. A deregulation of miRNA expression might be involved in neurological dysfunction or neurodegenerative processes. Interestingly, Liang et al. [14] showed that the expression pattern of brain and blood PBMC cluster together which might be an indication that a specific blood based expression signature might prove to be useful as biomarker for AD and other neurological diseases. MiRNA expression analyses can be readily applied for *in vitro* diagnostic testing by molecular diagnostics and CLIA (Clinical Laboratory Improvement Amendments) laboratories.

While altered miRNA patterns have been exhaustively investigated in AD patients' tissue samples or cell cultures [15-18], less information on circulating miRNAs in AD is known. A recent serum profiling of AD patients provided first evidence that expression changes of circulating miRNAs may be valuable biomarkers for AD [19].

We describe our results obtained by applying the next-generation sequencing (NGS) approach to screen the expression of all human miRNAs in blood from extensively characterized AD patients and healthy controls. Patient blood was obtained from the SAMPLE (Serial Alzheimer diseaseand MCI Prospective Longitudinal Evaluation) Registry of PrecisionMed (San Diego, CA, USA) and blood from age-matched healthy donors from the ACE (Aging Cognition Evaluation) Registry, a PrecisionMed- UBC (The University of British Columbia) collaboration. We identified 140 unique differentially expressed miRNAs between AD patients and controls. Validation of a 12-miRNA signature was carried out by RT-qPCR in a cohort of 202 samples encompassing patients suffering from other neurological disorders including mild cognitive impairment as a potential preliminary stage of AD, and other neurodegenerative diseases like Parkinson disease and multiple sclerosis as well as mental diseases like schizophrenia (SCHIZ), major depression (DEP), and bipolar disorder (BD).

A combination of AD-specific miRNA expression signatures with the rapidly developing and expanding amyloid load imaging techniques may be useful as non-invasive diagnostic tools in AD diagnosis in the future [20].

Results

Initial biomarker screening using next-generation sequencing

To detect potential AD biomarkers we examined blood from well-characterized patients and controls. We obtained blood from the SAMPLE (Serial Alzheimer diseaseand MCI Prospective Longitudinal Evaluation) Registry of PrecisionMed (San Diego, CA, USA). SAMPLE is a sample depository resulting from a longitudinal study that evaluates cognition in women and men, who are recruited, evaluated, cognitively studied, and sampled from 12 to 15 experienced investigative sites in USA. All participants underwent several tests (that is, Alzheimer Disease Assessment Scale-cognitive subscale (ADAS-Cog), Clinical Dementia Rating (CDR), Wechsler Memory Scale, and Mini-Mental State Exam (MMSE)) to evaluate cognition. Blood from age-matched healthy donors was obtained from the Ace Registry, which is a biological sample bank of serial patient samples with linked serial cognition data, based on a cognition battery selected from UBC's proprietary computerized testing platform.

We carried out high-throughput NGS of 22 healthy control samples (C) and 48 AD patient samples using IlluminaHiSeq 2000 sequencing with eight multiplexed samples on each sequencing lane. We detected not only known human miRNAs, but also novel miRNA candidates that have previously not been included in the miRBase

v18 [21,22]. These miRNA candidates are, however, much less abundant compared to the known human miRNAs. After removing the least abundant miRNAs (that is, all miRNAs with <50 read counts summed up across all samples of each group) we detected a total of 383 different miRNA precursors resulting in 416 unique mature miRNA forms.

To compare the NGS results of the AD patient samples with the samples from healthy donors we first computed Wilcoxon-Mann-Whitney (WMW) test and adjusted the significance values for multiple testing using Benjamini-Hochberg adjustment. All miRNAs with adjusted significance values <0.05 were considered statistically significant. We also computed the area under the receiver operator characteristics curve (AUC). In total, we detected 180 significantly dys-regulated miRNAs (140 unique mature miRNAs) including 90 miRNAs (58 unique mature miRNAs) that were downregulated and 90 miRNAs (82 unique mature miRNAs) that were upregulated in AD samples compared to healthy control samples (see Additional file 1-Table S1). Additional file 2-Figure S1 shows a heatmap for 180 significantly dys-regulated miRNAs. The most upregulated miRNA was hsa-miR-30d-5p (AUC of 0.0819) with a *P* value of $8.35*10^{-6}$and the most downregulated miRNA was hsa-miR-144-5p (AUC of 0.9138) with *P* value of $8.35*10^{-6}$. While the high AUC value indicates that each of these miRNAs has sufficient power to differentiate between AD and healthy controls, they are not specific for AD since both miRNAs have already been described for many other human pathologies, including different neoplasms [13]. Among the significantly dys-regulated miRNAs are also 15 novel miRNA candidates (called brain-miR) that were all upregulated in AD compared to controls. A list of all novel mature miRNAs is provided in Additional file 3-Table S2. To gain first insight into the biological function of the mature miRNAs that were dys-regulated between AD patients and healthy control individuals, we applied a miRNA over-representation analysis for these miRNAs using the TAM (tool for annotations of human miRNAs) database [23,24]. The TAM database classifies over- or under-represented miRNAs according to the categories miRNA family, miRNA cluster, miRNA function, miRNA associated diseases, and tissue specificity. We detected for all dys-regulated miRNAs 56 significant categories (*P* value <0.05 after adjustment for multiple testing), with the interesting categories miR-30 family with five miRNAs being upregulated (*P* value $6.64*10^{-4}$), the let-7 family with nine downregulated miRNAs (*P* value $5.65*10^{-7}$), and the disease category Alzheimer disease for which six dys-regulated miRNAs were relevant, including hsa-miR-21, hsa-miR-17, hsa-miR-29a, hsa-miR-29b, hsa-miR-106b, and hsa-miR-107 (*P* value 0.0139).

To determine whether the 140 unique differentially expressed miRNAs between AD patients and healthy controls cluster together within a same genomic region, which would suggest presence of common regulatory mechanisms for their expression, we sorted all miRNAs according to their position on each chromosome. Then, we assigned the miRNAs to one of the following three classes: not dys-regulated; upregulated in AD; and downregulated in AD. Finally, we searched for regions that contain at least three different dys-regulated mature miRNAs by applying window sizes varying between 1,000 and 100,000 base pairs. Within regions encompassing <1,000 base pairs we detected two clusters including one on chromosome 19 with the upregulated miRNAs hsa-miR-99b-5p and hsa-miR-125a-5p and the downregulated miRNA hsa-let-7e-5p and a second cluster on chromosome 22 with the downregulated miRNAs hsa-let-7a-5p and hsa-let-7b-5p and the upregulated miRNA hsa-let-7b-3p. Analyzing regions of up to 5,000 base pairs, we found on chromosome 9 a dense cluster with a total of five dys-regulated miRNAs including the downregulated miRNAs hsa-let-7a-5p, hsa-let-7f-5p, and hsa-let-7d-5p and the upregulated miRNAs hsa-let-7f-1-3p and hsa-let-7d-3p. For regions up to 30,000 base pairs, we discovered one region on chromosome 6 with three co-located miRNAs including hsa-miR-30c-5p, hsa-miR-30a-3p, and hsa-miR-30a-5p, all of which were upregulated. To understand whether the miRNAs are regulated by specific transcription factors (TF), we extracted potential TF binding sites from the UCSC genome browser but found no evidence for a significant enrichment for specific TF binding sites.

In the next step, we performed classification of AD and control samples using a standard machine learning approach. In a cross-validation loop, we stepwise added the miRNAs with lowest significance values and repeatedly carried out radial basis function support vector machines (SVM). As shown in Figure 1, a signature of 250 miRNAs yields an accuracy, specificity, and sensitivity of 90%. Since this set of miRNAs contains a significant amount of redundant miRNAs with largely identical information and high correlation among many miRNAs, a significantly smaller set of miRNAs is likely to yield comparably accurate distinction between AD samples and samples from healthy controls. We selected 12 miRNAs with limited cross-correlation, including strongly dys-regulated miRNAs that show a potential to separate AD from controls. We furthermore compared our NGS results with previous studies on different types of cancer and non-cancer diseases [13] in order to ensure that the selected miRNAs are not dys-regulated in several other diseases. Besides known miRNAs we also included two unknown miRNAs, namely brain-miR-112 and brain-miR-161. Finally, the selected 12-miRNA signature

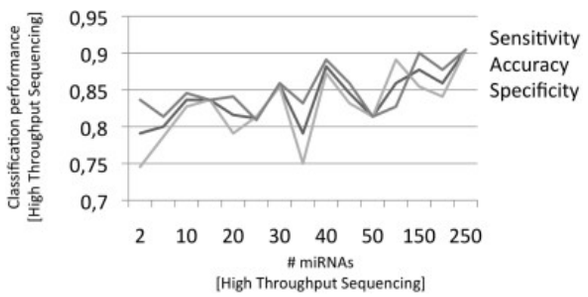


Figure 1 Classification performance dependent on miRNA combinations. With increasing number of miRNAs the accuracy, specificity, and sensitivity increases towards convergence at 90%.

contains the miRNAs brain-miR-112, brain-miR-161, hsa-let-7d-3p, hsa-miR-5010-3p, hsa-miR-26a-5p, hsa-miR-1285-5p, and hsa-miR-151a-3p, all of which are upregulated in AD and the downregulated miRNAs hsa-miR-103a-3p, hsa-miR-107, hsa-miR-532-5p, hsa-miR-26b-5p, and hsa-let-7f-5p.

Validation of a 12-miRNA signature by RT-qPCR

To validate the 12-miRNA signature we employed RT-qPCR and included not only additional patients with AD, but also patients with other diseases including neurological disorders. In total, we analyzed 12 miRNAs in 202 samples as detailed in Table 1.

We first considered the miRNA fold quotients that were obtained for AD samples and controls. We compared the fold quotients of each of the 12 miRNAs between initial NGS screening cohort and the RT-qPCR validation cohort. All but two of the 12 miRNAs, namely hsa-miR-1285-5p and hsa-miR-26a-5p, have been dysregulated in the same direction in both approaches, indicating a high degree of concordance between screening and validation study. Both hsa-miR-1285-5p and hsa-miR-26a-5p have been significantly upregulated in AD in the NGS screening experiment while downregulated in the RT-qPCR validation (see Figure 2). This discrepancy might be due to the duplication of the AD sample cohort. However, SVM classification on the RT-qPCR data to separate AD and controls using linear kernels in 10-fold cross-validations with 100 repetitions reached on average an accuracy of 93.3%, a specificity of 95.1%, and a sensitivity of 91.5%. The computed means, standard deviations, and confidence intervals for the repetitions concerning specificity, sensitivity, and accuracy are presented in Table 2, as well as the results for the control classifications with the randomly permuted class labels.

To evaluate whether the selected miRNAs are stage-dependent we further grouped the AD patients according to their MMSE score into mild AD (MMSE >19, n = 39) and moderate AD (MMSE 12-19, n = 46). The MMSE is a short test of 30 questions used to screen for cognitive impairment. Each question to be answered is scored with points, with a maximum possible score of 30 points. This questionnaire can be used to estimate the severity of cognitive impairment and to follow the course of cognitive changes in an individual over time. Normally, patients reaching 27 to 30 points do not suffer from dementia, 10 to 26 points are indicative for mild-to-moderate dementia, and less than 9 points indicates severe dementia. We found no significant expression differences of the 12-miRNA signature between the mild AD group and the moderate AD group.

As patients with other neurological disorders can show similar symptoms as AD patients, we decided to validate our AD NGS results also with samples from patients with several neurological diseases. Specifically, we asked whether other neurological disorders show significant deviations in the expression of the 12 miRNAs. The results of this validation help to determine whether the investigated miRNAs have the potential for clinical applications. We analyzed patients with neurodegenerative diseases (MCI, Parkinson disease (PD), multiple sclerosis (clinically isolated syndrome, CIS)) and patients with psychiatric disorders (SCHIZ, BD, and DEP) for the signature of 12 miRNAs. The pattern, which was closest to

Table 1 Overview of the blood samples analyzed using NGS and RT-qPCR

Sample group	*N*	Age (mean ± SD)	Sex (female/male)	MMSE (mean ± SD)	Cohort size NGS	Cohort size RT-qPCR
AD	106	72.7(10.4)	53/53	18.9 (3.4)	48	94
Healthy control	22	67.1 (7.5)	11/11	29.3 (1.2)	22	21
Mild cognitive impairment	18	73.9 (6.2)	9/9	25.3 (1.4)	-	18
Multiple sclerosis	16	32.3 (10.7)	12/4	NA	-	16
PD	9	69.7 (9.0)	1/8	25.2 (4.2)	-	9
DEP	15	45.2 (9.1)	0/15	NA	-	15
BD	15	41.9 (13.7)	14/1	29.5 (1.6)	-	15
SCHIZ	14	41.7 (7.9)	1/13	26.1 (4.3)	-	14

AD: Alzheimer disease; BD: bipolar disorder; DEP: major depression; MMSE: Mini-Mental State Exam; NA: not available; NGS: next-generation sequencing; PD: Parkinson's disease; RT-qPCR: quantitative real-time PCR; SCHIZ: schizophrenia.

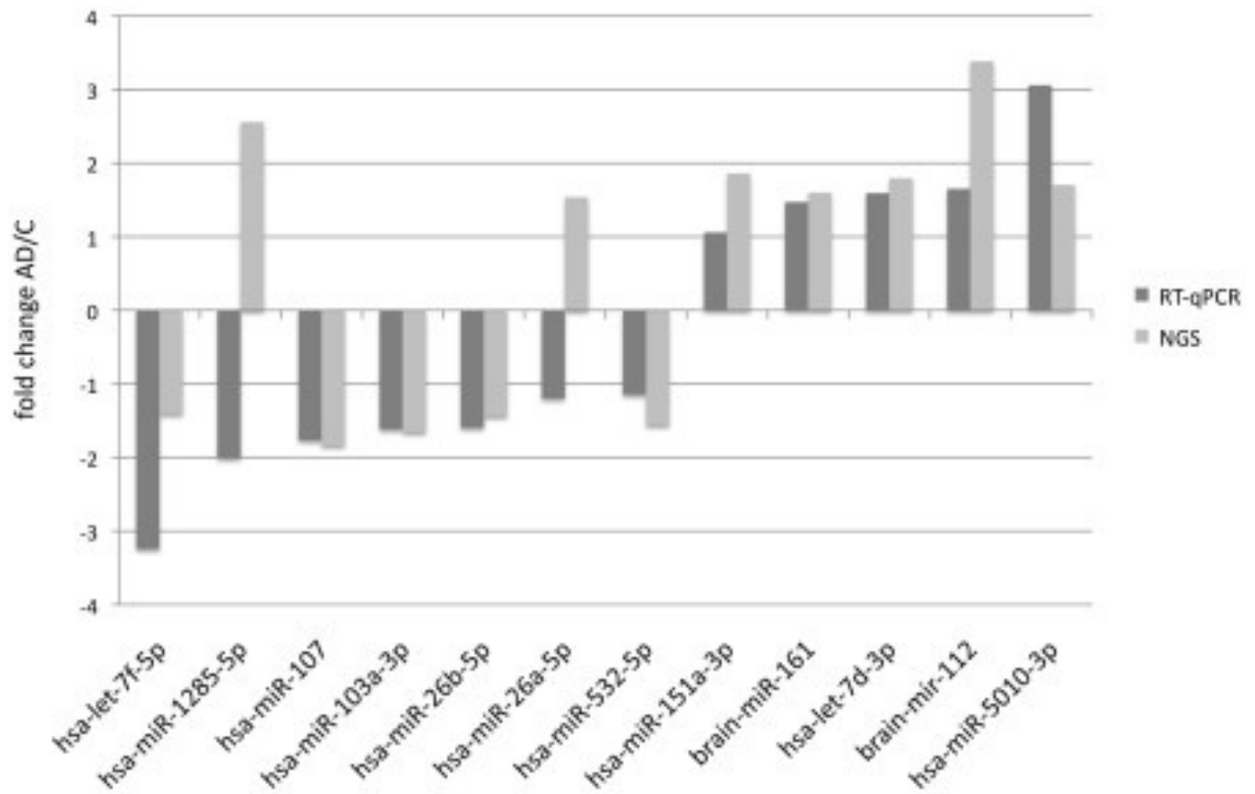


Figure 2 Comparison of the expression analysis results of AD patients *versus* healthy controls (fold changes) obtained by NGS and RT-qPCR for the 12-miRNA signature.

Table 2 Summary of the SVM classifications containing the means, standard deviations, and 95% confidence intervals (CI) of the accuracy (acc), specificity (spec), sensitivity (sens) running 100 repetitions of 10-fold cross-validations with linear kernel.

Comparison	Classification			Permutation test		
	Acc	Spec	Sens	Acc	Spec	Sens
AD *vs*. control	93.3% ± 4.6 CI:92.4-94.2%	95.1% ± 5.4 CI:94.1-96.2%	91.5% ± 5.8 CI:90.4-92.7%	50.7% ± 12.5 CI:48.2-53.1%	50.7% ± 13.3 CI:48.1-53.3%	50.7% ± 14.1 CI:47.9-53.4%
MCI *vs*. control	84.2% ± 3.7 CI:83.4-84.9%	81.1% ± 5.6 CI:80.0-82.2%	87.7% ± 3.7 CI:87.0-88.5%	51.3% ± 11.4 CI:49.0-53.5%	52.0% ± 12.2 CI:50.0-54.4%	50.4% ± 13.5 CI:47.8-53.1%
PSY *vs*. control	97.1% ± 1.6 CI:96.8-97.4%	95.3% ± 1.7 CI:95.0-95.6%	99.0% ± 2.4 CI:98.5-99.4%	48.7% ± 10.6 CI:46.7-50.8%	48.5% ± 12.4 CI:46.0-50.9%	49.0% ± 12.1 CI:46.6-50.8%
Other ND *vs*. control	82.8% ± 5.0 CI:81.8-83.7%	84.0% ± 5.8 CI:83.0-85.2%	81.4% ± 6.7 CI:80.1-82.7%	50.3% ± 10.3 CI:48.3-52.3%	50.7% ± 11.7 CI:48.4-53.0%	50.0% ± 12.0 CI:47.6-52.3%
NEURO *vs*. control	86.1% ± 5.7 CI:85.0-87.2%	88.7% ± 6.8 CI:87.3-90.0%	83.6% ± 6.6 CI:82.3-84.9%	49.9% ± 10.8 CI:47.8-52.1%	50.1% ± 11.5 CI:47.9-52.3%	49.8% ± 13.3 CI:47.2-52.3%
AD *vs*. MCI	75.6% ± 7.8 CI:74.1-77.2%	76.7% ± 8.3 CI:75.1-78.4%	74.6% ± 9.7 CI:72.7-76.5%	50.6% ± 9.4 CI:48.7-52.4%	51.2% ± 10.4 CI:49.1-53.2%	49.9% ± 11.7 CI:47.7-52.2%
AD *vs*. PSY	77.8% ± 4.0 CI:77.0-78.5%	76.3% ± 4.8 CI:75.4-77.3%	79.2% ± 5.4 CI:78.1-80.2%	50.0% ± 8.0 CI:48.5-51.6%	49.1% ± 9.3 CI:47.3-50.9%	51.1% ± 10.3 CI:49.1-53.1
AD *vs*. other ND	73.8% ± 4.4 CI:72.9-74.7%	75.2% ± 4.7 CI:74.2-76.1%	72.4% ± 6.4 CI:71.2-73.7%	50.1% ± 7.3 CI:48.7-51.5%	49.2% ± 9.4 CI:47.4-51.1%	51.0% ± 8.5 CI:49.3-52.7%

The right part of the table contains the results for the permuted class labels. PSY = psychological disorders (DEP, BD, SCHIZ), other ND = other neurodegenerative disorders (PD, MS, MCI), NEURO = PSY + other ND

AD was SCHIZ, where we found six up- and six downregulated miRNAs. We found a strong overall downregulation for most of the selected 12 miRNAs for patients with DEP and PD and a strong overall upregulation for patients with MCI, CIS, and BD (Figure 3).

In addition, we also applied machine learning procedures using SVM to estimate the accuracy, sensitivity, and specificity of the 12-miRNA signature regarding the other neurological diseases in comparison to the control group and to AD. The results of these classifications are also listed in Table 2. Interestingly, while the 12 miRNAs were chosen for their potential to separate AD and controls, this signature also separates the group of the psychological disorders (DEP, BD, SCHIZ) from controls with an accuracy of 97.1%, a specificity of 95.3%, and a sensitivity of 99.0% whereas other neurodegenerative diseases (PD, multiple sclerosis, mild cognitive impairment) were separated from controls with a worse accuracy of 82.8%, a specificity of 84.0%, and a sensitivity of 81.4%. The average accuracy for the other classifications against controls (that is, MCI *versus* control and neurodegenerative and psychological disorders *versus* control) reached values of 84.2% and 86.1%, respectively. Furthermore, we tested how well the 12-miRNA signature separates AD from MCI, AD from psychological disorders, and AD from other neurodegenerative diseases, respectively. The average accuracy for these comparisons was between 73.8% and 77.8%. Since the 12-miRNA signature has been tailored to differentiate between AD and controls, other miRNAs may likely contribute to a signature that permits also a better differentiation between the other tested diseases and AD.

Prediction of miRNA targets and over-representation analysis

Target gene prediction of the 10 known miRNAs from the 12-miRNA signature revealed 2,354 genes that may be regulated by those miRNAs. These target genes were used to perform an over-representation analysis and identified 73 computed Gene Ontology (GO) categories with *P* values <0.05 and FDR adjustment. Interestingly, we found a significant enrichment of miRNA targets in the GO categories nervous system development, neuron projection, neuron projection development, and neuron projection morphogenesis. These GO categories are listed in Table 3 together with the predicted miRNA target genes involved in these categories. Furthermore, target genes that have already been related to AD or other neurological diseases are also listed in the table in separate columns.

Target gene prediction for the two unknown target genes brain-miR-112 and brain-miR-161 revealed 234 target genes for brain-miR-112, but only six target genes

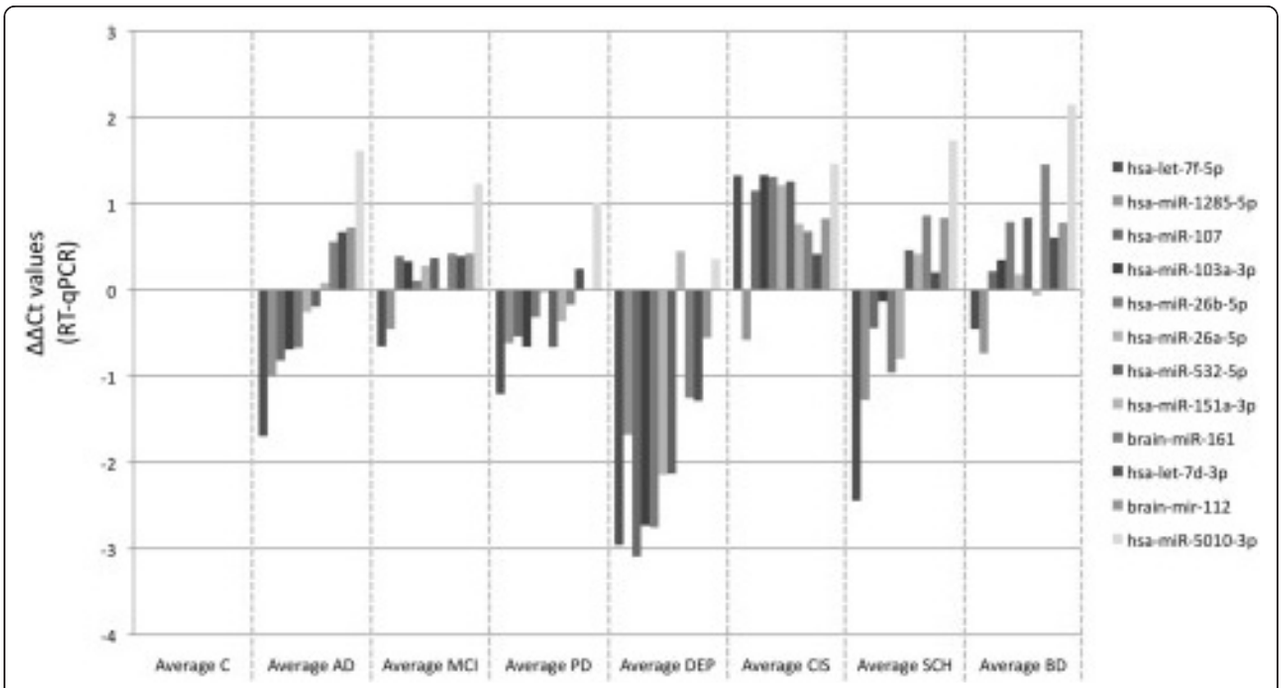


Figure 3 Results of the RT-qPCR analysis of the 12 miRNAs in each of the analyzed diseases. ΔΔCT values of the patients groups as compared to controls are shown on the y-axis. A lower expression in the patients than in controls is indicated by the bars <0 and higher expression in the patients as compared to controls is indicated by the bars >0. (C = healthy control, AD = Alzheimer disease, MCI = mild cognitive impairment, PD = Parkinson disease, DEP = major depression, CIS = multiple sclerosis (clinically isolated syndrome), SCH = schizophrenia, BD = bipolar disorder).

Table 3 Results of the over-representation analysis of the predicted target genes of the 10 known miRNAs.

Subcategory	Subcategory alternative name	Expected	Observed	*P*value (FDR)	Target genes	AD	BD	DEP	PD	SCHIZ	Multiple sclerosis
Nervous system development	GO:0007399	159,555	215	0,000921692	ARSB ATM GJA1 JAG1 LEP NDP PTEN PAFAH1B1 TWIST1 DRD1 IGF1R GDF6 SMPD1 KCNMA1 NTRK2 CTNS NF1 INSC SLC6A3 FBXO45 IGF1 ADM APC DLG4 GRIN2A PAX7 PPT1 GPSM1 FEZF1 TSC1 DISC1 GLRB BMPR1B CDK6 CX3CR1 CELSR2 ID4 ERBB3 FGF2 AFF2 GLRA2 **GSK3B** HOXB3 LAMC1 LRP6 LSAMP NGF NPAS2 OPHN1 P2RY1 PEX13 POU3F1 PTPRZ1 SALL1 SMARCC1 STRN T TFAP2A TGFB2 TIAM1 NR2C2 YWHAH ZIC1 ULK1 ENC1 IRS2 ADAM23 KALRN SEMA5A EDNRB DMD AQP4 GMFB SDHA SLC1A2 GDA VCAN DVL1 EPHA4 EPHA7 KIF5C LRP2 POU4F2 RPS6KA3 SPOCK1 TGFBR1 AXIN2 DCLK1 MED1 ONECUT2 SIM1 CNTN2 ATF1 DLX6 ERBB4 SMAD4 SIX3 NHLH1 POU3F2 REST ABI2 PURA SMAD1 NAB1 SIX1 PPARD PRKCQ CHERP MAB21L2 TBR1 CHL1 FRS2 FKTN BTG2 SHOX2 SLC5A3 ZNF24 WWP1 STMN2 RAPGEF5 PIP5K1C ATXN10 RACGAP1 GREM1 NRG1 CNTNAP2 RPS6KA6 CYFIP1 ULK2 NLGN1 RUFY3 ARHGAP26 NFASC CLASP2 NIPBL SUFU PDGFC HPCAL4 RAPGEFL1 SHC3 FZD3 SIX4 BAIAP2 CSGALNACT1 PCDHB10 NMUR2 VANGL2 SEMA6A CNTN3 LRRC4C RET GNAO1 SCN2A FGF12 XRCC5 NTN4 BCR ADAM22 ACSL4 FGFR1 HTR5A NOTCH2 TTLL7 PGAP1 JHDM1D ATXN3 ZEB1 NDEL1 MAP2 B3GNT5 CHD6 SLITRK6 ELAVL3 HOOK3 ATOH8 WNT3A ZIC5 FGF1 SOX6 PDE5A SNAP25 GRIN3A CREB1 NRXN1 NRXN3 TPM3 FYN SEMA6D HOXA1 BDNF ALDH5A1 UNC5B DMBX1 IL6ST UHMK1 DCX CUX1 ATL1 GLDN RNF6 FAM5C CCNG1 NRP2 GAS7 ACSL3 RCAN1 SYNJ1 PCDH9 MOG RTN1 QKI LIG4 MBNL1 CCDC64 WNK1	DRD1 IGF1R GSK3B FGF1 FYN BDNF	DRD1 DISC1 GSK3B BCR HTR5A BDNF SYNJ1	BCR SNAP25 CREB1 BDNF	BDNF	LEP DRD1 SLC6A3 GRIN2A DISC1 YWHAH SLC1A2 CHL1 NRG1 FZD3 HTR5A SNAP25 BDNF SYNJ1	JAG1
Neuron projection	GO:0043005	49,9516	79	0,00411039	ADRB2 CA2 PAFAH1B1 ATP1A2 DRD1 GABRA6 GAD1 GRM3 IGF1R KCNJ2 NPY1R PGR AR KCNMA1 NF1 TACR1 MYO5B ACTN2 GRM1 APC GRIN2A ATXN1L MYO5A PPT1 OPRM1 TSC1 HTR2A CALCR OPHN1 STRN TGFB2 ULK1 PRSS12 KALRN BNIP3 SLC1A2 DVL1 EPHA7 KIF5C NCAM2 KIF5A CNTN2 ABI2 PURA CAPRIN1 IGF2BP1 SCN1A STMN2 SNCA STAT1 EPB41L3 ATXN10 CNTNAP2 RUFY3 NFASC ERC2 KIAA1598 SEMA6A SCN2A GAN TTLL7 CPEB1 NDEL1 MAP2 PSD2 CALD1 SNAP25 GRIN3A TPM3 AQP11 UHMK1 EXOC8 DICER1 ATL1 ANKS1B RNF6 CCNG1 CACNA1C NRP2	DRD1 IGF1R HTR2A SNCA	DRD1 HTR2A	HTR2A	SNCA	DRD1 GAD1 GRM3 GRIN2A HTR2A SLC1A2 SNAP25	ADRB2

Table 3 Results of the over-representation analysis of the predicted target genes of the 10 known miRNAs. *(Continued)*

Neuron projection development	GO:0031175	43,1466	67	0,0162232	GJA1 PTEN PAFAH1B1 IGF1R ADM APC FEZF1 DISC1 BMPR1B CELSR2 ERBB3 LAMC1 NGF OPHN1 PTPRZ1 STRN TIAM1 YWHAH ULK1 KALRN DMD VCAN DVL1 EPHA4 EPHA7 KIF5C POU4F2 DCLK1 CNTN2 ATF1 POU3F2 ABI2 SMAD1 TBR1 STMN2 PIP5K1C ATXN10 CYFIP1 ULK2 RUFY3 NFASC FZD3 BAIAP2 SEMA6A LRRC4C GNAO1 ACSL4 FGFR1 NDEL1 MAP2 SLITRK6 WNT3A SNAP25 GRIN3A CREB1 NRXN1 NRXN3 HOXA1 BDNF UNC5B UHMK1 DCX ATL1 RNF6 NRP2 GAS7 CCDC64	IGF1R BDNF	DISC1 BDNF	SNAP25 CREB1 BDNF	BDNF	DISC1 YWHAH FZD3 SNAP25 BDNF	–
Neuron projection morphogenesis	GO:0048812	33,5906	52	0,0462928	GJA1 PAFAH1B1 IGF1R ADM APC FEZF1 BMPR1B CELSR2 ERBB3 NGF OPHN1 PTPRZ1 TIAM1 YWHAH ULK1 KALRN DMD VCAN DVL1 EPHA4 EPHA7 KIF5C POU4F2 DCLK1 CNTN2 POU3F2 SMAD1 TBR1 PIP5K1C CYFIP1 ULK2 RUFY3 NFASC FZD3 BAIAP2 SEMA6A LRRC4C NDEL1 SLITRK6 WNT3A SNAP25 CREB1 NRXN1 NRXN3 HOXA1 BDNF UNC5B DCX ATL1 RNF6 NRP2 GAS7	BDNF IGF1R	BDNF	SNAP25 CREB1 BDNF	BDNF	FZD3 SNAP25 BDNF YWHAH	–

This table lists interesting Gene Ontology (GO) subcategories and over-represented target genes with *P*values <0.05 and FDR adjustment related to nervous system development. Target genes associated with AD and other neurological diseases are listed separately.

AD: Alzheimer disease; BP: bipolar disorder; DEP: major depression; PD, Parkinson's disease; SCHIZ: schizophrenia

for brain-miR-161. Over-representation analysis was done for both brain-miRNAs separately. Here, we identified 126 GO categories with *P* value <0.05 for brain-miR-112, with significant enrichment of miRNA targets in GO categories associated with nervous system and neuron function (see Table 4). For brain-miR-161 no significant GO categories were found.

Discussion

At present, there is no single molecular test that is suitable to reliably diagnose AD with adequate specificity and sensitivity. Tests for the analysis of CSF proteins like Aß42 or tau have high specificity and sensitivity, but are only indicated as confirmation of AD diagnosis based on clinical symptoms or as differential diagnosis to differentiate between AD and other forms of diseases that can cause symptoms like dementia. The analysis of SNPs in certain genes (for example, ApoE) yields too low diagnostic accuracy and is therefore not recommended as diagnostic test for AD. Furthermore, Ray et al. yielded promising results by the identification of 18 proteins in blood plasma that could differentiate AD patients from controls with 90% accuracy [25].

Here, we investigate whether blood-borne miRNA expression signatures might contribute to AD diagnosis. Until now, many efforts have been made to understand the role of miRNAs in neurodegenerative disorders, as summarized by Eacker et al. [26]. However, there are only two publications dealing with the miRNA expression in peripheral blood mononuclear cells (PBMC) of AD patients. The study by Villa et al. analyzed the expression of heterogeneous nuclear ribonucleoprotein (hnRNP)-A1, that is involved in the maturation of APP mRNA, and showed that the decreased expression of hsa-miR-590-3p is negatively correlated with the increased hnRNP-A1 mRNA levels [27]. The study by Schipper et al. [28] investigated the expression of 462 different miRNAs in PBMCs of 16 AD patients and 16 healthy controls to identify miRNAs that are responsible for the regulation of transcription of mRNA species that were previously reported to be downregulated in PBMCs of AD patients [29]. Only a modest relative increase of miRNA expression in AD PBMC in the range of 1.1- to 1.4-fold was found for nine miRNAs, namely hsa-miR-34a, hsa-miR-579, hsa-miR-181b, hsa-miR-520h, hsa-miR-155, hsa-miR-517*, hsa-let-7f, hsa-miR-200a, and hsa-miR-371. These data link the development of AD pathology to systemic dysfunction in the cellular stress/antioxidant response and genomic maintenance [28].

Using high throughput sequencing, we identified 140 unique miRNAs from 180 precursors that were differentially expressed between whole blood obtained from AD patients and healthy controls. It is incumbent upon the investigator, who proposes a set of miRNAs as done here to examine whether there is any known connection of these miRNAs and their target genes to neurodegeneration. Below we discuss this aspect in respect to our findings of dys-regulated miRNAs in blood of AD patients compared to healthy controls.

According to our TAM analysis out of the downregulated miRNAs, six were associated with the disease category Alzheimer disease including hsa-miR-21, hsa-miR-17, hsa-miR-29a, hsa-miR-29b, hsa-miR-106b, and hsa-miR-107. In a mouse model, Wang et al. investigated the involvement of hsa-miR-106b in the TGF-β signaling pathway that plays a key role in the pathogenesis of AD and found an inverse correlation between the expression of hsa-miR-106b and TGF-β type II receptor (TβR II) protein level [30]. In addition, Hebert et al. showed that hsa-miR-106b affects the expression of Amyloid precursor protein (APP) *in vitro*. Furthermore, they found a statistically significant decrease in hsa-miR-106b expression in sporadic AD patients, but the correlation between miR-106b and APP expression in AD brain was not significant [31]. The same group showed an inverse correlation between increased BACE1 levels and decreased miR-29a/b-1 expression [15]. Shioya et al. also observed a decreased expression of hsa-miR-29a in brain tissue of AD patients [32]. They also identified neuron navigator 3 (NAV3), a regulator of axon guidance, as a target of hsa-miR-29a and found elevated NAV3 mRNA levels in AD brains [32]. Hsa-miR-17 was shown to regulate APP expression *in vitro* and under physiological conditions in cells [31,33]. MiR-21 was shown to be downregulated in time-course assays of mature murine primary hippocampal cell cultures after neuronal Aβ treatments [34].

We further performed over-representation analysis with the 2,354 predicted targets of the 10 known miRNAs of our 12-miRNA signature. Here, several GO categories, with significant enrichment of miRNA targets in the GO categories linked to the nervous system, were found. Most interestingly, some of these target genes have already been related to AD or other of the investigated neurological diseases. One of the most prominent examples is DRD1 that encodes the Dopamine receptor D1, which is the most abundant dopamine receptor in the central nervous system. DRD1 is associated with AD, BD, and SCHIZ. Another example, DISC1 (disrupted in SCHIZ), associated with BD and SCHIZ, encodes a protein involved in neurite outgrowth and cortical development. BDNF (brain-derived neurotrophic factor) important for survival of striatal neurons in the brain is known to be downregulated in AD patients and also associated with BD, DEP, PD, and SCHIZ. IGF1R is the only target gene that was exclusively found to be associated with AD. The protein encoded by this gene is increased in temporal cortex surrounding and within Aß-containing plaques, but a significantly lower number of neurons of AD patients express IGF1R [35].

Table 4 Results of the over-representation analysis of the predicted target genes of brain-miR-112.

Subcategory	Subcategory alternative name	Expected	Observed	*P*value (FDR)	Target genes	AD	BD	DEP	PD	SCHIZ	Multiple sclerosis
Neurogenesis	GO:0022008	119.612	31	0.000284421	ONECUT2 ANK3 CACNB3 CDK6 CELSR3 FGFR2 MEF2A NFIB PICALM PLAG1 PLXNA1 PSD4 PTPRR RAB11A RPS6KA4 SIX4 COL4A4 DFNB31 DISC1 SRF STX3 ADCY1 CDKN1C CNP ENAH HOXC10 LIF LRP6 ROCK1 RPS6KA3 TFAP2A	–	DISC1	–	–	DISC1	–
Neuron differentiation	GO:0030182	103.937	27	0.000801251	ONECUT2 ANK3 CACNB3 CELSR3 FGFR2 MEF2A NFIB PICALM PLXNA1 PSD4 PTPRR RAB11A RPS6KA4 COL4A4 DFNB31 SRF STX3 ADCY1 CDKN1C CNP ENAH HOXC10 LIF LRP6 ROCK1 RPS6KA3 TFAP2A	–	–	–	–	–	–
Neuron development	GO:0048666	845.125	23	0.00159317	ONECUT2 ANK3 CACNB3 CELSR3 FGFR2 MEF2A NFIB PICALM PLXNA1 RAB11A RPS6KA4 COL4A4 DFNB31 SRF STX3 ADCY1 CDKN1C CNP ENAH LIF ROCK1 RPS6KA3 TFAP2A	–	–	–	–	–	–
Nervous system development	GO:0007399	184.019	37	0.00268227	ONECUT2 ANK3 CACNB3 CDK6 CELSR3 FGFR2 MEF2A NFIB PICALM PLAG1 PLXNA1 PSD4 PTPRR RAB11A RPS6KA4 SEMA5B SIX4 COL4A4 DFNB31 DISC1 FGF1 SRF STX3 SULF1 ADCY1 ARHGEF15 CDKN1C CNP ENAH HOXC10 LIF LPHN1 LRP6 MEN1 ROCK1 RPS6KA3 TFAP2A	FGF1	DISC1	–	–	DISC1	–
Neuron projection development	GO:0031175	736.077	19	0.00946322	ANK3 CACNB3 CELSR3 FGFR2 MEF2A NFIB PICALM PLXNA1 RAB11A RPS6KA4 COL4A4 SRF STX3 ADCY1 CNP ENAH LIF ROCK1 RPS6KA3	–	–	–	–	–	–
Neuron projection	GO:0043005	633.844	16	0.0262424	ALOX5 ANK3 MYLK2 NFIB SLC38A7 DFNB31 DISC1 FRMPD4 GRIA4 SLC6A1 STX3 AAK1 ALDOC ARHGEF15 BACE1 LPHN1	BACE1	DISC1	–	–	DISC1 GRIA4	–
Neurotransmitter: sodium symporter activity	GO:0005328	0.215825	3	0.0330004	SLC6A20 SLC6A1 SLC6A6	–	–	–	–	–	–
Neuron projection morphogenesis	GO:0048812	624.756	15	0.0391162	ANK3 CACNB3 CELSR3 FGFR2 MEF2A NFIB PICALM PLXNA1 RPS6KA4 COL4A4 ADCY1 CNP ENAH ROCK1 RPS6KA3	–	–	–	–	–	–
Neurotransmitter transporter activity	GO:0005326	0.272621	3	0.0412241	SLC6A20 SLC6A1 SLC6A6	–	–	–	–	–	–
Neuroblast division	GO:0055057	0.0795144	2	0.0412241	FGFR2 LRP6	–	–	–	–	–	–
Forebrain neuroblast division	GO:0021873	0.0795144	2	0.0412241	FGFR2 LRP6	–	–	–	–	–	–
Generation of neurons	GO:0048699	112.797	31	0.000188727	ONECUT2 ANK3 CACNB3 CDK6 CELSR3 FGFR2 MEF2A NFIB PICALM PLAG1 PLXNA1 PSD4 PTPRR RAB11A RPS6KA4 SIX4 COL4A4 DFNB31 DISC1 SRF STX3 ADCY1 CDKN1C CNP ENAH HOXC10 LIF LRP6 ROCK1 RPS6KA3 TFAP2A	–	DISC1	–	–	DISC1	–
Cell morphogenesis involved in neuron differentiation	GO:0048667	616.805	15	0.0366659	ANK3 CACNB3 CELSR3 FGFR2 MEF2A NFIB PICALM PLXNA1 RPS6KA4 COL4A4 ADCY1 CNP ENAH ROCK1 RPS6KA3	–	–	–	–	–	–

This table lists interesting Gene Ontology (GO) subcategories with *P*values <0.05 and FDR adjustment related to nervous system development.
AD: Alzheimer disease; BP: bipolar disorder; DEP: major depression; PD, Parkinson's disease; SCHIZ: schizophrenia

This suggests that IGF1R signaling normally controlling vital growth, survival, and metabolic functions in the brain is disturbed in AD brains. The two unknown miRNAs revealed 234 target genes for brain-miR-112, but only six target genes for brain-miR-161. In the over-representation analysis for brain-miR-112 we also identified GO categories linked to the nervous system, including targets like DISC1 as discussed above. For brain-miR-161 we found no significant GO categories. However, a literature review of the six target genes of brain-miR-161 revealed some interesting findings. GRID1 (glutamate receptor, ionotropic, delta 1), predicted to be a target gene of brain-miR-161, encodes a gene product that is a subunit of glutamate receptor channels which mediate most of the fast excitatory synaptic transmission in the central nervous system and play key roles in synaptic plasticity. Interestingly, GRID1 has previously been associated with SCHIZ and BD [36-38]. Another predicted target gene CCDN2 (Cyclin D2) plays a role in corticogenesis [39].

However, we have to point out that our analysis is based on whole blood. Previous findings on cancer suggest that the miRNA expression pattern between blood cells and cancer tissue do not necessarily show the same expression pattern but some overlaps can be found [40-42]. Unfortunately, tissue and blood samples of the same patients were not available for the present study. Nevertheless, we performed database analysis and extracted all miRNAs deregulated in AD and the corresponding literature out of the Human MiRNA& Disease Database [43]. In total, we found 18 different publications, with 15 publications on AD brain tissue and/or cell culture models. Out of those studies, 29 different miRNAs deregulated in AD are listed in the HMDD. Comparing these miRNAs with our data revealed eight of the 29 miRNAs that were significantly dys-regulated in blood cells in our study. There is, however, no evidence whether these overlaps were found by chance or not. Any link between deregulated miRNAs in blood of patients with neurological diseases and the disease itself has to be considered with caution.

Since a large set of miRNAs often contains a significant amount of redundant miRNAs with largely identical information content the differentiation between AD samples and healthy controls using a reduced set of miRNAs may likely yield comparably accurate results. Therefore, a panel of 12 miRNAs with limited cross-correlation, including most strongly dys-regulated miRNAs that show a potential to separate AD from controls, was selected. Some of these 12 miRNAs have already been related to AD. For example, Wang et al. showed in a computational analysis that the 3'-untranslated region (UTR) of beta-site amyloid precursor protein-cleaving enzyme 1 (BACE1) mRNA is targeted by hsa-miR-107 and that BACE1 mRNA levels tended to increase as miR-107 levels decreased in the progression for AD. An increased BACE1 expression is an important risk factor for sporadic AD [15]. Nelson et al. also showed a negative correlation between the expression of hsa-miR-107 and BACE1 [44]. Interestingly, hsa-miR-107 that was also part of our 12-miRNA signature investigated in the presented study was also downregulated in blood of AD patients compared to healthy controls. Augustin et al. [45] recently investigated miRNAs that are predicted to target another AD-related gene, namely ADAM10, which controls the proteolytic processing of *APP* and the formation of the amyloid plaques. Database analyses prompted them to further investigate two miRNAs that were also included in our 12-miRNA signature, namely hsa-miR-107 and hsa-miR-103. They found that predicted target genes of hsa-miR-107 and hsa-miR-103 showed significant overlap with the AlzGene database. In a reporter assay ADAM10 expression was reduced by both miRNAs. These two miRNAs were also investigated in relation to the expression of cofilin protein in a transgenic mouse model [46]. Cofilin binds to actin resulting in the formation of Hirano bodies, which may play an essential role in AD pathogenesis. In APP transgenic mouse brains hsa-miR-107 and hsa-miR-103 levels were decreased while cofilin levels were increased and in a luciferase assay it was demonstrated that hsa-miR-107 and hsa-miR-103 were able to reduce the expression of cofilin. In our RT-qPCR approach both miRNAs hsa-miR-107 and hsa-miR-103 showed the same expression pattern, that is, both were downregulated in blood of AD, PD, DEP, and SCHIZ patients and upregulated in mild cognitive impairment, multiple sclerosis, and BD patients. All other miRNAs of our 12-miRNA signature have not been identified or investigated so far in relationship to AD.

While we showed the 12-miRNA signature's potential to separate AD patients from controls with an accuracy of 93.3%, we also tested its applicability as differential diagnostic biomarker to separate AD from other neurological diseases. As we expected, the accuracy decreased when trying to use this signature for separating other neurodegenerative diseases from controls or separating AD from other neurological disorders. Remarkably, the classification of psychiatric disorders *versus* controls yielded an even better accuracy than for AD *versus* controls. These findings suggest a relevance of the considered 12 miRNAs also for psychological disorders. The association of the 12-miRNA signature with neurological diseases in general is further underlined by the results of our over-representation analysis using GeneTrail. Here, we found four significant GO categories related to nervous system and neurons with an over-representation of target genes of the 10 known miRNAs from our 12-miRNA signature. In addition, out of the 10 known miRNAs nine miRNAs are already included in the HMDD and five of those miRNAs that were previously associated with neurological diseases

including AD, PD, and SCHIZ. As mentioned above, Yao et al. [46] showed that reduced levels of hsa-miR-103 or hsa-miR-107 are associated with elevated cofilin protein levels and formation of rod-like structures in a transgenic mouse model of AD. Both miRNAs were also downregulated in our study. Martins et al. [47] showed that hsa-miR-151a-3p and hsa-miR-26a-5p are differentially expressed in PBMCs (peripheral blood mononuclear cells) of PD patients and controls. In prefrontal cortex tissue of individuals with SCHIZ hsa-miR-26b was downregulated [48]. Target analysis of the miRNA that was not included in HMDD, hsa-miR-5010-3p, revealed target genes involved in nervous system processes. For example, predicted targets of hsa-miR-5010-3p include the NFASC (neurofascin), that functions in neurite outgrowth, and organization of nodes of Ranvier on axons, NPY (Neuropeptide Y), that is one of the most abundant neuropeptides in the mammalian central nervous system [49], NLGN1 (neuroligin 1), that may be involved in the formation and remodeling of central nervous system synapses, NRXN3 (neurexin3), that functions in the nervous system as receptors and cell adhesion molecule, and NCAN (neurocan), that seems to be a genetic risk factor for BD.

Finally, one has to take into account that AD is a complex progressive neurodegenerative disease causing cognitive, behavioral, and functional problems that are also found in other neurological diseases. Furthermore, dementia is not only caused by AD but can result from other neurological disorders. Dementia patients often suffer from other additional mental and behavioral problems like depression, anxiety, psychosis, agitation, and aggression further complicating correct classification. As AD shares common neuropsychiatric symptoms with other neurological diseases there might be an overlap with the associated medication.

Most importantly one needs to point out that as the patients included in our study are not treatment-naïve, we cannot exclude the influence of administered drugs on the miRNA signature. As an example, Bocchio-Chiavetto et al. showed that chronic anti-depressant treatment has effects on the blood miRNA profile [50]. Furthermore, we have to point out that we do not have a birth cohort. Nevertheless, the age distribution between the AD samples and the control samples used for NGS is not significantly different (*P* value 0.1147). The age distribution of AD patients, MCI patients, PD patients, and controls is quite similar. Patients suffering from multiple sclerosis, DEP, BD, or SCHIZ are about 20 to 30 years younger. The differences in the age distribution are due to the differences between the onsets of the diseases. In previous studies [51] we already investigated the influence of age and gender on the miRNA expression profile of whole blood. We did not find any statistically significant deregulated miRNAs between men and women. The miRNA with the lowest *P* value was hsa-miR-423 (*P* value 0.78). To test for the influence of age we compared the profiles obtained from old *versus* young patients by splitting the total group in half based on the age. Here, the miRNA with the lowest *P* value was hsa-miR-890 (*P* value 0.87). Again, we did not find any deregulated miRNAs. In summary, we found no evidence that age and gender have a substantial influence on the miRNA profiles. Both miRNAs mentioned above were not significant in the present study on AD.

Conclusion

Here we identified 140 unique differentially expressed miRNAs between AD patients and healthy controls. Using a signature of 12 miRNAs differentially expressed between AD patients and healthy controls we were not only able to distinguish with high diagnostic accuracies between AD patients and healthy controls, but also between AD patients and patients suffering from other neurological disorders including mild cognitive impairment as a potential preliminary stage of AD, and other neurodegenerative diseases like PD and multiple sclerosis as well as mental diseases like SCHIZ, DEP, and BD. However, additional work will be needed to elucidate the applicability of this 12-miRNA signature as a potential diagnostic test for AD and the above-mentioned effects of the drug treatments commonly used in the treatment of the disease. Hopefully, tests of this non-invasive and relatively cheap kind will be applicable to prodromal AD cases and to MCI patients with the aim to recognize early AD to initiate treatment.

Materials and methods

Patient details

We analyzed the expression of miRNAs in peripheral blood of a total of 215 patients and healthy controls, either by NGS or by RT-qPCR or by both methods (see Table 1). In detail, we obtained 2.5 mL blood collected in PAXgene Blood RNA tubes (PreAnalytiX) from patients with AD (n = 106), patients with mild cognitive impairment (MCI) (n = 18), patients with multiple sclerosis (clinically isolated syndrome, CIS) (n = 16), patients with PD (n = 9), patients with DEP (n = 15), patients with BD (n = 15), patients with SCHIZ (n = 14), and from healthy controls (C) (n = 22). Samples from patients with AD stem from the Biorepository and Tissue Bank PrecisionMed (San Diego, CA, USA) (n = 97) and the University Clinic of Erlangen (Germany) (n = 9), samples from healthy controls and from patients with MCI, PD, DEP, BD, and SCHIZ stem from PrecisionMed (San Diego, CA, USA) and samples from patients with CIS stem from Charité Berlin (Germany). Detailed patient characteristics are listed in Additional file 4-Table S3. AD and MCI patients were diagnosed by using state of the art criteria. In detail, in order to be included in the 'probable AD' group, patients fulfilled the following criteria of the

NINCDS-ADRDA (National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer disease and Related Disorders Association) [52]: MMSE >14 and <26, deficit in two or more areas of cognition, progressive worsening of memory and other cognitive functions, no disturbance of consciousness, onset between the ages of 40 and 90 years, most often after 65 years, and absence of systemic disorders or other brain diseases that could account for the progressive deterioration in cognition. Furthermore, MRI or CT reports that were compatible with AD are available. The median MMSE score for the AD patients was 18.9 (3.4).

Samples included in the MCI group fulfilled the following criteria: MMSE >22 and <28, not demented, memory complaint, preserved general cognitive function, intact activities of daily living: (allowed problems with 2 or less of the following: phone calls, meal preparation, handling money, completing chores), abnormal memory function documented by scoring below the education adjusted cutoff on the Logical Memory II subscale (delayed paragraph recall) from the Wechsler Memory Scale-Revised (maximum score = 25) with (a) <8 for 16 years or more of education, (b) <4 for 8-15 years of education, (c) <2 for 0-7 years of education. The median MMSE score for the MCI patients was 25.3 (±1.4).

The study was approved by the institutional review boards of Charité - Universitätsmedizin Berlin (EA1/182/10) and the study was performed in accordance with the Helsinki declaration. Written informed consent was obtained from all patients participating in the study.

Samples and clinical data supplied by PrecisionMed are handled in strictest compliance with all applicable rules and regulations including the recommendations of the Council of the Human Genome Organization (HUGO) Ethical, Legal, and Social Issues Committee (HUGO-ELSI, 1998); with the United Nations Educational, Scientific, and Cultural Organization's (UNESCO) Universal Declaration on the Human Genome and Human Rights (1997); and with recommendations guiding physicians in biomedical research involving human subjects adopted by the 18th World Medical Assembly, Helsinki, Finland, 1964 and later revisions.

RNA isolation

Total RNA including miRNA was isolated using the PAXgene Blood miRNA Kit (Qiagen) following the manufacturer's recommendations. Isolated RNA was stored at -80° C until use. RNA integrity was analyzed using Bioanalyzer 2100 (Agilent) and concentration and purity were measured using NanoDrop 2000 (Thermo Scientific).

Library preparation and next-generation sequencing

We first analyzed samples from AD patients (n = 48) and healthy controls (n = 22) by NGS.

For the library preparation, 200 ng of total RNA was used per sample, as determined with a RNA 6000 Nano Chip on the Bioanalyzer 2100 (Agilent). Preparation was performed following the protocol of the TruSeq Small RNA Sample Prep Kit (Illumina). Concentration of the ready prepped libraries was measured on the Bioanalyzer using the DNA 1000 Chip. Libraries were then pooled in batches of six samples in equal amounts and clustered with a concentration of 9 pmol in one lane each of a single read flowcell using the cBot (Illumina). Sequencing of 50 cycles was performed on a HiSeq 2000 (Illumina). Demultiplexing of the raw sequencing data and generation of the fastq files was done using CASAVA v.1.8.2.

NGS data analysis

The raw Illumina reads were first preprocessed by cutting the 3' adapter sequence using the program fastx_clipper from the FASTX-Toolkit [53]. Reads shorter than 18 nts after clipping were removed. The remaining reads are reduced to unique reads and their frequency per sample to make the mapping steps more time efficient. For the remaining steps, we used the miRDeep2 pipeline [54]. These steps consist of mapping the reads against the genome (hg19), mapping the reads against miRNA precursor sequences from miRBase release v18, summarizing the counts for the samples, and the prediction of novel miRNAs. Since the miRDeep2 pipeline predicts in our case the novel miRNAs per sample, we merged the miRNAs afterwards as follows: first, we extract the novel miRNAs per sample that have a signal-to-noise ratio >10. Subsequently, we merge only those novel miRNAs that are located on the same chromosome, and both their mature forms share an overlap of at least 11 nucleotides. The remaining putative novel miRNAs were mapped with BLAST (v 2.2.24, [55]) against known ncRNA and miRNA sequences from diverse sources (miRBase v18 [56], snoRNA-LBME-db [57], ncRNAs from Ensembl 'Homo_sapiens.GRCh37.67.ncrna.fa'[58], NONCODE v3.0[59]). We excluded sequences that aligned with >90% of their length (allowing 1 mismatch) to any of the ncRNA sequences. All NGS data are publicly available in GEO database (GSE46579 [60]).

Bioinformatics analysis

For the NGS analysis, we excluded miRNAs with <50 read counts summed up across all samples of each group (AD or control), since these were considered lowly abundant. We normalized the read counts using standard quantile normalization. Next, we calculated for each miRNA the area under the receiver operator characteristic curve (AUC), the fold-change, and the significance value (P value) using Wilcoxon-Mann-Whitney (WMW) test. All significance values were adjusted for multiple testing using the Benjamini-Hochberg approach [61,62].

The bioinformatics analyses have been carried out using the freely available tool R [63]. For classification purposes, we used support vector machines (SVM) from the R package e1071. If not stated otherwise, we computed the group-wise classifications using linear kernels in 10-fold cross-validations with 100 repetitions. In addition, we computed the classification of permuted class labels with the same parameters as control. If group sizes were unbalanced, we randomly selected samples from the bigger group to match the sample sizes in the smaller group in each repetition.

Database analysis

MiRNA enrichment analysis was performed using the TAM tool [23,24]. The miRNA targets of the known miRNAs were predicted using miRDB [64-66]. Targets for the unknown brain-miRs were predicted using TargetScan [67,68]. TargetScan is able to predict targets of miRBase miRNAs as well as targets of other sequences by using the heptamer seed sequence (nucleotides 2-8) of a potential miRNA. For brain-miR-161 we used the heptamer UUCGAAA, for brain-mir-112 GCUCUGU. With the predicted miRNA target genes we performed an over-representation analysis using the gene set analysis tool GeneTrail [69,70] with default settings. The *P* values for the biological categories were adjusted by False Discovery Rate (FDR) [71] and were considered significant if <0.05. Furthermore, we searched for miRNA-disease interactions using the Human MiRNA& Disease Database (HMDD [43,72]).

Quantitative real time-PCR (RT-qPCR)

For validation purposes we analyzed the expression of single miRNAs using quantitative real time-polymerase chain reaction (RT-qPCR) in the same samples as used for NGS, if sufficient amounts of RNA were available. We used the miScript PCR System (Qiagen) for reverse transcription and RT-qPCR. A total of 200 ng RNA was converted into cDNA using the miScript Reverse Transcription Kit according to the manufacturer's protocol. The RT-qPCR was performed with the miScript SYBR® Green PCR Kit in a total volume of 20 μL per reaction containing 1 μL cDNA according to the manufacturer's protocol. For each miScript Primer Assay we additionally prepared a PCR negative-control with water instead of cDNA (non-template control).

We further expanded the number of samples by further samples from patients with AD, MCI, CIS, PD, DEP, BD, and SCHIZ, resulting in a total of 202 samples analyzed by RT-qPCR (see Table 1). In detail, we analyzed with RT-qPCR a total of 94 samples from AD patients, 18 samples from MCI patients, 16 samples from CIS patients, nine samples from PD patients, 15 samples from DEP patients, 15 samples from BD patients, 14 samples from SCHIZ patients, and 21 samples from healthy controls.

Out of the NGS results we selected 12 miRNAs deregulated between patients with AD and healthy individuals. The set contained the following miRNAs: The upregulated miRNAs brain-miR-112, brain-miR-161, hsa-let-7d-3p, hsa-miR-5010-3p, hsa-miR-26a-5p, hsa-miR-1285-5p, and hsa-miR-151a-3p as well as the downregulated miRNAs hsa-miR-103a-3p, hsa-miR-107, hsa-miR-532-5p, hsa-miR-26b-5p, and hsa-let-7f-5p, respectively.

While 10 of the 12 miRNAs have already been annotated in the miRBase, two miRNAs, namely brain-miR-112 and brain-miR-161, were newly identified and not yet included in miRBase [21,22]. As endogenous control we used the small nuclear RNA RNU48.

Additional material

Additional file 1: Table S1. Table listing the 180 significantly dys-regulated miRNAs (140 unique mature miRNAs).

Additional file 2: Figure S1. Heatmap for the 180 miRNAs significantly dys-regulated in AD patients compared to control individuals.

Additional file 3: Table S2. Table listing all novel mature miRNAs.

Additional file 4: Table S3. Table listing patient characteristics and indicates which samples are included in NGS analysis and/or in the RT-qPCR.

List of abbreviations

ACE: Aging Cognition Evaluation; AD: Alzheimer disease; ADAS-Cog: Alzheimer disease Assessment Scale-cognitive subscale; AUC: area under the receiver operator characteristics curve; BD: bipolar disorder; C: healthy control; CDR: Clinical Dementia Rating; CIS: clinically isolated syndrome; CLIA: Clinical Laboratory Improvement Amendments; CNS: central nervous system; CSF: cerebrospinal fluid; DEP: major depression; FDR: False Discovery Rate; GO: Gene Ontology; HMDD: Human MiRNA& Disease Database; HUGO: Human Genome Organization; MCI: mild cognitive impairment; miRNA: micro Ribo Nucleic Acid; MMSE: Mini-Mental State Exam; mtDNA: mitochondrial DNA; NGS: next generation sequencing; PBMC: peripheral blood mononuclear cells; PD: Parkinson disease; RT-qPCR: quantitative Real Time Polymerase Chain Reaction; SAMPLE: Serial Alzheimer disease and MCI Prospective Longitudinal Evaluation; SCHIZ: schizophrenia; SNP: single nucleotide polymorphisms; SVM: support vector machines; TAM: tool for annotations of human miRNAs; TF: transcription factor; UBC: University of British Columbia; UCSC: University of California Santa Cruz; UNESCO: United Nations Educational, Scientific, and Cultural Organization; WMW: Wilcoxon-Mann-Whitney.

Competing interests

AK and CS are employees of Siemens Healthcare. Siemens Healthcare in part supported this work.

Authors' contributions

The work presented here was carried out in collaboration between all authors. EM, AK, and CS initiated the study. EM, AK, and BM designed the study and KR and FP participated in the design of the study. PL, SD, KS, JH, KF, and BM performed the laboratory experiments. CB and AK analyzed and interpreted the data. SM assisted in data analysis. CL provided the clinical samples collected in Erlangen. PL and CB performed database analysis. TB interpreted the data and revised the manuscript critically. PL, CB, AK and EM wrote the paper. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by Siemens Healthcare, and in part by DFG grants LE2783/1-1 and ME917/20-1.

Authors' details
[1]Department of Human Genetics, Saarland University, Kirrbergerstraße, Building 60, 66421 Homburg, Germany. [2]Internal Medicine II, Heidelberg University, Im Neuenheimer Feld 350, 69120 Heidelberg, Germany. [3]Clinical and Experimental Multiple Sclerosis Research Center, Charité - University Medicine Berlin, Campus Mitte, Charitéplatz 1, 10117 Berlin, Germany. [4]NeuroCure Clinical Research Center, Charité - University Medicine Berlin, Campus Mitte, Charitéplatz 1,10117 Berlin, Germany. [5]Siemens Healthcare, Strategy, Hartmannstr. 16, 91052 Erlangen, Germany. [6]Neurological Unit, University of Erlangen, Schwabachanlage 6, 91054 Erlangen, Germany. [7]Department of Chemical Physiology, The Scripps Research Institute, 10550 N Torrey Pines Rd, La Jolla, CA 92037, USA.

Received: 14 February 2013 Revised: 18 June 2013
Accepted: 29 July 2013 Published: 29 July 2013

References

1. Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM: **Forecasting the global burden of Alzheimer's disease.** *Alzheimers Dement* 2007, **3**:186-191.
2. Brookmeyer R, Gray S, Kawas C: **Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset.** *Am J Public Health* 1998, **88**:1337-1342.
3. Gandy S: **Perspective: prevention is better than cure.** *Nature* 2011, **475**: S15.
4. Deutsche Gesellschaft für Psychiatrie PuND, (DGN) DGfrN: *S3-Leitlinie "Demenzen"* 2009.
5. Fita IG, Enciu AM, Stanoiu BP: **New insights on Alzheimer's disease diagnostic.** *Rom J Morphol Embryol* 2011, **52**:975-979.
6. Frankfort SV, Tulner LR, van Campen JP, Verbeek MM, Jansen RW, Beijnen JH: **Amyloid beta protein and tau in cerebrospinal fluid and plasma as biomarkers for dementia: a review of recent literature.** *Curr Clin Pharmacol* 2008, **3**:123-131.
7. Malaplate-Armand C, Desbene C, Pillot T, Olivier JL: **[Biomarkers for early diagnosis of Alzheimer's disease: current update and future directions].** *Rev Neurol (Paris)* 2009, **165**:511-520.
8. Gasparini L, Racchi M, Binetti G, Trabucchi M, Solerte SB, Alkon D, Etcheberrigaray R, Gibson G, Blass J, Paoletti R, Govoni S: **Peripheral markers in testing pathophysiological hypotheses and diagnosing Alzheimer's disease.** *FASEB J* 1998, **12**:17-34.
9. Gerrish A, Russo G, Richards A, Moskvina V, Ivanov D, Harold D, Sims R, Abraham R, Hollingworth P, Chapman J, Hamshere M, Pahwa JS, Dowzell K, Williams A, Jones N, Thomas C, Stretton A, Morgan AR, Lovestone S, Powell J, Proitsi P, Lupton MK, Brayne C, Rubinsztein DC, Gill M, Lawlor B, Lynch A, Morgan K, Brown KS, Passmore PA, *et al*: **The role of variation at AbetaPP, PSEN1, PSEN2, and MAPT in late onset Alzheimer's disease.** *J Alzheimers Dis* 2012, **28**:377-387.
10. Hudson G, Sims R, Harold D, Chapman J, Hollingworth P, Gerrish A, Russo G, Hamshere M, Moskvina V, Jones N, Thomas C, Stretton A, Holmans PA, O'Donovan MC, Owen MJ, Williams J, Chinnery PF: **No consistent evidence for association between mtDNA variants and Alzheimer disease.** *Neurology* 2012, **78**:1038-1042.
11. Doecke JD, Laws SM, Faux NG, Wilson W, Burnham SC, Lam CP, Mondal A, Bedo J, Bush AI, Brown B, De Ruyck K, Ellis KA, Fowler C, Gupta VB, Head R, Macaulay SL, Pertile K, Rowe CC, Rembach A, Rodrigues M, Rumble R, Szoeke C, Taddei K, Taddei T, Trounson B, Ames D, Masters CL, Martins RN: **Blood-based protein biomarkers for diagnosis of Alzheimer disease.** *Arch Neurol* 2012, **69**:1-8.
12. Tan M, Wang S, Song J, Jia J: **Combination of p53(ser15) and p21/p21 (thr145) in peripheral blood lymphocytes as potential Alzheimer's disease biomarkers.** *Neurosci Lett* 2012, **516**:226-231.
13. Keller A, Leidinger P, Bauer A, Elsharawy A, Haas J, Backes C, Wendschlag A, Giese N, Tjaden C, Ott K, Werner J, Hackert T, Ruprecht K, Huwer H, Huebers J, Jacobs G, Rosenstiel P, Dommisch H, Schaefer A, Muller-Quernheim J, Wullich B, Keck B, Graf N, Reichrath J, Vogel B, Nebel A, Jager SU, Staehler P, Amarantos I, Boisguerin V, *et al*: **Toward the blood-borne miRNome of human diseases.** *Nat Methods* 2011, **8**:841-843.
14. Liang Y, Ridzon D, Wong L, Chen C: **Characterization of microRNA expression profiles in normal human tissues.** *BMC Genomics* 2007, **8**:166.
15. Hebert SS, Horre K, Nicolai L, Papadopoulou AS, Mandemakers W, Silahtaroglu AN, Kauppinen S, Delacourte A, De Strooper B: **Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/beta-secretase expression.** *Proc Natl Acad Sci USA* 2008, **105**:6415-6420.
16. Wang WX, Rajeev BW, Stromberg AJ, Ren N, Tang G, Huang Q, Rigoutsos I, Nelson PT: **The expression of microRNA miR-107 decreases early in Alzheimer's disease and may accelerate disease progression through regulation of beta-site amyloid precursor protein-cleaving enzyme 1.** *J Neurosci* 2008, **28**:1213-1223.
17. Geekiyanage H, Chan C: **MicroRNA-137/181c regulates serine palmitoyltransferase and in turn amyloid beta, novel targets in sporadic Alzheimer's disease.** *J Neurosci* 2011, **31**:14820-14830.
18. Long JM, Lahiri DK: **MicroRNA-101 downregulates Alzheimer's amyloid-beta precursor protein levels in human cell cultures and is differentially expressed.** *Biochem Biophys Res Commun* 2010, **404**:889-895.
19. Geekiyanage H, Jicha GA, Nelson PT, Chan C: **Blood serum miRNA: non-invasive biomarkers for Alzheimer's disease.** *Exp Neurol* 2011, **235**:491-496.
20. Braskie MN, Toga AW, Thompson PM: **Recent advances in imaging Alzheimer's disease.** *J Alzheimers Dis* 2013, , **Suppl 1:** S313-327.
21. **miRBase..** .
22. Kozomara A, Griffiths-Jones S: **miRBase: integrating microRNA annotation and deep-sequencing data.** *Nucleic Acids Res* 2010, **39**:D152-157.
23. **TAM - tool for annotations of human miRNAs.** [http://202.38.126.151/hmdd/tools/tam.html].
24. Lu M, Shi B, Wang J, Cao Q, Cui Q: **TAM: a method for enrichment and depletion analysis of a microRNA category in a list of microRNAs.** *BMC Bioinformatics* 2010, **11**:419.
25. Ray S, Britschgi M, Herbert C, Takeda-Uchimura Y, Boxer A, Blennow K, Friedman LF, Galasko DR, Jutel M, Karydas A, Kaye JA, Leszek J, Miller BL, Minthon L, Quinn JF, Rabinovici GD, Robinson WH, Sabbagh MN, So YT, Sparks DL, Tabaton M, Tinklenberg J, Yesavage JA, Tibshirani R, Wyss-Coray T: **Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins.** *Nat Med* 2007, **13**:1359-1362.
26. Eacker SM, Dawson TM, Dawson VL: **Understanding microRNAs in neurodegeneration.** *Nat Rev Neurosci* 2009, **10**:837-841.
27. Villa C, Fenoglio C, De Riz M, Clerici F, Marcone A, Benussi L, Ghidoni R, Gallone S, Cortini F, Serpente M, Cantoni C, Fumagalli G, Martinelli Boneschi F, Cappa S, Binetti G, Franceschi M, Rainero I, Giordana MT, Mariani C, Bresolin N, Scarpini E, Galimberti D: **Role of hnRNP-A1 and miR-590-3p in neuronal death: genetics and expression analysis in patients with Alzheimer disease and frontotemporal lobar degeneration.** *Rejuvenation Res* 2011, **14**:275-281.
28. Schipper HM, Maes OC, Chertkow HM, Wang E: **MicroRNA expression in Alzheimer blood mononuclear cells.** *Gene Regul Syst Bio* 2007, **1**:263-274.
29. Maes OC, Xu S, Yu B, Chertkow HM, Wang E, Schipper HM: **Transcriptional profiling of Alzheimer blood mononuclear cells by microarray.** *Neurobiol Aging* 2007, **28**:1795-1809.
30. Wang H, Liu J, Zong Y, Xu Y, Deng W, Zhu H, Liu Y, Ma C, Huang L, Zhang L, Qin C: **miR-106b aberrantly expressed in a double transgenic mouse model for Alzheimer's disease targets TGF-beta type II receptor.** *Brain Res* 2010, **1357**:166-174.
31. Hebert SS, Horre K, Nicolai L, Bergmans B, Papadopoulou AS, Delacourte A, De Strooper B: **MicroRNA regulation of Alzheimer's Amyloid precursor protein expression.** *Neurobiol Dis* 2009, **33**:422-428.
32. Shioya M, Obayashi S, Tabunoki H, Arima K, Saito Y, Ishida T, Satoh J: **Aberrant microRNA expression in the brains of neurodegenerative diseases: miR-29a decreased in Alzheimer disease brains targets neurone navigator 3.** *Neuropathol Appl Neurobiol* 2010, **36**:320-330.
33. Delay C, Calon F, Mathews P, Hebert SS: **Alzheimer-specific variants in the 3'UTR of Amyloid precursor protein affect microRNA function.** *Mol Neurodegener* 2011, **6**:70.
34. Schonrock N, Ke YD, Humphreys D, Staufenbiel M, Ittner LM, Preiss T, Gotz J: **Neuronal microRNA deregulation in response to Alzheimer's disease amyloid-beta.** *PLoS One* 2010, **5**:e11070.
35. Moloney AM, Griffin RJ, Timmons S, O'Connor R, Ravid R, O'Neill C: **Defects in IGF-1 receptor, insulin receptor and IRS-1/2 in Alzheimer's disease indicate possible resistance to IGF-1 and insulin signalling.** *Neurobiol Aging* 2010, **31**:224-243.

36. Fallin MD, Lasseter VK, Avramopoulos D, Nicodemus KK, Wolyniec PS, McGrath JA, Steel G, Nestadt G, Liang KY, Huganir RL, Valle D, Pulver AE: **Bipolar I disorder and schizophrenia: a 440-single-nucleotide polymorphism screen of 64 candidate genes among Ashkenazi Jewish case-parent trios.** *Am J Hum Genet* 2005, **77**:918-936.
37. Guo SZ, Huang K, Shi YY, Tang W, Zhou J, Feng GY, Zhu SM, Liu HJ, Chen Y, Sun XD, He L: **A case-control association study between the GRID1 gene and schizophrenia in the Chinese Northern Han population.** *Schizophr Res* 2007, **93**:385-390.
38. Treutlein J, Muhleisen TW, Frank J, Mattheisen M, Herms S, Ludwig KU, Treutlein T, Schmael C, Strohmaier J, Bosshenz KV, Breuer R, Paul T, Witt SH, Schulze TG, Schlosser RG, Nenadic I, Sauer H, Becker T, Maier W, Cichon S, Nothen MM, Rietschel M: **Dissection of phenotype reveals possible association between schizophrenia and Glutamate Receptor Delta 1 (GRID1) gene promoter.** *Schizophr Res* 2009, **111**:123-130.
39. Tsunekawa Y, Britto JM, Takahashi M, Polleux F, Tan SS, Osumi N: **Cyclin D2 in the basal process of neural progenitors is linked to non-equivalent cell fates.** *EMBO J* 2012, **31**:1879-1892.
40. Bauer AS, Keller A, Costello E, Greenhalf W, Bier M, Borries A, Beier M, Neoptolemos J, Buchler M, Werner J, Giese N, Hoheisel JD: **Diagnosis of pancreatic ductal adenocarcinoma and chronic pancreatitis by measurement of microRNA abundance in blood and tissue.** *PLoS One* 2012, **7**:e34151.
41. Fenoglio C, Ridolfi E, Galimberti D, Scarpini E: **MicroRNAs as active players in the pathogenesis of multiple sclerosis.** *Int J Mol Sci* 2012, **13**:13227-13239.
42. Waters PS, McDermott AM, Wall D, Heneghan HM, Miller N, Newell J, Kerin MJ, Dwyer RM: **Relationship between circulating and tissue microRNAs in a murine model of breast cancer.** *PLoS One* 2012, **7**:e50459.
43. **Human MiRNA & Disease Database**.
44. Nelson PT, Wang WX: **MiR-107 is reduced in Alzheimer's disease brain neocortex: validation study.** *J Alzheimers Dis* 2010, **21**:75-79.
45. Augustin R, Endres K, Reinhardt S, Kuhn PH, Lichtenthaler SF, Hansen J, Wurst W, Trumbach D: **Computational identification and experimental validation of microRNAs binding to the Alzheimer-related gene ADAM10.** *BMC Med Genet* 2012, **13**:35.
46. Yao J, Hennessey T, Flynt A, Lai E, Beal MF, Lin MT: **MicroRNA-related cofilin abnormality in Alzheimer's disease.** *PLoS One* 2010, **5**:e15546.
47. Martins M, Rosa A, Guedes LC, Fonseca BV, Gotovac K, Violante S, Mestre T, Coelho M, Rosa MM, Martin ER, Vance JM, Outeiro TF, Wang L, Borovecki F, Ferreira JJ, Oliveira SA: **Convergence of miRNA expression profiling, alpha-synuclein interacton and GWAS in Parkinson's disease.** *PLoS One* 2011, **6**:e25443.
48. Perkins DO, Jeffries CD, Jarskog LF, Thomson JM, Woods K, Newman MA, Parker JS, Jin J, Hammond SM: **microRNA expression in the prefrontal cortex of individuals with schizophrenia and schizoaffective disorder.** *Genome Biol* 2007, **8**:R27.
49. Unden A, Tatemoto K, Mutt V, Bartfai T: **Neuropeptide Y receptor in the rat brain.** *Eur J Biochem* 1984, **145**:525-530.
50. Bocchio-Chiavetto L, Maffioletti E, Bettinsoli P, Giovannini C, Bignotti S, Tardito D, Corrada D, Milanesi L, Gennarelli M: **Blood microRNA changes in depressed patients during antidepressant treatment.** *EurNeuropsychopharmacol* 2013, **23**:602-611.
51. Keller A, Leidinger P, Lange J, Borries A, Schroers H, Scheffler M, Lenhof HP, Ruprecht K, Meese E: **Multiple sclerosis: microRNA expression profiles accurately differentiate patients with relapsing-remitting disease from healthy controls.** *PLoS One* 2009, **4**:e7440.
52. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM: **Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease.** *Neurology* 1984, **34**:939-944.
53. **FASTX-Toolkit.** [http://hannonlab.cshl.edu/fastx_toolkit/].
54. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N: **miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades.** *Nucleic Acids Res* 2012, **40**:37-52.
55. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *JMolecular Biol* 1990, **215**:403-410.
56. Kozomara A, Griffiths-Jones S: **miRBase: integrating microRNA annotation and deep-sequencing data.** *Nucleic Acids Res* 2011, **39**:D152-157.
57. Lestrade L, Weber MJ: **snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs.** *Nucleic Acids Res* 2006, **34**:D158-162.
58. **Ensembl.** [ftp://ftp.ensembl.org/pub/release-67/fasta/homo_sapiens/ncrna/].
59. Bu D, Yu K, Sun S, Xie C, Skogerbo G, Miao R, Xiao H, Liao Q, Luo H, Zhao G, Zhao H, Liu Z, Liu C, Chen R, Zhao Y: **NONCODE v3.0: integrative annotation of long noncoding RNAs.** *Nucleic Acids Res* 2012, **40**:D210-215.
60. **GEO Database.** [http://www.ncbi.nlm.nih.gov/geo/].
61. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I: **Controlling the false discovery rate in behavior genetics research.** *Behav Brain Res* 2001, **125**:279-284.
62. Hochberg Y: **A sharper bonferroni procedure for multiple tests of significance.** *Biometrica* 1988, **75**:185-193.
63. Team R: **R: A Languageand Environment for Statistical Computing.** *Vienna: R Foundation for Statistical Computing* 2008.
64. **miRDB.** [http://mirdb.org/miRDB/].
65. Wang X: **miRDB: a microRNA target prediction and functional annotation database with a wiki interface.** *RNA* 2008, **14**:1012-1017.
66. Wang X, El Naqa IM: **Prediction of both conserved and nonconserved microRNA targets in animals.** *Bioinformatics* 2008, **24**:325-332.
67. **TargetScan.** [http://www.targetscan.org/vert_50/seedmatch.html].
68. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.** *Cell* 2005, **120**:15-20.
69. **GeneTrail.** [http://genetrail.bioinf.uni-sb.de/].
70. Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, Muller R, Meese E, Lenhof HP: **GeneTrail–advanced gene set enrichment analysis.** *Nucleic Acids Res* 2007, **35**:W186-192.
71. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.** *J Roy Stat SocB* 1995, **57**:289-300.
72. Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W, Cui Q: **An analysis of human microRNA and disease associations.** *PLoS One* 2008, **3**:e3420.

doi:10.1186/gb-2013-14-7-r78
Cite this article as: Leidinger *et al.*: A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biology* 2013 **14**:R78.

Papers in Press. Published June 30, 2014 as doi:10.1373/clinchem.2014.224238
The latest version is at http://hwmaint.clinchem.org/cgi/doi/10.1373/clinchem.2014.224238

Clinical Chemistry 60:9
000–000 (2014)

Influence of the Confounding Factors Age and Sex on MicroRNA Profiles from Peripheral Blood

Benjamin Meder,[1,3,6] Christina Backes,[2] Jan Haas,[1,3] Petra Leidinger,[2] Cord Stähler,[4] Thomas Großmann,[5] Britta Vogel,[1] Karen Frese,[1] Evangelos Giannitsis,[1] Hugo A. Katus,[1,3,6] Eckart Meese,[2] and Andreas Keller[5*]

BACKGROUND: MicroRNAs (miRNAs) measured from blood samples are promising minimally invasive biomarker candidates that have been extensively studied in several case-control studies. However, the influence of age and sex as confounding variables remains largely unknown.

METHODS: We systematically explored the impact of age and sex on miRNAs in a cohort of 109 physiologically unaffected individuals whose blood was characterized by microarray technology (stage 1). We also investigated an independent cohort from a different institution consisting of 58 physiologically unaffected individuals having a similar mean age but with a smaller age distribution. These samples were measured by use of high-throughput sequencing (stage 2).

RESULTS: We detected 318 miRNAs that were significantly correlated with age in stage 1 and, after adjustment for multiple testing of 35 miRNAs, remained statistically significant. Regarding sex, 144 miRNAs showed significant dysregulation. Here, no miRNA remained significant after adjustment for multiple testing. In the high-throughput datasets of stage 2, we generally observed a smaller number of significant associations, mainly as an effect of the smaller cohort size and age distribution. Nevertheless, we found 7 miRNAs that were correlated with age, of which 5 were concordant with stage 1.

CONCLUSIONS: The age distribution of individuals recruited for case-control studies needs to be carefully considered, whereas sex may be less confounding. To support the translation of miRNAs into clinical application, we offer a web-based application (http://www.ccb.uni-saarland.de/mirnacon) to test individual miRNAs or miRNA signatures for their likelihood of being influenced.

The potential of microRNAs (miRNAs)[7] as biomarkers on the basis of tissue or body fluids is increasingly recognized. Since their discovery, miRNA profiles from serum, plasma, or blood cells have been generated and statistically evaluated for a multitude of human pathogenic processes, including almost all cancer entities but also many noncancer diseases, such as multiple sclerosis, acute myocardial infarction, Alzheimer disease, and chronic obstructive pulmonary disease *(1–11)*.

The majority of the existing biomarker studies have been carried out by use of case-control designs. One would expect that matching of both groups for confounding factors in these studies was a prerequisite. However, biomaterials from existing retrospective cohorts often do not meet the high requirements for RNA-based molecular analysis, and the buildup of adequately large matched disease and control cohorts can be problematic. Especially in diseases effecting elderly persons, it can be highly challenging to recruit suitable healthy control cohorts of the same age distribution. Consequently, many published studies fail to match the 2 most basic confounders, age and sex.

The influence of these fundamental confounding variables, age and sex, on miRNA profiles from bodily fluids has not been fully explored. However, various miRNAs are known to exert key roles in aging, and other miRNAs are encoded on sex chromosomes *(12)*, which already suggests that a relevant portion of human miRNA profiles will depend on the age and sex distribution of samples. In our analysis, we systematically investigated the influence of age and sex on miRNA profiles in a large cohort of physiologically unaffected individuals. We detected a statistically signifi-

[1] Department of Internal Medicine III, University of Heidelberg, Heidelberg, Germany; [2] Department of Human Genetics, Saarland University, Homburg, Germany; [3] DZHK (German Centre for Cardiovascular Research); [4] Siemens AG, Strategy Division; [5] Clinical Bioinformatics, Saarland University, Saarbrucken, Germany; [6] Klaus Tschira Institute for Integrative Computational Cardiology, Heidelberg, Germany.

* Address correspondence to this author at: Clinical Bioinformatics, Saarland University, Building 60 Homburg, Saarbrucken, Germany. Fax +49-173-1484638; e-mail ack@bioinf.uni-sb.de.

Received March 19, 2014; accepted May 29, 2014.
Previously published online at DOI: 10.1373/clinchem.2014.224238

[7] Nonstandard abbreviations: miRNA, microRNA; NGS, next-generation sequencing.

cant number of miRNAs that were influenced by age or sex of the respective individuals. We validated our initial findings using an independent cohort of 58 samples from physiologically unaffected controls by applying high-throughput sequencing.

Materials and Methods

STUDY DESIGN AND BLOOD SAMPLE COLLECTION

In this study, we included 109 physiologically unaffected individuals (stage 1), whose blood has been partially measured as part of the human bloodborne miRNome project *(1)*. This collection contains whole miRNome-wide measurements according to Sanger miRBase *(13, 14)* version 14, for 454 samples. The remaining control samples have been included in the second version of the human bloodborne miRNome project, containing a total of 1050 samples measured by the same microarray technology. In a second cohort (stage 2), we measured an additional 58 physiologically unaffected controls using an independent technology, high-throughput sequencing on Illumina HiSeq 2000.

All blood samples were collected by use of a standard operating procedure in PAXgene Blood RNA tubes (Becton Dickinson). All blood donors participating in this study provided written informed consent, and the local ethics committee approved the study.

miRNA EXTRACTION AND MICROARRAY SCREENING (STAGE 1)

We carried out miRNA extraction and microarray measurement as previously described *(1)*. In brief, 2.5–5 mL of venous blood was collected in PAXgene Blood RNA tubes. Total RNA, including small RNAs, was extracted and stored at −70 °C. All samples of stage 1 were screened by use of the Geniom RT Analyzer system (Febit Biomed) with the Geniom Biochip miRNA *Homo sapiens* covering 848 common miRNAs in versions 12–14 of the Sanger miRBase. Each miRNA was represented by at least 7 replicated features on the microarray; for each miRNA, the median signal intensity was calculated.

HIGH-THROUGHPUT SEQUENCING (STAGE 2)

For library preparation, we used 70 ng total RNA per sample, as determined with a RNA 6000 Pico Chip on the Bioanalyzer 2100 (Agilent). For preparation, we used the TruSeq Small RNA Sample Prep Kit (Illumina). Ready prepped libraries were measured with the Bioanalyzer by use of the DNA 1000 Chip and subsequently pooled in batches of 6 samples in equal amounts. Sequencing libraries were then clustered with a final concentration of 9 pmol in 1 lane each of a single-read flow cell by use of the cBot instrument (Illumina). Sequencing of 50 cycles was performed on a HiSeq 2000 (Illumina). Demultiplexing of the raw sequencing data and generation of the fastq files was done with CASAVA version 1.8.2.

BIOSTATISTICAL ANALYSIS

To account for variations between different microarrays, we applied standard quantile normalization to the raw expression intensities. All downstream analyses were carried out on the normalized intensity values. We performed all bioinformatics calculations using the free and publicly available statistical language R (http://www.r-project.org/), if not mentioned otherwise.

For next-generation sequencing (NGS) data analysis, we preprocessed the raw Illumina reads by cutting the 3′ adapter sequence by use of the program fastx_clipper from the FASTX-Toolkit. After that, we used the miRDeep2 pipeline using the standard parameters for retrieving the miRBase counts for release version 20 and prediction of novel miRNAs. We applied HG20 as the reference genome for this analysis. Because the microarray experiments were measured by use of previous miRBase versions, whereas high-throughput sequencing results relied on the most recent version of the miRBase, we used the sequence of miRNAs as unique identifiers to match between the different miRBase versions.

To assess significance values for quantifying differences between males and females, we applied the parametric unpaired two-tailed *t*-test after verifying that data were approximately normally distributed by use of the Shapiro–Wilk test. To compute significance values for correlation coefficients, we applied test statistics on the basis of Pearson's product–moment correlation coefficient. Cluster analysis was carried out by use of R. Hierarchical clustering on the basis of the Bioconductor package Heatplus was applied to calculate heat maps and dendrograms. The hclust and cuttree functions were used to extract clusters out of the dendrogram. By use of this clustering information, contingency tables were generated, and Fisher test was applied to calculate significance values.

WEB SERVICE

To make the calculations available for other researchers, we implemented a web-based tool that is freely available for noncommercial usage (http://www.ccb.uni-saarland.de/mirnacon). Our tool receives as input either the IDs of a set of miRNAs along with the respective miRBase version or a set of miRNA sequences. These sequences are then matched to the most recent version 20 of miRBase and mapped to our experimental data. For further input parameters, the user can select the significance threshold (standard value 0.05) and a lower boundary for significance values (standard value 0.001). This boundary is just applied for the graphical representation of the results, i.e.,

miRNAs with significance below this value are presented at the cutoff. Because the analysis itself may be biased by different age distributions in the test set, users can input the mean age and SD of the study population used for their study. Our tool then automatically extracts a subcohort of our samples that best matches the user's age distribution.

For output, our tool generates a scatterplot showing the background distribution of miRNAs in our study (gray dots); the miRNAs uploaded by the user (red dots); and whether the miRNAs are not influenced by age and sex (green area of the scatter plot), influenced by either age or sex (gray area of the scatter plot), or influenced by age and sex (red area of the scatter plot). Furthermore, we generate a tabular output containing the uploaded ID, the respective sequence, the ID in the most recent version 20 of the Sanger miRBase, and whether this miRNA is influenced by age and/or sex.

Results

We included a total of 109 controls in the microarray-based miRNA assessment of stage 1. The miRNA biomarkers were profiled by use of miRNA microarrays covering 848 miRNAs across versions 12–14 of the Sanger miRBase. The cohort contained samples from 65 women and 44 men with a mean age of 57.3 (SD 25.5) years, range 19–105 years.

Likewise, 58 independently collected and measured controls were processed by use of high-throughput sequencing and mapped to miRBase version 20. The cohort contained 12 women and 46 men. Considering age, the individuals in the second cohort had a similar mean age as the stage 1 cohort (58.3 years); however, this cohort had a much smaller age variance (SD 8.6 years, range 44–75 years). Age distribution metrics for both cohorts are provided in Supplemental Table 1, which accompanies the online version of this article at http://www.clinchem.org/content/vol60/issue9.

IMPACT OF AGE ON miRNA PROFILES

First, we calculated the correlation of each miRNA to the age of the individuals, providing us with 848 different correlation coefficients. Additionally, we calculated significance values for the respective correlations and considered unadjusted as well as adjusted *P* values (Bonferroni adjustment). Of the 848 miRNAs, 318 were significantly correlated with age (raw *P* value of <0.05). Notably, around one-third (107) were negatively correlated with age, whereas two-thirds (211) were positively correlated with age. This shift in the distribution toward positive correlation can be seen in Fig. 1 (right side of the histogram). In this figure, all miRNAs with a correlation coefficient >0.5 are given. Notably, even after adjustment for multiple testing by use of the conservative Bonferroni approach, 35 miRNAs remained significant (adjusted *P* value <0.05). The 35 miRNAs along with the raw significance values and the correlation coefficients are detailed in Table 1.

Regarding the stage 2 cohort with a more narrow age distribution but comparable mean age, we calculated a substantially smaller amount of significantly associated miRNAs. Whereas the stage 1 cohort quartiles are 34, 57.5, and 71 years, the validation cohort quartiles are 53, 56, and 65 years. Despite these differences, we observed 7 of the originally detected miRNAs that were expressed and significantly correlated with age in the second cohort. Of these 7 miRNAs, 5 showed the same direction of dysregulation as in stage 1. The miRNAs significantly influenced by age in both stages include hsa-miR-1284, hsa-miR-93–3p, hsa-miR-1262, hsa-miR-34a-5p, and hsa-miR-145–5p, meaning that these miRNAs may be most strongly affected by aging. The markers are summarized in Table 2 together with the respective correlation values. Fig. 2 shows scatter plots for the most significantly downregulated miRNA, namely hsa-miR-106a, and the miRNA with the best fit between microarrays and NGS, hsa-miR-93–3p. Each image shows a significant positive or negative correlation of miRNA expression with age of individuals, respectively.

Next, we investigated the dependency of age–miRNA correlations on the age distribution. To this end, for the 23 most significant correlations from Table 1, we calculated the correlation for subcohorts with approximated mean age of 40, 50, 60, and 70 years. As the spider diagram in Fig. 3 shows, the significance for these miRNAs substantially changed with different subcohorts. The most significant results were detected for the subcohort with a mean age of 60 years, where 10 miRNAs were significantly correlated (inside of the gray-shaded area of the spider diagram). By contrast, for the 40- and 70-year subcohorts, just 2 of the miRNAs remained significant.

IMPACT OF SEX ON miRNA PROFILES

We also calculated significance values for the sex of all individuals. In this analysis, we detected much lower numbers of significantly associated miRNAs. Although 318 significant markers were found in stage 1 for age, we found only 144 miRNAs significant before adjustment for multiple testing in case of sex. Although this number is much higher than the expected number of significant miRNAs at an α level of 0.05, no miRNA remained significant after Bonferroni adjustment (*P* value <0.05 after adjustment; smallest *P* value after adjustment: 0.09).

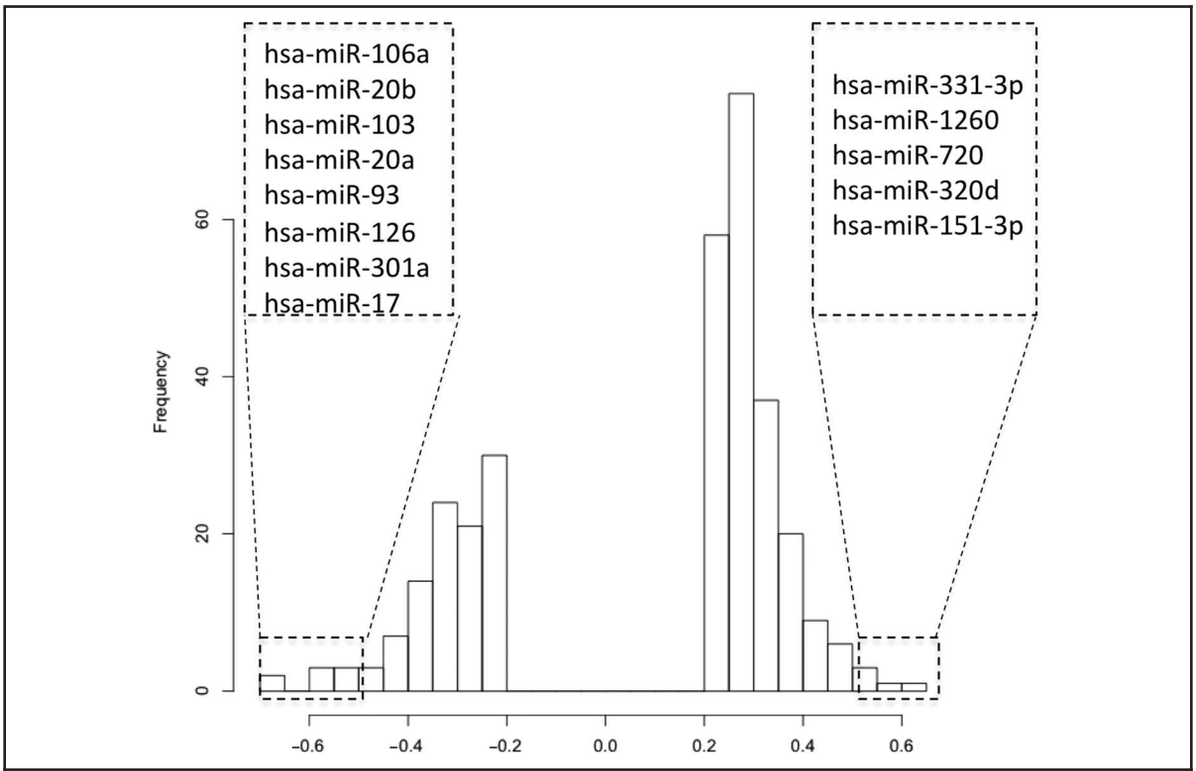


Fig. 1. Histogram of positive and negative correlated miRNA.
The histogram shows the left and right tail of the correlation of miRNAs versus the age of patients and reveals a clear shift for positive correlation.

For stage 2, we calculated significantly different expression levels depending on sex for 6 of the differentially expressed miRNAs, with only 3 miRNAs (hsa-miR-219a-1-3p, hsa-miR-548c-3p, and hsa-miR-130a-3p) being concordant in both cohorts, demonstrating that differences in miRNA were mainly due to the confounding factor age, but less to sex.

miRNA PATTERNS CLUSTER INDIVIDUALS REGARDING AGE AND SEX

In addition to the above correlation analysis, we carried out unsupervised and supervised cluster approaches. First, we extracted the 10 most variable miRNAs and calculated whether these miRNAs separate the individuals with respect to sex (male vs female) or age (young vs old, cutoff mean age). With respect to sex, we reached a significance value of 0.01 after separating the data into 2 clusters. With respect to age, we found an even more significant clustering with a P value of 0.0067, confirming our initial findings that the age of individuals has a larger impact on miRNA than their sex (see online Supplemental Fig. 1).

Additionally, we applied a supervised clustering approach. Here, we included the most significantly correlated miRNAs (raw P value <0.05) for the clustering, limiting the analysis, however, to the 50 miRNAs with highest data variance. As expected, the significance of the clustering substantially improved. With respect to age, the significance value went down to 0.0004; for sex, down to 0.0006 (see online Supplemental Figs. 2 and 3).

REPRESENTATION OF RESULTS AND WEB-BASED ANALYSIS

Because our results indicated a moderate influence of sex and a substantial influence of age on bloodborne miRNA profiles, we implemented a web-based solution for providing other researchers with easy access to the respective data and the ability to visualize the degree of influence of age and sex on candidate miRNA biomarkers of their studies. As input, users can choose between the miRNA sequence or miRBase miRNA IDs. For the latter, all versions starting from miRBase 16 are implemented. Furthermore, the user can specify a significance threshold as well as parameters for improved graphical representation. As demonstrated above, the overall age distribution has the highest impact on miRNAs. Thus, users can also specify the average age and SD of their cohort. Our algorithm, which relies on

Table 1.
Significantly age-correlated miRNAs (stage 1, adjusted *P* value <0.05).

miRNA	Correlation	Raw *P* value
hsa-miR-106a	−0.680	4.54E−14
hsa-miR-20b	−0.657	6.43E−13
hsa-miR-151-3p	0.628	1.21E−11
hsa-miR-103	−0.592	3.41E−10
hsa-miR-320d	0.578	1.06E−09
hsa-miR-20a	−0.569	2.19E−09
hsa-miR-93	−0.557	5.56E−09
hsa-miR-720	0.548	1.07E−08
hsa-miR-126	−0.533	3.13E−08
hsa-miR-301a	−0.530	4.00E−08
hsa-miR-1260	0.526	5.14E−08
hsa-miR-17	−0.524	5.83E−08
hsa-miR-331-3p	0.505	2.12E−07
hsa-miR-30c	0.491	5.18E−07
hsa-miR-590-5p	−0.486	6.97E−07
hsa-miR-320c	0.485	7.22E−07
hsa-miR-30d	0.477	1.16E−06
hsa-miR-107	−0.471	1.68E−06
hsa-miR-24	−0.470	1.79E−06
hsa-miR-1262	0.470	1.81E−06
hsa-miR-526b*	0.456	3.94E−06
hsa-miR-664	0.452	4.75E−06
hsa-miR-548i	0.444	7.40E−06
hsa-miR-197	0.433	1.28E−05
hsa-miR-892a	0.433	1.29E−05
hsa-miR-30a	0.429	1.58E−05
hsa-miR-20a*	−0.427	1.79E−05
hsa-miR-374a	−0.425	1.94E−05
hsa-miR-29c*	0.425	1.95E−05
hsa-miR-15b	−0.423	2.15E−05
hsa-miR-144	−0.421	2.44E−05
hsa-miR-520c-3p	0.420	2.50E−05
hsa-miR-96	−0.413	3.45E−05
hsa-miR-339-5p	0.404	5.38E−05
hsa-miR-106b	−0.403	5.71E−05

the results of stage 1, then searches dynamically for a subcohort that matches the requested parameters of the user.

In the tabular output, the miRNA-ID is shown, as well as its mapping to the most recent ID in miRBase version 20 and the sequence of the miRNA. In the fourth and fifth column of the output table, the potential influence of age and sex are documented. We tested our tool on a hypothetical disease signature with 9 markers (chosen from studies on different diseases). Of these, 2 (miR-144 and miR-20b) are potentially influenced by age, whereas 7 (miR-1, miR-127–5p, miR-1270, miR-1271, miR-1272, miR-144*, and miR-20a*) are not influenced significantly. Fig. 4 presents the graphical output of the tool. The miRNAs in the upper right quadrant (green) are not significantly influenced. The miRNAs in the lower right quadrant are potentially influenced by age and thus highlighted by red points. The red-shaded lower right quadrant would contain miRNAs that are influenced by age and sex. All miRNAs uploaded by the user are shown as colored dots, and the distribution of the miRNAs from stage 1 of this study are represented by gray dots. The tool is freely available for noncommercial applications at http://www.ccb.uni-saarland.de/mirnacon/.

Table 2. Overlap between stage 1 and stage 2 with respect to age.

miRNA	Array correlation	NGS correlation	*P* value
hsa-miR-1284	0.258594528	0.386691307	0.002713735
hsa-miR-23a-5p[a]	0.238861262	−0.370263342	0.004224409
hsa-miR-652-3p[a]	−0.204746216	0.367868659	0.004497567
hsa-miR-93-3p	0.390291436	0.320291239	0.014241407
hsa-miR-1262	0.469537126	0.301844226	0.021293944
hsa-miR-34a-5p	0.269430053	0.279969853	0.033292267
hsa-miR-145-5p	0.284076368	0.272182316	0.038738333

[a] Discordant between microarray and NGS experiments.

Discussion

miRNAs are increasingly recognized as biomarkers for various diseases, including almost all cancer entities and metabolic, neurological, and cardiovascular disorders. We investigated here the role of the confounding variables age and sex on the miRNA profiles observed in whole peripheral blood.

Despite the euphoria about the potential clinical application of miRNAs in disease detection and estimation of prognosis, many miRNA biomarkers show discrepant results in independent investigations of the same disease. In addition to technical challenges such as sample handling, RNA processing, and storage, as well as differences in the underlying measurement technology such as microarrays or high-throughput sequencing, many obstacles remain that could additionally affect this observation. In recent publications, some confounding factors for miRNAs from serum or

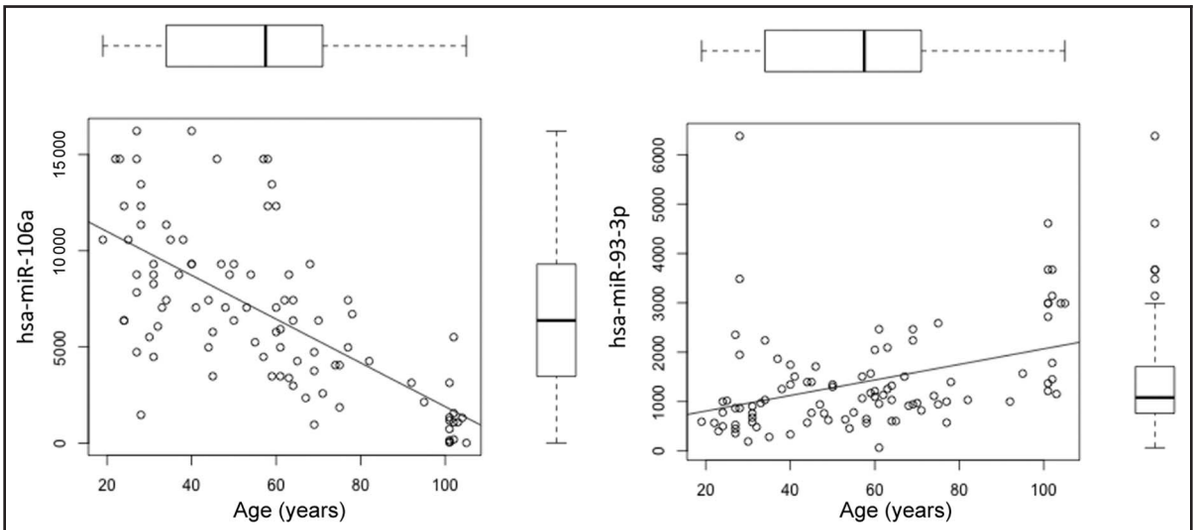


Fig. 2. Example scatter plots for positive and negative correlation of 2 miRNAs with age.
Distribution of miRNA and age are presented above and on the right of the plots, respectively.

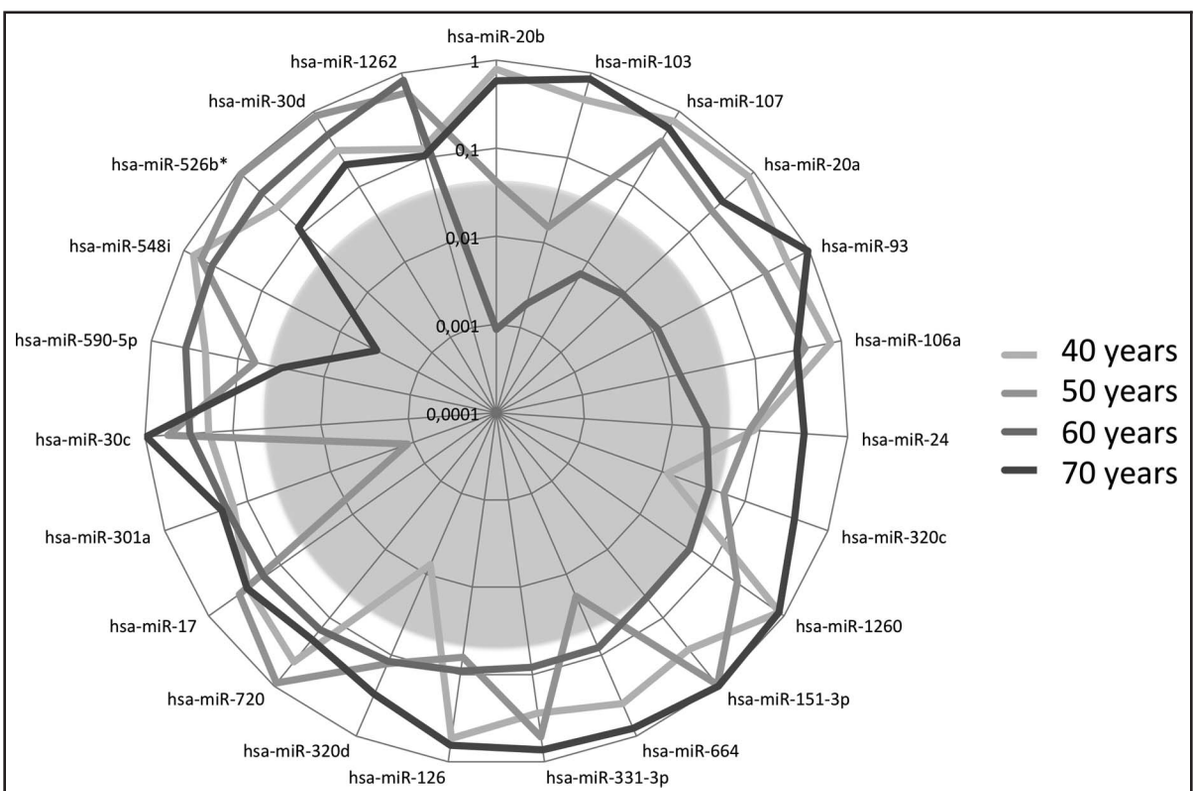


Fig. 3. Spider diagram showing the variance of significance depending on mean age.
On a logarithmic scale, the diagram presents the significance of correlation with the age for 4 age groups. The highest significance (closest proximity to the center) was detected for the age cohort of 60-year-old individuals.

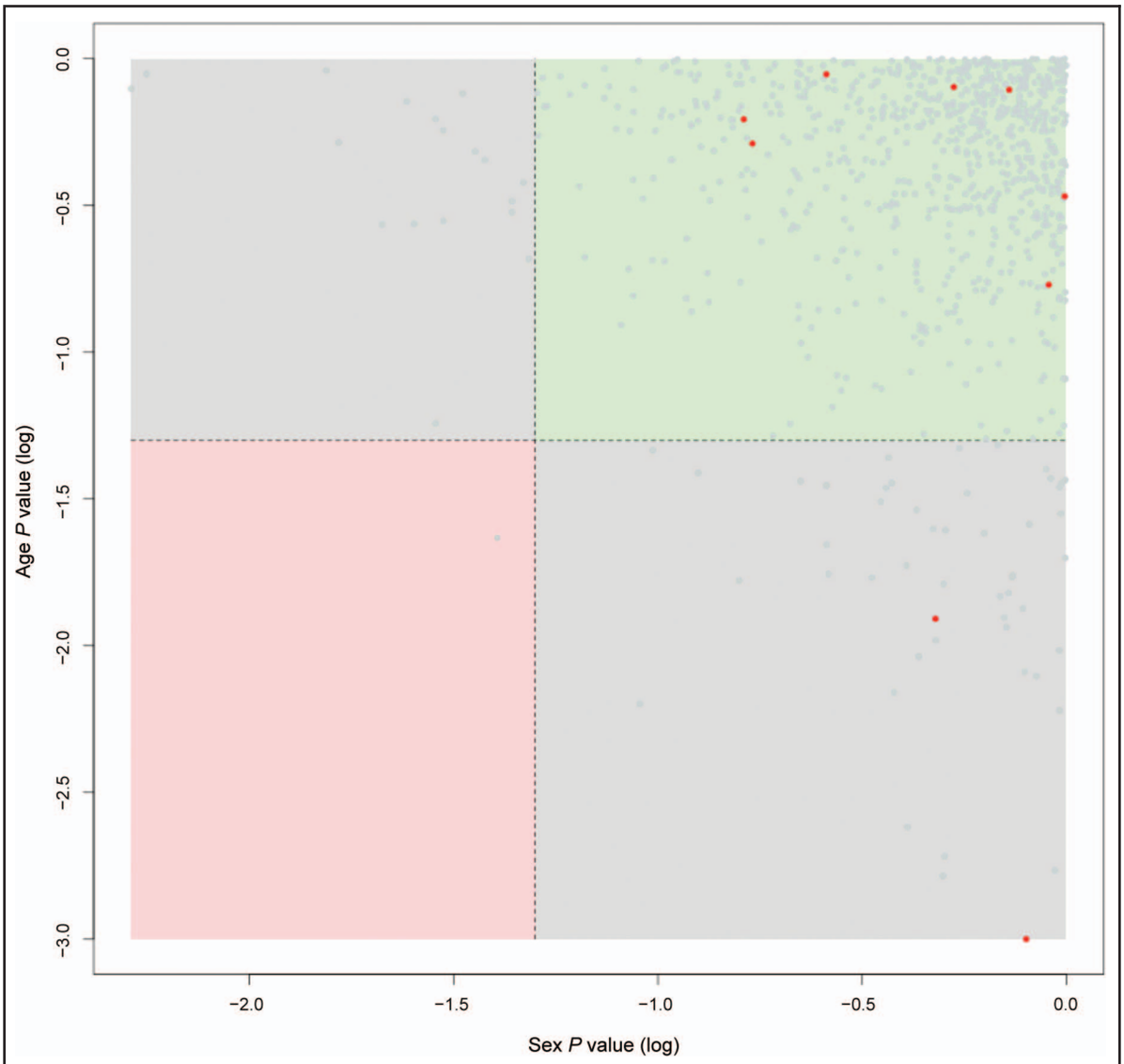


Fig. 4. Graphical output of miRNACon, the web service for dynamic calculation of potential confounding variables.
x axis, *P* value for sex; *y* axis, *P* value for age. Gray dots belong to background miRNAs, green dots to user-specified miRNAs that seem to be not affected by sex and age, and red dots to affected miRNAs.

plasma were found to include patient treatment *(15)* and comorbidities. However, the role of the fundamental confounders age and sex are only partially understood. We show here that age potentially has a higher influence on the expression of miRNAs than sex.

Knowledge about the confounding factors and their influence on certain miRNAs has considerable consequences. First, a well-designed study with adequately sized case and control cohorts should be a prerequisite. However, often it is very challenging to have suitable control cohorts of healthy individuals matching the age distribution of cases. This is obviously most important when studying diseases of the elderly, such as neurodegenerative diseases or chronic heart failure. Another way to circumvent potential bias due to non–equally distributed variables would be the implementation of the findings presented here in statistical approaches to select appropriate subcohorts in silico, to use, e.g., the ages as additional input variables for machine-learning methods or to dynamically build models. A more straightforward approach would be to

verify miRNAs by use of the provided online application and exclude strongly influenced candidates from the respective signatures.

Although we found age to be a strong confounder, all analyses revealed only a limited influence of sex on miRNA patterns. Nevertheless, we investigated the most differentially expressed miRNAs between males and females in more detail to support the hypothesis of sex-dependent miRNA regulation by comparing them to those experimentally affected by estrogen. We investigated blood cells, and the study by Maillot et al. *(16)* relied on breast cell cultures exposed to estrogen. They found 23 miRNAs that are significantly downregulated after estrogen signaling has been induced and thus depend indirectly on sex. Of these 23 miRNAs, 18 are expressed higher in males (78.3%) in our cohorts. Although these miRNAs were not significantly differentially expressed after adjustment for multiple testing, these results hint at a limited sex-dependent miRNA signature in blood cells. Here, larger cohorts may reveal whether the differences are actually significant.

A challenge in generalizing our findings is technical variation between different platforms. Most frequently, array technology and NGS are applied to screen for mRNA or miRNA biomarkers, and quantitative reverse-transcription PCR is applied to validate the results. In 2010, Git et al. presented a technological evaluation *(17)*, concluding that the actual overlap between the platforms was low. As a consequence, our web service currently incorporates only the microarray data of the larger cohort with the higher age variation. The extension to NGS is planned for one of the succeeding versions, as well as to provide similar functionality for serum and plasma.

Another interesting observation of our study is that age-related miRNAs may also have biological meaning. For several miRNAs, animal studies could provide evidence for their role in senescence or antiaging. For instance, hsa-miRNA-34a was recently recognized as positively correlating with age, suppressing important downstream targets and leading to telomere shortening and cardiomyocyte dysfunction/apoptosis *(12)*. In line with these results, Li et al. describe miR-34a as being upregulated in tissue and blood of older mice *(18)*. This makes miR-34a a good positive control for our study. Indeed, miR-34a is significantly correlated with age in our results (raw *P* value 0.0086). In contrast, Li et al. *(18)* describe miR-196a to be independent of age. Concordant with these results, this miRNA is not significantly regulated with age in our study (raw *P* value 0.08). Thus, miR-196a represents a valuable negative control for our study. The miRNAs correlating to age identified here might also harbor functional properties that are important for age research and represent potential pharmaceutical targets. Hence, our repertoire of miRNAs includes appealing targets for further functional workup.

In summary, our study provides evidence that especially age is an important confounding variable for miRNA biomarker profiles in human blood samples, whereas sex shows just a limited effect on bloodborne miRNA patterns. We make the results of this study available to researchers through an easy-to-use web-based tool. Clearly, there should be a focus on additional common confounders, such as smoking, kidney and liver function, and others, to systematically dissect their influence on miRNAs from body fluids and tissues. With these precautions, miRNAs have the potential to proceed into clinical application for many diseases that are currently difficult to diagnose.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: C. Stähler, Siemens AG–Healthcare Sector.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: B. Meder, European Union (FP7 BestAgeing), Bundesministerium für Bildung und Forschung (BMBF): German Center for Cardiovascular Research (DZHK); C. Backes, European Union (FP7 BestAgeing); J. Haas, European Union (FP7 BestAgeing), Bundesministerium für Bildung und Forschung (BMBF): German Center for Cardiovascular Research (DZHK); P. Leidinger, European Union (FP7 BestAgeing); T. Groβmann, European Union (FP7 BestAgeing); B. Vogel, European Union (FP7 BestAgeing); K. Frese, European Union (FP7 BestAgeing); E. Giannitsis, European Union (FP7 BestAgeing); H.A. Katus, European Union (FP7 BestAgeing), Bundesministerium für Bildung und Forschung (BMBF): German Center for Cardiovascular Research (DZHK); E. Meese, European Union (FP7 BestAgeing); A. Keller, European Union (FP7 BestAgeing).
Expert Testimony: None declared.
Patents: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

Acknowledgments: We thank Rouven Nietsch, Christina Scheiner, and Sabine Marquart for excellent technical assistance.

References

1. Keller A, Leidinger P, Bauer A, Elsharawy A, Haas J, Backes C, et al. Toward the blood-borne miRNome of human diseases. Nat Methods 2011; 8:841–3.
2. Keller A, Leidinger P, Lange J, Borries A, Schroers H, Scheffler M, et al. Multiple sclerosis: microRNA expression profiles accurately differentiate patients with relapsing-remitting disease from healthy controls. PLoS One 2009;4:e7440.
3. Keller A, Leidinger P, Steinmeyer F, Stahler C, Franke A, Hemmrich-Stanisak G, et al. Comprehensive analysis of microRNA profiles in multiple sclerosis including next-generation sequencing. Mult Scler 2013.
4. Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, et al. A blood based 12-miRNA signature of Alzheimer disease patients. Genome Biol 2013;14:R78.
5. Leidinger P, Keller A, Borries A, Huwer H, Rohling M, Huebers J, et al. Specific peripheral miRNA profiles for distinguishing lung cancer from COPD. Lung Cancer 2011;74:41–7.
6. Leidinger P, Keller A, Borries A, Reichrath J, Rass K, Jager SU, et al. High-throughput miRNA profiling of human melanoma blood samples. BMC Cancer 2010;10:262.
7. Meder B, Keller A, Vogel B, Haas J, Sedaghat-Hamedani F, Kayvanpour E, et al. MicroRNA signatures in total peripheral blood as novel biomarkers for acute myocardial infarction. Basic Res Cardiol 2011;106:13–23.
8. Vogel B, Keller A, Frese KS, Leidinger P, Sedaghat-Hamedani F, Kayvanpour E, et al. Multivariate miRNA signatures as biomarkers for non-ischaemic systolic heart failure. Eur Heart J 2013;34:2812–23.
9. Fenoglio C, Ridolfi E, Cantoni C, De Riz M, Bonsi R, Serpente M, et al. Decreased circulating miRNA levels in patients with primary progressive multiple sclerosis. Mult Scler 2013;19:1938–42.
10. Sayed AS, Xia K, Yang TL, Peng J. Circulating microRNAs: a potential role in diagnosis and prognosis of acute myocardial infarction. Dis Markers 2013;35:561–6.
11. Kiko T, Nakagawa K, Tsuduki T, Furukawa K, Arai H, Miyazawa T. MicroRNAs in plasma and cerebrospinal fluid as potential markers for Alzheimer's disease. J Alzheimers Dis 2014;39:253–9.
12. Boon RA, Iekushi K, Lechner S, Seeger T, Fischer A, Heydt S, et al. MicroRNA-34a regulates cardiac ageing and function. Nature 2013;495:107–10.
13. Griffiths-Jones S. The microRNA registry. Nucleic Acids Res 2004;32:D109–11.
14. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. Mirbase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 2006;34:D140–4.
15. Boeckel JN, Thome CE, Leistner D, Zeiher AM, Fichtlscherer S, Dimmeler S. Heparin selectively affects the quantification of microRNAs in human blood samples. Clin Chem 2013;59:1125–7.
16. Maillot G, Lacroix-Triki M, Pierredon S, Gratadou L, Schmidt S, Benes V, et al. Widespread estrogen-dependent repression of microRNAs involved in breast tumor cell growth. Cancer Res 2009;69: 8332–40.
17. Git A, Dvinge H, Salmon-Divon M, Osborne M, Kutter C, Hadfield J, et al. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. RNA 2010; 16:991–1006.
18. Li X, Khanna A, Li N, Wang E. Circulatory miR34a as an RNAbased, noninvasive biomarker for brain aging. Aging 2011;3:985–1002.

Papers in Press. Published January 23, 2015 as doi:10.1373/clinchem.2014.232165
The latest version is at http://hwmaint.clinchem.org/cgi/doi/10.1373/clinchem.2014.232165

Clinical Chemistry 61:4
000–000 (2015)

MicroRNA In Vitro Diagnostics by Use of Immunoassay Analyzers

Andreas Kappel,[1†] Christina Backes,[2†] Yiwei Huang,[1] Sachli Zafari,[2] Petra Leidinger,[3] Benjamin Meder,[4] Herbert Schwarz,[6] Walter Gumbrecht,[1] Eckart Meese,[3] Cord Stähler,[5] and Andreas Keller[2*]

BACKGROUND: The implementation of new biomarkers into clinical practice is one of the most important areas in medical research. Besides their clinical impact, novel in vitro diagnostic markers promise to have a substantial effect on healthcare costs. Although numerous publications report the discovery of biomarkers, only a fraction of those markers are routinely used. One key challenge is a measurement system that is compatible with clinical workflows.

METHODS: We designed a new immunoassay for microRNA (miRNA) quantification. The assay combines streptavidin-linked microparticles, a biotinylated catcher oligonucleotide complementary to a single miRNA species, and finally, a monoclonal antibody to DNA/RNA heterohybrids labeled with acridinium ester. Importantly, our assay runs on standard immunoassay analyzers. After a technical validation of the assay, we evaluated the clinical performance on 4 Alzheimer disease miRNAs.

RESULTS: Our assay has an analytical specificity of 99.4% and is at the same time sensitive (concentrations in the range of 1 pmol/L miRNA can be reliably profiled). Because the novel approach did not require amplification steps, we obtained high reproducibility for up to 40 biological replicates. Importantly, our assay prototype exhibited a time to result of <3 h. With human blood samples, the assay was able to measure 4 miRNAs that can detect Alzheimer disease with a diagnostic accuracy of 82% and showed a Pearson correlation >0.994 with the gold standard qRT-PCR.

CONCLUSIONS: Our miRNA immunoassay allowed the measurement of miRNA signatures with sufficient analytical sensitivity and high specificity on commonly available laboratory equipment.

A substantial number of molecules, including DNA, RNA, microRNAs (miRNAs),[7] proteins, and methylated sites in the genome or metabolites, are reported as disease markers for various human pathologies, but only a small fraction will be translated to clinical routine use. One challenge is often poor diagnostic specificity or sensitivity, which can be overcome in some instances by combining biomarkers. The second major challenge is the reliable measurement of novel markers on platforms that are commonly used in clinical laboratories. Although current molecular methods used to measure DNA or miRNA biomarkers, such as quantitative RT-PCR (qRT-PCR) and next-generation sequencing (NGS), are available in selected clinical laboratories, they are rather expensive. Moreover, compatibility with clinical high-throughput workflows is challenging. The adaption of miRNA assays to platforms and technologies that would overcome those issues may foster their use.

Small noncoding RNAs such as miRNAs have important functions in nearly all cellular processes owing to their ability to regulate the expression of many protein-coding genes *(1)*. Associations have been described for a large fraction of the >2000 known miRNA diseases, which have been collected in databases such as the Human miRNA and Diseases Database *(2)*. Because of their ability to regulate target gene translation through either silencing or degradation of the target mRNA, miRNAs are involved in pathological processes such as cancer, neurological disorders, and heart disease *(3–5)*. Furthermore, complex miRNA signatures have been increasingly recognized as stable and powerful biomarkers for human pathologies *(6–14)*, making them ideal biomarker candidates. For the application of biomarkers in routine clinical settings, body fluids such as serum, urine, and cerebrospinal fluid represent preferable sources for biomarkers. Notably, blood cells contain a rich repertoire of disease-related markers.

Specific miRNA expression signatures for many human cancer and noncancer diseases have been identified

[1] Corporate Technology, [5] Strategy, Siemens AG, Erlangen, Germany; [2] Clinical Bioinformatics, Medical Faculty, [3] Department of Human Genetics, Saarland University, Saarbrücken, Germany; [4] Internal Medicine III, University Hospital Heidelberg, Heidelberg, Germany; [6] Siemens Healthcare Diagnostics Products GmbH, Marburg, Germany.
* Address correspondence to this author at: E-mail andreas.keller@ccb.uni-saarland.de.
[†] A. Kappel and C. Backes contributed equally to this study.
Received August 21, 2014; accepted December 18, 2014.
Previously published online at DOI: 10.1373/clinchem.2014.232165
[7] Nonstandard abbreviations: miRNA, microRNA; qRT-PCR, quantitative RT-PCR; NGS, next-generation sequencing; AD, Alzheimer disease; RLU, relative light unit.

(6, 15–18). Following biomarker discovery studies with limited sample cohorts, the suitability of blood-based miRNA expression signatures as early disease detection biomarkers is increasingly being investigated in larger validation studies, either in comparison to or in combination with known serum protein biomarkers *(18)*. In particular, the first tissue-based tests for measurement of specific miRNA expression signatures are already commercially available on qRT-PCR platforms (Rosetta Genomics). However, such tissue-based qRT-PCR tests have important downsides. First, they require substantial hands-on time. Second, qRT-PCR platforms are not used in many clinical laboratories, and tests performed on these platforms are usually less integrated into workflows than immunoassays. The lesser penetration of qRT-PCR and other molecular methods in the clinical laboratory compared with immunoassays is also reflected by the fact that molecular methods other than blood bank tests made up only 5% of all in vitro diagnostic sales in 2011, compared with a 25% market share of immunoassays (excluding blood bank tests) *(19)*. Third, tissue-based miRNA expression signatures require invasive sampling and are therefore more complicated to implement than blood-based tests in routine diagnostic applications. Given these downsides to tissue-based tests, blood-based miRNA diagnostics by use of immunoassay represents an interesting opportunity to introduce miRNA testing into clinical laboratories.

To promote the translation of miRNA tests further into routine use, and to address the challenges mentioned above, we developed a new miRNA measurement principle on the basis of an immunoassay format. Immunoassay platforms are already routinely used in clinical laboratories worldwide, and many immunological tests such as cardiac troponin are carried out on these commercial systems. After successfully setting up the assay format, we evaluated the assay performance on an Alzheimer disease (AD) miRNA panel *(11)*.

Methods

SAMPLE COLLECTION

We carried out miRNA measurements with PAXgene Blood RNA tubes (Preanalytix, Becton Dickinson). These tubes can be used to collect 2.5 mL blood from donors, according to the manufacturer's recommendations. We collected blood samples from 40 healthy volunteers. The Institutional Ethics Committee of the University Erlangen-Nuremberg approved the study. All donors met the relevant guidelines *(20, 21)* and tested negative for human immunodeficiency virus, hepatitis B virus, and hepatitis C virus.

miRNA EXTRACTION

The pellets from 2.5 mL blood collected in PAXGene tubes were obtained by 10-min centrifugation at 4500*g* according to the manufacturer's instructions, and the supernatant was removed immediately. The pellets were then resuspended in 4 mL RNase-free water by vortex-mixing and collected by 10-min centrifugation at 4500*g*. We then isolated total RNA including miRNA from the pellets with the miRNeasy Mini Kit (Qiagen) according to the manufacturer's recommendations. Isolated RNA was pooled, divided into aliquots, and stored at −80 °C until use.

miRNA qRT-PCR MEASUREMENT

We analyzed the miRNAs using stem-loop primers for qRT-PCR with TaqMan® probes on a Stratagene MX-3005P real-time cycler, essentially as previously described *(22)*. The master mix for real-time PCR, M-MuLV H Plus Reverse Transcriptase, dNTPs, and RNase inhibitor were obtained from Peqlab and stored at −20 °C. The synthetic miRNAs were obtained from Biomers.net. The sequences of primers are described in Supplemental Table 1, which accompanies the online version of this article at http://www.clinchem.org/content/vol61/issue4. We dissolved the synthetic miRNAs in Rnase-free water with 30 mU/μL RNase inhibitor to a concentration of 100 μmol/L and divided the miRNA solution to 5 μL/tube to be stored at −80 °C until use. The calibration curve was determined by qRT-PCR with the synthetic miRNA from 0.1 pmol/L to 1 nmol/L. The primer sets for qRT-PCR were obtained from Biomer.net. The sequences of primers for measuring the synthetic miRNAs are described in online Supplemental Table 2.

miRNA IMMUNOASSAY

The miRNA immunoassay presented in this study is a 2-step nucleic acid capture immunoassay adapted to the Advia Centaur® Immunoassay System (Siemens Healthcare Diagnostics). This immunoassay analyzer platform can be used to measure protein and small molecule analytes by respective assays with acridinium ester technology *(23)*. The components of our assay prototype consisted of the solid phase (containing streptavidin-linked microparticles), a biotinylated catcher oligonucleotide complementary to a single miRNA species (the biotinylated catchers are described in online Supplemental Table 3), and finally a monoclonal antibody to DNA/RNA heterohybrids *(24)* labeled with acridinium ester. The antibody, which was developed in the 1980s, specifically binds to DNA/RNA hybrids without any obvious bias toward a specific sequence *(24, 25)*.

In the assay, the purified miRNA from a blood sample is first hybridized to the biotinylated catcher oligonucleotide, generating perfectly matched DNA/RNA heterohybrids. In a second step, these biotinylated DNA/RNA heterohybrids are then incubated with and bound to the streptavidin-labeled solid phase. In the next step, the acridinium ester–labeled antibody to DNA/RNA

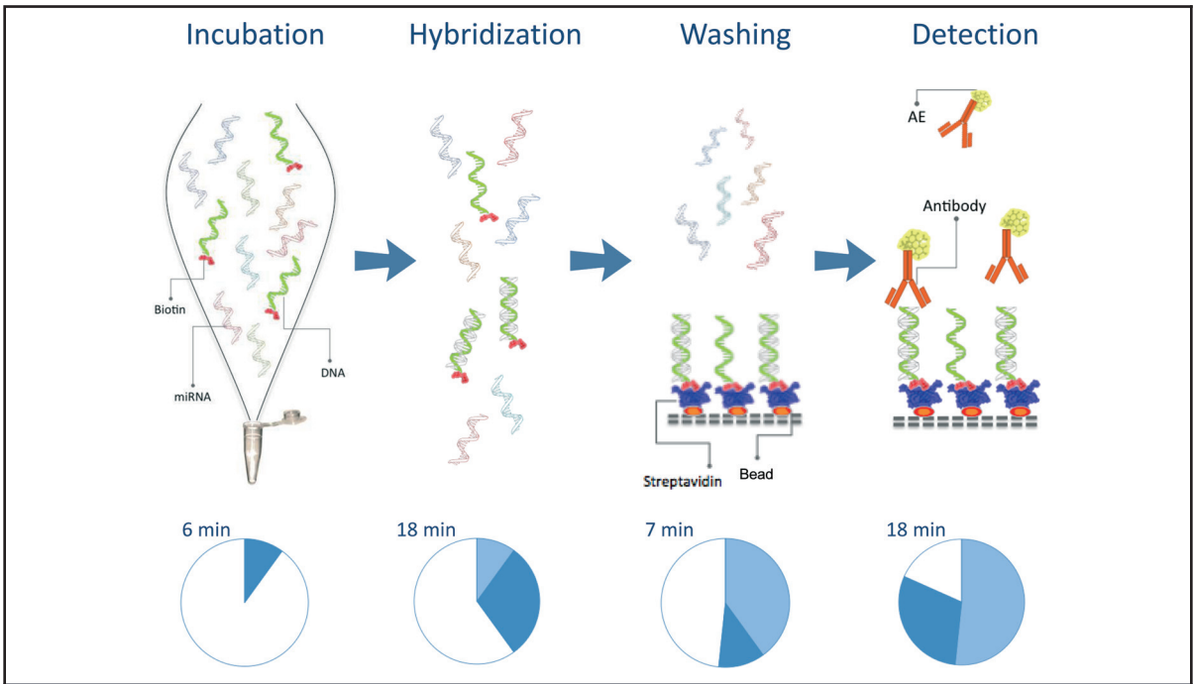


Fig. 1. The 4 fully automated main steps carried out on the immunoassay analyzer. AE, acridinium ester.

heterohybrids is added. This antibody can bind only to perfectly matched heterohybrids and does not bind to mismatched heterohybrids. The amount of antibody bound will therefore be proportional to the amount of perfectly matched heterohybrids present in the reaction, which again is proportional to the amount of that specific miRNA species present in the blood sample. Chemiluminescence is then triggered by addition of acid and base reagent *(26)*.

The following 9 automated steps were carried out with the Advia Centaur system. (*a*) Pipetting 75 μL samples in a cuvette. (*b*) Pipetting 75 μL reagent (20 mmol/L sodium phosphate, pH 7.2, 300 mmol/L NaCl, 0.1% Triton X-100, 0.5% bovine serum albumin, 0.02% sodium azide) containing biotinylated oligonucleotides (10 nmol/L) and incubating for 6 min at 37 °C. (*c*) Pipetting 150 μL solid phase and incubating for 18 min at 37 °C. (*d*) Separating solid phase from the mixture and removing the liquid phase. (*e*) Washing the cuvette with washing solution 1 and incubating for 6.75 min at 37 °C. (*f*) Pipetting 95 μL antibody reagent and incubating for 18 min at 37 °C. (*g*) Separating solid phase from the mixture and removing the liquid phase. (*h*) Washing the cuvette with wash solution 1. (*i*) Pipetting 300 μL reagent A (acid) and 300 μL reagent B (base) to generate a chemiluminescence signal. The workflow is presented schematically in Fig. 1. The concepts and information presented in this article represent research and are not commercially available.

CALIBRATION CURVES AND CALCULATION OF CONCENTRATIONS

We measured the calibration curve with synthetic miRNAs from a concentration of 1 pmol/L to 1 nmol/L on an the Advia Centaur system, carrying out a second-degree polynomial analysis to determine the equation of the relationship between relative light unit (RLU) counts and miRNA concentration. We then measured the biological samples on the same Advia Centaur system. The concentration of a certain miRNA of biological samples was calculated from the RLU counts on the basis of the equation of the calibration curve.

STATISTICS

We carried out all statistical calculations with the freely available R programming language (version 3.0.2). Hypothesis tests were carried out, if not mentioned explicitly, as 2-tailed unpaired tests. In cases where the parametric *t*-test was applied (evaluating the null hypothesis that the means of 2 normally distributed populations are equal), approximate normal distribution was verified by Shapiro–Wilk test (evaluating the null hypothesis that measurements come from a normally distributed population).

To show the distribution of miRNA measurements, we generated box-whisker plots, and to provide a per-measurement representation, we provided bee swarm plots as included in the beeswarm R package.

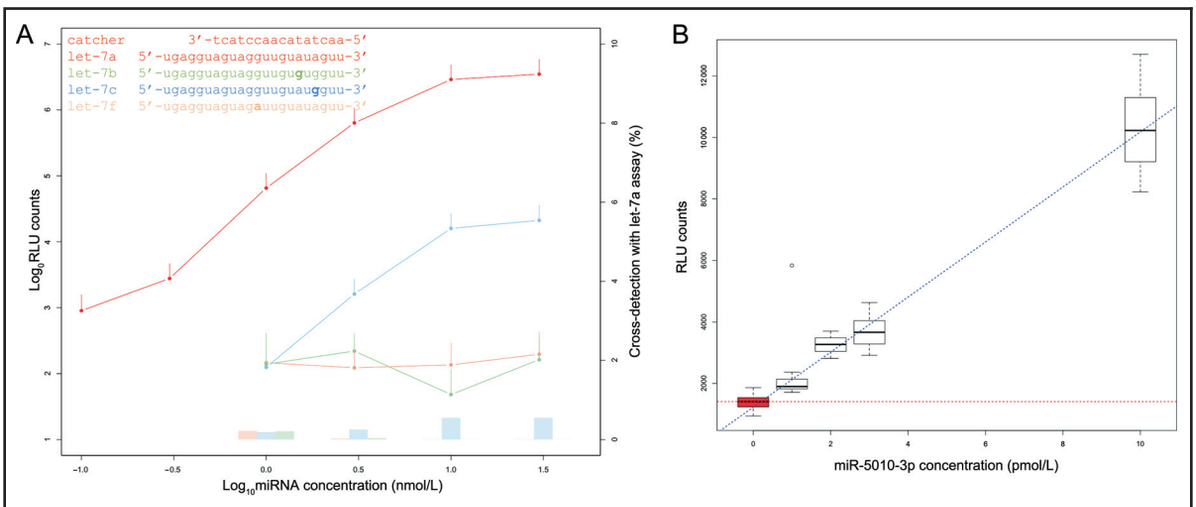


Fig. 2. Technical specificity and analytical sensitivity.

(A), Technical specificity of the assay for miRNA let-7a. The red line shows the response for let-7a. The detected false-positive signals for 3 miRNAs with a single base difference (let-7b, let-7c, and let-7f) are also shown. The bar charts represent the respective percentages of false-positive signals (scale to the right of the plot). (B), Analytical sensitivity of the miRNA assay derived from dilution series of synthetic miR-5010–3p miRNAs. At a concentration of 1 pmol/L, the signal intensity is above the background (red). The blue dashed line indicates linear quantification up to a concentration of 10 pmol/L.

Results

miRNA IMMUNOASSAY

As shown in Fig. 1, our assay works as follows. Total RNA isolated from PAXgene blood is used for the hybridization assay. The total RNA is hybridized with a biotinylated DNA catcher and forms a DNA-miRNA duplex. Streptavidin-coupled magnetic beads are added to the solution, and the DNA catcher binds to the beads through biotin–streptavidin interaction. Unbound miRNAs and other RNAs are washed away so that just the DNA-miRNA duplex remains. A monoclonal antibody specific to DNA-miRNA hybrids labeled with acridinium ester is added to the solution, binding to the DNA-miRNA hybrids. A light signal proportional to the number of DNA-miRNA hybrids is monitored and reported. Altogether, the entire experimental setup, including RNA purification and miRNA profiling, requires <3 h.

SPECIFICITY OF THE IMMUNOASSAY

To evaluate the analytical specificity of the immunoassay, we distinguished members of the let-7 family that differed by just a single base. The miRNA to be quantified was selected to be hsa-let-7a. Synthetic molecules of this miRNA were added in 6 concentrations from 0.1 to 30 nmol/L to the respective catcher, leading to background-corrected results between 1201 counts (0.1 nmol/L) and 4.6 million counts (30 nmol/L). Next, we carried out the same measurement with the 3 miRNAs hsa-let-7b, hsa-let-7c, and hsa-let-7f. For the lowest concentrations, signals were beyond the detection limit; for the higher concentrations, we measured up to 21 222 counts (hsa-let-7c, 30 nmol/L). The results are shown in Fig. 2A. In this figure, the lines represent log10 values of raw counts and the bars correspond to the percentage of crosstalk (false-positive light signals) with hsa-let-7a. Whereas the signals for hsa-let-7b and hsa-let-7f remained in the range of the background even for the highest concentrations, for hsa-let-7c, low signals at very high concentrations could be measured. The crosstalk never exceeded 0.6%, demonstrating a specificity of 99.4% for the miRNA immunoassay.

SENSITIVITY AND LOWER LIMIT OF DETECTION OF THE IMMUNOASSAY

Next, we systematically evaluated the limit of detection of the immunoassay. We selected 1 of the miRNAs included in our AD panel *(11)*, namely hsa-miR-5010–3p. With a catcher probe, we performed 20 replicates for different concentrations between 1 and 10 pmol/L. Additionally, we performed 20 replicates of blank controls representing the background signal. As shown in Fig. 2B, we were able to measure signals substantially exceeding the background noise even for miRNAs at concentrations of 1 pmol/L. Whereas the blank controls (shown in red) had a median intensity of 1411 RLU counts (SD 211) (horizontal red dashed line), 1 pmol/L hsa-miR-

5010–3p resulted in 1904 RLU counts (SD 863) ($P = 0.001$, 2-tailed unpaired t-test). For 2 pmol/L hsa-miR-5010–3p, 3270 RLU counts (SD 314) were reported; for 3 pmol/L 3666 RLU (SD 493), and for 10 pmol/L 10 226 RLU (SD 1208). Altogether, the concentration of hsa-miR-5010–3p correlated significantly with the counts measured by our assay (Pearson correlation 0.998, $P = 0.0001$). In all measurements carried out with our immunoassay, we recorded just a single outlier (Fig. 2B, concentration of 1 pmol/L, Grubbs test $P < 0.001$).

ABILITY TO MEASURE MODERATE CHANGES IN miRNA ABUNDANCE

The variation in blood-based miRNA concentrations in diseases is frequently limited. We have found that variations in circulating miRNA patterns are usually moderate (2-fold expression changes). We thus explored the potential of the miRNA immunoassay to measure changes in concentrations typical for miRNAs found in previous studies. Specifically, we carried out 2 experiments on different concentration scales. First, we started at a concentration of 3 pmol/L and increased the concentration by 0.3 pmol/L in each step until we reached an absolute concentration of 4.8 pmol/L after 7 dilution steps. The R^2 between the concentration and RLU counts reached 0.91 (see online Supplemental Fig. 1). For all measured data points, we found deviation between the expected measurement given the linear regression line and the actual measurement to be <5%. For the 3.6 pmol/L data point, a slightly higher difference was observed (expected according to regression line, 3760 RLU; actually measured, 3995 RLU). Nevertheless, our assay was able to measure even 10% changes reliably in the lower concentration range. To demonstrate that this could also be achieved for the higher-abundance miRNAs, we performed similar experiments, increasing abundance by an order of magnitude. Specifically, we started at 30 pmol/L concentration and increased it by 3 pmol/L up to 60 pmol/L in the 11th step. The R^2 value was even higher and reached 0.98 (Fig. 3). These results demonstrate the linearity of measurements for concentrations of >3 orders of magnitude and also provide evidence that even small changes in miRNA abundance can be quantified by our prototype assay.

MULTIPLEX IMMUNOASSAY

Originally, the assay format was designed as single-plex assay. Although this setup does not prevent routine application, an automated measurement of several miRNAs from the same sample would be beneficial. Therefore, we explored the potential of serial multiplexing. We mixed 8 synthetic miRNAs (miR-5010–3p, miR-151a-3p, let-7d-3p, miR-107, miR-26b-5p, miR-103a, miR-26a-5p, and let-7f-5p) in increasing concentrations. Starting from the miRNA with lowest concentration, we performed measurement of the single-plex assay. The supernatant, however, was not discharged but reentered the measurement cycle with the next miRNA. The same experiments were also done for aliquots of the single-plex assay. The results of single-plex vs multiplex are shown in online Supplemental Fig. 2. Generally, we observed a good correlation; however, those miRNAs with just a single mismatch, such as miR-26a and miR-26b, showed slight variations. Additionally, the experiments revealed a lower performance for let-7f-5p. These preliminary results demonstrate that 8-plex measurements are possible but that increasing the degree of multiplexing decreases the analytical specificity and sensitivity of the assay.

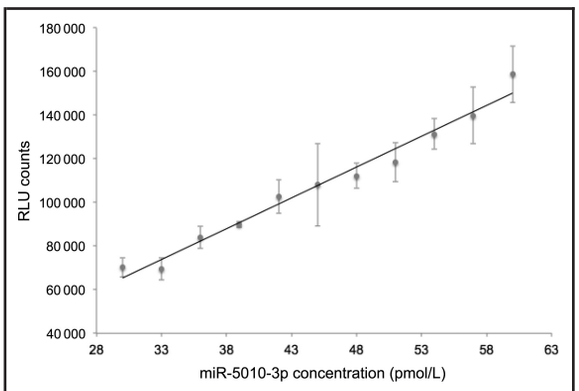


Fig. 3. Increased resolution of quantification in the range between 30 and 60 pmol/L derived from synthetic miR-5010–3p miRNAs.

TRANSFER TO BIOLOGICAL MEASUREMENTS

After exploring the limit of detection, analytical sensitivity, and specificity of our miRNA immunoassay with synthetic miRNAs, we tested 4 miRNAs of our AD miRNA panel, hsa-miR-5010–3p, hsa-miR-26a-5p, hsa-miR-151a-3p, and hsa-let-7d-3p, with 40 replicates of biological samples to evaluate their potential for clinical application beyond the measurement of the synthetic miRNAs presented above. The miRNAs were selected so that most informative markers of the signature were combined while ensuring that lower-abundance markers were also included. Thus, we purposely selected the three -3p mature and the higher-abundance -5p mature form of miR-26a. We generated calibration curves for all 4 miRNAs, as described in Methods, to enable quantification with our novel assay.

These previously published miRNAs allow for detecting patients with AD with diagnostic accuracy, specificity, and sensitivity of 82%, 85%, and 80%, respectively (area under the curve 0.91) *(11)*. On the immunoassay analyzer system, we measured 40 replicates for the 4 miRNAs and controlled the process with 20

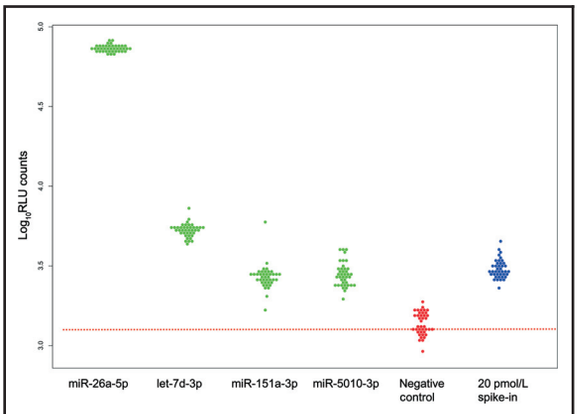


Fig. 4. Signal intensity of 4 AD miRNAs (green dots), negative controls (red dots), and spike-in miRNAs (blue dots) from pooled blood samples.

The signal intensity is presented for 40 samples. Background is shown by the red dashed line.

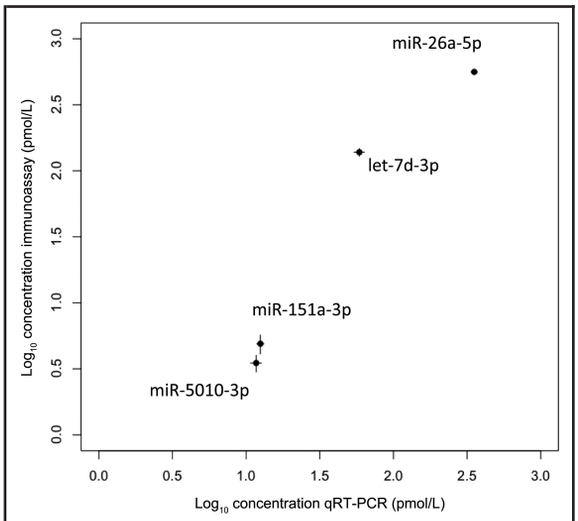


Fig. 5. Correlation with qRT-PCR for 4 Alzheimer miRNAs in blood samples.

pmol/L spike-in controls (Fig. 4). The measurements were carried out with aliquots of the pooled samples by use of the single-plex assay. Again, even for the lowest-abundance miRNAs hsa-miR-5010–3p and hsa-miR-151a-3p, stable signals above the background were observed. For the background, we calculated 1391 RLU (SD 222). For miR-5010–3p, RLU counts were already 2835 (SD 516) (2-tailed unpaired *t*-test between background and miR-5010–3p, $P < 10^{-20}$). For miR-151a-3p, 2738 RLU (SD 604) was found, with 2-tailed unpaired *t*-test significance of $<10^{-20}$, indicating that the difference between this miRNA and the background was highly significant.

In all 240 measurements, 2 outliers (0.8%) were observed. For miR-26a-5p, the mean concentration was 561.3 pmol/L (SD 19.9), let-7d-3p had a mean concentration of 38.3 pmol/L (SD 9), miR-151a-3p had a mean concentration of 5 pmol/L (SD 0.8), and miR-5010–3p had a mean concentration of 3.5 pmol/L (SD 0.5). Given these mean values and SDs, we calculated CV values of 0.04 (miR-26a-5p), 0.24 (let-7d-3p), 0.16 (miR-151a-3p), and 0.13 (miR-5010–3p). Although the CV values were generally low (miR-26a-5p showed a CV of 0.04), let-7d-3p showed an increased CV. The CV values of the blood samples were in the same range as the technical evaluation CV values.

In developing a new test, it is important to benchmark it against the gold standard, in this case qRT-PCR. We quantified the same samples by qRT-PCR as described in Methods. We found a high correlation between qRT-PCR and the Advia Centaur system (Pearson correlation >0.994, $P = 0.006$) (Fig. 5). For hsa-miR-5010–3p, hsa-miR-151a-3p, let-7d, and hsa-miR-26a-5p, the concentrations on the immunoanalyzer system were 3.5, 5, 38.3, and 561.3 pmol/L, respectively, and on qRT-PCR the concentrations were 11.7, 12.5, 58.7, and 335.6 pmol/L. Although these results indicated differences between the technologies, the results showed a general concordance.

Discussion

Our novel method involves hybridization of miRNA from a patient sample to complementary biotinylated DNA oligonucleotides, followed by detection of the DNA-miRNA hybrids by a monoclonal antibody that specifically binds to DNA-miRNA hybrids. Using this setup, we were able to obtain a prototype assay that can measure miRNAs from biological samples without any preamplification step. Our assay has an analytical specificity of 99.4%, a limit of detection in the range of 1 pmol/L, and a time to result of <3 h, including RNA purification and miRNA profiling. We obtained stable results over a dynamic range of 4 orders of magnitude. Additionally, the amplification-free detection allows for less biased miRNA measurements. This advantage, however, results in a current lower limit of detection of 1 pmol/L. Although many blood-based miRNAs can be profiled with the proposed assay, the sensitivity has to be further improved to measure other samples with lower miRNA concentrations, such as serum. Another drawback of our assay is the currently limited multiplexing capability. We demonstrate first results on a multiplexing concept here, but more work is required to obtain the same specificity as for the single-plex assay. Another point that has to be taken into account is that the used

antibody can react with different DNA-RNA hybrids with different affinity *(24)*, influencing the sensitivity of the assay for this miRNA and requiring additional calibration.

In a test on clinical samples, we found an outstanding correlation with qRT-PCR data (Pearson correlation >0.994), which as of now represents the gold standard for miRNA expression analysis. Our assay is currently a research assay that aims to lay the basis for further development, with the challenging goal to promote the usage of miRNAs as clinical IVD tests.

Besides its application to measure miRNAs, our assay design bears the potential to be extended to other nucleic acid test formats, in particular to those that still require preamplification of the target nucleic acid. For example, the method described by Yehle et al. *(25)*, which allows bacterial typing by hybridization of 16s rRNA to strain-specific oligonucleotides, could be adapted to our automated assay format. Moreover, high-abundance mRNAs or rRNAs could be quantified by hybridization to complementary DNA oligonucleotides in the assay format described in this article.

Our miRNA immunoassay has a low time-to-result, comparable to that of qRT PCR, and is still faster than NGS, for which typically at least 1 day (and frequently several days) is required. At the same time, our assay is inexpensive, with costs in the same range as established and marketed immunoassays, which are below those of qRT-PCR or even NGS, and microarrays, which are still in the range of several hundred dollars. In turn, NGS has a much higher multiplexing capability and allows for integrative screening of all miRNAs, even those that are not annotated in databases. NGS is thus a perfect biomarker discovery tool, whereas our assay is tailored for much higher throughput in terms of samples at a decreased degree of multiplexing. Among the most important points with respect to our miRNA immunoassay is that the required hardware is installed in many central laboratories of hospitals worldwide.

In summary, we developed a method that has the potential to change the current practice to measure miRNAs, by providing a means to analyze miRNAs on commonly used immunoassay analyzers, thus providing substantial advantages over existing methodologies.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: A. Kappel, Siemens AG; Y. Huang, Siemens; H. Schwarz, Siemens Healthcare Diagnostics Products GmbH—Predevelopment; W. Gumbrecht, Siemens AG. C.F. Staehler, Siemens Healthcare.
Consultant or Advisory Role: A. Keller, Siemens Healthcare.
Stock Ownership: A. Kappel, Siemens AG; W. Gumbrecht, Siemens.
Honoraria: None declared.
Research Funding: A. Keller, Siemens Healthcare.
Expert Testimony: None declared.
Patents: A. Kappel, patent no. WO2013/135581; B. Meder, several on miRNA; H. Schwarz, 201204166.

Role of Sponsor: The funding organizations played a direct role in the final approval of manuscript.

References

1. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 2009;19:92–105.
2. Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W, Cui Q. An analysis of human microRNA and disease associations. PLoS One 2008;3:e3420.
3. Osbourne A, Calway T, Broman M, McSharry S, Earley J, Kim GH. Downregulation of connexin43 by microRNA-130a in cardiomyocytes results in cardiac arrhythmias. J Mol Cell Cardiol 2014;74:53–63.
4. Okada N, Lin CP, Ribeiro MC, Biton A, Lai G, He X, et al. A positive feedback between p53 and miR-34 miRNAs mediates tumor suppression. Genes Dev 2014;28:438–50.
5. Long JM, Ray B, Lahiri DK. MicroRNA-339–5p down-regulates protein expression of beta-site amyloid precursor protein-cleaving enzyme 1 (BACE1) in human primary brain cultures and is reduced in brain tissue specimens of Alzheimer disease subjects. J Biol Chem 2014;289:5184–98.
6. Keller A, Leidinger P, Bauer A, Elsharawy A, Haas J, Backes C, et al. Toward the blood-borne miRNome of human diseases. Nat Methods 2011;8:841–3.
7. Keller A, Leidinger P, Borries A, Wendschlag A, Wucherpfennig F, Scheffler M, et al. miRNAs in lung cancer: studying complex fingerprints in patient's blood cells by microarray experiments. BMC Cancer 2009;9:353.
8. Keller A, Leidinger P, Gislefoss R, Haugen A, Langseth H, Staehler P, et al. Stable serum miRNA profiles as potential tool for non-invasive lung cancer diagnosis. RNA Biol 2011;8:506–16.
9. Keller A, Leidinger P, Lange J, Borries A, Schroers H, Scheffler M, et al. Multiple sclerosis: microRNA expression profiles accurately differentiate patients with relapsing-remitting disease from healthy controls. PLoS One 2009;4:e7440.
10. Keller A, Leidinger P, Steinmeyer F, Stahler C, Franke A, Hemmrich-Stanisak G, et al. Comprehensive analysis of microRNA profiles in multiple sclerosis including next-generation sequencing. Mult Scler 2014;20:295–303.
11. Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, et al. A blood based 12-miRNA signature of Alzheimer disease patients. Genome Biol 2013;14:R78.
12. Leidinger P, Keller A, Backes C, Huwer H, Meese E. MicroRNA expression changes after lung cancer resection: a follow-up study. RNA Biol 2012;9:900–10.
13. Leidinger P, Keller A, Borries A, Huwer H, Rohling M, Huebers J, et al. Specific peripheral miRNA profiles for distinguishing lung cancer from COPD. Lung Cancer 2011;74:41–7.
14. Leidinger P, Keller A, Borries A, Reichrath J, Rass K, Jager SU, et al. High-throughput miRNA profiling of human melanoma blood samples. BMC Cancer 2010;10:262.
15. Hausler SF, Keller A, Chandran PA, Ziegler K, Zipp K, Heuer S, et al. Whole blood-derived miRNA profiles as potential new tools for ovarian cancer screening. Br J Cancer 2010;103:693–700.
16. Santos JI, Teixeira AL, Dias F, Mauricio J, Lobo F, Morais A, Medeiros R. Influence of peripheral whole-blood microRNA-7 and microRNA-221 high expression levels on the acquisition of castration-resistant prostate cancer: evidences from in vitro and in vivo studies. Tumour Biol 2014;35:7105–13.
17. Schrauder MG, Strick R, Schulz-Wendtland R, Strissel PL, Kahmann L, Loehberg CR, et al. Circulating microRNAs as potential blood-based markers for early stage breast cancer detection. PLoS One 2012;7:e29770.
18. Schultz NA, Dehlendorff C, Jensen BV, Bjerregaard JK, Nielsen KR, Bojesen SE, et al. MicroRNA biomarkers in whole blood for detection of pancreatic cancer. JAMA

2014;311:392–404.
19. The worldwide market for in vitro diagnostic (IVD) tests. 8th ed. New York: Kalorama Information; 2012.
20. German Medical Association and Paul Ehrlich Institute. Guidelines for the collection of blood and blood components and the use of blood products (hemotherapy). Cologne, Germany: Deutscher Aerzteverlag; 2007.
21. Guide to the use, preparation and quality control of blood components. Recommendation no. R(95) 15. 15th ed. Strasbourg, France: Council of Europe; 2009.
22. Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT, et al. Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic Acids Res 2005;33:e179.
23. Dwyer R. The ADVIA Centaur infectious disease assays: a technical review. J Clin Virol 2004;30(Suppl 1): S1–5.
24. Boguslawski SJ, Smith DE, Michalak MA, Mickelson KE, Yehle CO, Patterson WL, Carrico RJ. Characterization of monoclonal antibody to DNA. RNA and its application to immunodetection of hybrids. J Immunol Methods 1986;89:123–30.
25. Yehle CO, Patterson WL, Boguslawski SJ, Albarella JP, Yip KF, Carrico RJ. A solution hybridization assay for ribosomal RNA from bacteria using biotinylated DNA probes and enzyme-labeled antibody to DNA:RNA. Mol Cell Probes 1987;1:177–93.
26. Weeks I, Campbell AK, Woodhead JS. Two-site immunochemiluminometric assay for human alpha 1-fetoprotein. Clin Chem 1983;29:1480–3.

Longitudinal study on circulating miRNAs in patients after lung cancer resection

Petra Leidinger[1], Valentina Galata[2], Christina Backes[2], Cord Stähler[3], Stefanie Rheinheimer[1], Hanno Huwer[4], Eckart Meese[1,*] and Andreas Keller[2,*]

[1] Department of Human Genetics, Saarland University, Homburg, Germany

[2] Chair for Clinical Bioinformatics, Saarland University, Saarbrücken, Germany

[3] Siemens AG, Strategy Division, Erlangen, Germany

[4] Department of Cardiothoracic Surgery, Heart Center, Völklingen, Germany

[*] These authors equally contributed as senior author

***Correspondence to**: Petra Leidinger, **email**: p.leidinger@mx.uni-saarland.de*

***Keywords**: microRNA, plasma, Lung cancer, metastases, follow-up*

***Received**: January 05, 2015* ***Accepted**: May 25, 2015* ***Published**: May 29, 2015*

ABSTRACT

There is an urgent need of comprehensive longitudinal analyses of circulating miRNA patterns to identify dynamic changes of miRNAs in cancer patients after surgery. Here we provide longitudinal analysis of 1,205 miRNAs in plasma samples of 26 patients after lung cancer resection at 8 time points over a period of 18 months and compare them to 12 control patients. First, we report longitudinal changes with respect to the number of detected miRNAs over time and identified a significantly increased number of miRNAs in patients developing metastases (p = 0.0096). A quantitative analysis with respect to the expression level of the detected miRNAs revealed more significant changes in the miRNA levels in samples from patients without metastases compared to the non-cancer control patients. This analysis provided further evidence of miRNA plasma levels that are changing over time after tumor resection and correlate to patient outcome. Especially hsa-miR-197 could be validated by qRT-PCR as prognostic marker. Also for this miRNA, patients developing metastases had levels close to that of controls while patients that did not develop metastases showed a significant up-regulation.

In conclusion, our data indicate that the overall miRNome of a patient that later develops metastases is less affected by surgery than the miRNome of a patient who does not show metastases. The relationship between altered plasma levels of specific miRNAs with the development of metastases would partially have gone undetected by an analysis at a single time point only.

INTRODUCTION

The fact that most non-small cell lung cancer (NSCLC) patients are diagnosed in late stages with locally advanced or metastatic disease, makes NSCLC to one of the most deadly cancers with a 5-year overall survival rate of around 17% [1]. The detection and resection of NSCLC in early stages is of profound relevance as it is normally correlated with a substantially improved prognosis [2]. Nevertheless, the rate of recurrences and metastases is high, even in early stage lung cancers. In a study on more than 900 patients who underwent early NSCLC curative-intend resection about 13% of patients developed lung cancer recurrence and 78% of the recurrences occurred within two years after operation. [3]. Disseminated tumor cells can already be present in early tumor stages before resection but they are not detected by conventional histopathology analysis and tumor staging and are often staged as N0 tumors [4]. The overall incidence of recurrence lies around 30% to 70% depending on lung cancer stage [5-7]. To improve the overall survival rate there is an urgent need for the

identification of new prognostic factors. Second, intensive follow-up is important to reduce lung cancer mortality by the detection of recurrences after surgery [8].

MicroRNAs (miRNAs) found in body fluids indicate a high impact as diagnostic and prognostic biomarker as they play a crucial role in many cellular processes by regulating an extended number of target genes due to mRNA degradation or inhibition of the translation of the target mRNA [9, 10]. Until now, substantial effort has been undertaken to identify disease-specific miRNA profiles suitable for early diagnosis of diseases and to predict disease outcome [11, 12]. While many case-control studies have revealed a plentitude of miRNAs as biomarker candidates, dynamic changes over extended time periods have not been explored for the majority of them. Most respective studies are either limited in the number of time-points, patients, or considered miRNAs.

An analysis of the physiological fluctuation of serum miRNA profiles of samples taken from 12 healthy individuals over varying time periods up to 17 months revealed miRNA profiles that showed a high correlation and no significantly differentially expressed miRNAs were found. This suggests that circulating miRNAs are stable over extended time periods in healthy individuals [13]. Thus, changes in the overall abundance of circulating miRNAs due to a certain disease make them to good biomarker candidates. Changes of few miRNAs have for example already been monitored in a kinetic study over months in serum of 15 colorectal cancer patients [14]. However, just few studies investigate circulating miRNA profiles for changes between lung cancer samples collected before and after cancer resection [15, 16]. We recently performed a first follow-up study on lung cancer patients over a period of 18 months after lung cancer resection to identify miRNA signatures that possibly contribute to disease monitoring [17]. Although we analyzed 8 different time points and profiled a large number of miRNAs, a major limitation of this study was the small cohort size of only 5 patients. We now screened 26 patients for up to 8 time points – prior to surgery, following surgery and subsequently in 3 months intervals. Additionally, we compare the miRNAs identified in plasma of the lung cancer patients to those measured in samples obtained from12 control patients that suffered from other non-cancer lung diseases. Altogether, 215 single complex miRNA profiles have been generated using a microarray approach. Since one key criterion for a potential application in clinics beyond technical sensitivity and specificity is the reproducibility of measurements we applied a microarray technology that has been described to be most reproducible among 12 commonly used commercial systems [18]. Following background correction, adjustment for batch effects and normalization, bioinformatics analysis was applied in order to identify and validate the most relevant regulated miRNAs towards their usefulness as potential prognostic lung cancer biomarker.

RESULTS

The main aim of our study was to provide a comprehensive longitudinal analysis of circulating miRNAs in plasma of lung cancer patients following surgery to identify miRNAs with prognostic relevance. In detail, we analyzed 1,205 different miRNAs in 26 lung cancer patients over a period of 18 months measured at 8 time points including one time point prior and up to seven time points after cancer resection. The expression profiles of the lung cancer samples were compared to 12 patients suffering from other non-cancer lung diseases that served as control.

miRNA repertoire in lung cancer patients over time and in non-cancer controls

We determined for all lung cancer patients and each time point (TP) and for all controls the average number of miRNAs detected in each sample (Figure 1). The samples obtained from lung cancer patients contained independent of the time point a lower number of miRNAs compared to the non-cancer controls (on average 295 miRNAs were detected in lung cancer samples and 331 in control samples). However, only for TP5 the difference between the average number of miRNAs detected in the lung cancer plasma samples compared to controls was significant (adjusted p-value 0.025). Lung cancer samples collected at TP2 showed with an average number of 321 detected miRNAs the lowest difference compared to controls (adjusted p-value of 0.67). Since the analysis of plasma samples obtained from the same individuals at different time points also enables paired testing of consecutive time-points we investigated whether significant changes of miRNA levels can be observed over time. Here, we found the most significant differences average number of detected miRNAs between TP1 and TP2 (raw p-value 0.019) and between TP5 and TP6 (raw p-value 0.016).

We also asked whether the miRNA repertoire differs in its quantity between lung cancer patients developing a metastases compared to those not developing metastases. The results are presented in Figure 1B, where for both groups and all time points the average number of miRNAs are shown. For patients not developing metastases we observed significant increase of miRNA repertoire from TP1 to TP2 and TP5 to TP6. For the other patients no significant alterations in the miRNA number were discovered, although the differences between different time points seems to be higher. But, as the standard deviation for the number of detected miRNAs is higher in the samples obtained from patients that developed metastases, the differences were not significant. But generally, we observed larger miRNA repertoire of

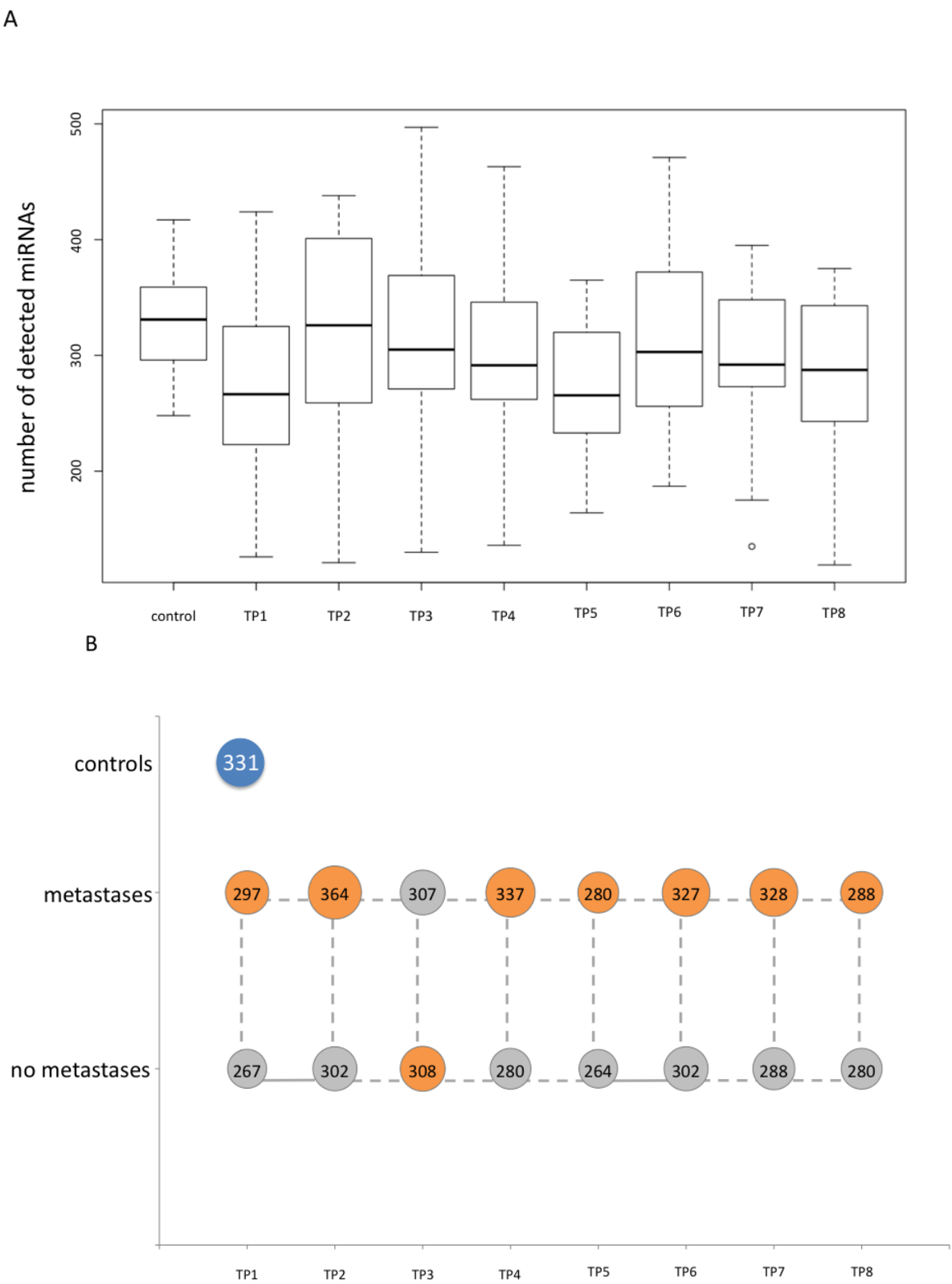


Figure 1: Comparisons of the overall numbers of detected miRNAs. A. Box plot showing the overall number of detected miRNAs for all non-cancer control samples and all lung cancer samples for each time point separately. **B.** Bubble plot indicating the overall number of detected miRNAs for the non-cancer control patients as well as for the lung cancer patients that developed metastases and the lung cancer patients that did not develop metastases for each time point, separately.

Table 1: Correlation analysis of miRNA pattern over time for all lung cancer patients combined and the non-cancer control patients

miRNA	Correlation	*p*-Value	Lower CI	Upper CI
hsa-miR-181d	-0.95	0.0003	-0.99	-0.80
hsa-miR-670	-0.81	0.0139	-0.95	-0.38
hsa-miR-196b	-0.80	0.0179	-0.95	-0.34
hsa-miR-3148	-0.78	0.0219	-0.95	-0.30
hsa-miR-762	-0.76	0.0290	-0.94	-0.25
hsa-miR-539	-0.74	0.0342	-0.93	-0.22
hsa-let-7d*	0.71	0.0467	0.16	0.93
hsa-miR-484	0.72	0.0432	0.17	0.93
hsa-miR-3663-5p	0.72	0.0429	0.18	0.93
hsa-miR-183	0.73	0.0385	0.20	0.93
hsa-miR-17*	0.74	0.0362	0.21	0.93
hsa-let-7c	0.74	0.0345	0.22	0.93
hsa-miR-548c-5p	0.75	0.0326	0.23	0.94
hsa-miR-3189	0.75	0.0325	0.23	0.94
hsa-miR-20b	0.75	0.0322	0.23	0.94
hsa-miR-29b	0.75	0.0321	0.23	0.94
hsa-miR-224	0.75	0.0317	0.24	0.94
hsa-miR-501-5p	0.76	0.0301	0.25	0.94
hsa-miR-20a	0.76	0.0280	0.26	0.94
hsa-miR-370	0.76	0.0272	0.26	0.94
hsa-miR-18a	0.78	0.0226	0.30	0.94
hsa-miR-532-5p	0.78	0.0220	0.30	0.95
hsa-miR-1915	0.78	0.0217	0.31	0.95
hsa-miR-146b-5p	0.78	0.0212	0.31	0.95
hsa-miR-3654	0.80	0.0177	0.34	0.95
hsa-miR-451	0.80	0.0161	0.36	0.95
hsa-miR-374a	0.81	0.0145	0.38	0.95
hsa-miR-3180-3p	0.84	0.0093	0.45	0.96
hsa-miR-10b*	0.84	0.0087	0.46	0.96
hsa-miR-184	0.85	0.0075	0.48	0.96
hsa-miR-141	0.85	0.0071	0.49	0.96
hsa-miR-4281	0.86	0.0061	0.51	0.97
hsa-miR-454	0.88	0.0038	0.57	0.97
hsa-miR-301a	0.88	0.0037	0.57	0.97

CI = confidence interval

patients that develop metastases. Independent of the time point we observed 286 miRNAs for patients not developing metastases while the remaining patients revealed 316 miRNAs (two-tailed unpaired *t*-test *p*-value of 0.0096). Interstingly, the analysis of the 12 non-cancer control samples revealed 331 detected miRNAs.

For the following quantitative analysis we only focused on the 485 miRNAs that were expressed in at least 5% of all tested 215 individual samples.

Correlation analysis of miRNA pattern over time for all lung cancer patients combined and the non-cancer control patients

To identify miRNAs that show an overall increase or decrease from the first to the last measurement we first calculated pair-wise significance values between the miRNA profiles of the 12 non-cancer controls and the profiles of the 26 lung cancer patients for each of the time points using two-tailed unpaired *t*-test. Next, we correlated the logarithm of the significance values obtained by the two-tailed unpaired *t*-test with the rank of the time points. We discovered 6 negatively and 28 positively correlated miRNAs (raw *p*-value of correlation below 0.05). These 34 miRNAs with correlation values, *p*-values and upper and lower confidence interval are provided in Table 1. Notably, a strong negative correlation indicates that the respective miRNA is not de-regulated in samples from lung cancer patients at the beginning of the time course (high *p*-values at early time points) but shows increasing difference in miRNA plasma levels from non-cancer controls over time (low *p*-values at the end). In contrast, strong positive correlation indicates that the respective miRNA is de-regulated at the beginning (low *p*-values at early time points) but shows decreasing difference to the non-cancer control miRNA level over time (high *p*-values at the end). The miRNAs with correlation values around zero do not show increasing or decreasing significance over time but are rather constantly expressed. Although no miRNA was significant following adjustment, we observed a substantial increased number of miRNAs significant prior to adjustment as compared to the expected number of 24 random miRNAs. Figure 2 presents exemplarily the miRNA plasma levels of hsa-miR-370 (representative for positive correlated miRNAs) and hsa-miR-181d (representative for negative correlated miRNAs).

Correlation analysis of miRNA pattern over time for single lung cancer patients

Beside the analysis of the miRNA changes for all patients combined, our study set-up also allows the analysis of the miRNA time courses for single patients. We calculated for each patient and each miRNA separately correlation values between miRNA expression and time-points and estimated the significance values for the respective correlation. We excluded miRNAs that did not revealed significant correlation for at least 10% of all patients. For the remaining miRNAs we calculated in how many patients a miRNA was positive or negative correlated over time and calculated the difference of positive and negative correlated patients for each of these miRNAs. We excluded miRNAs for which the number of patients with a positive correlation largely corresponded to the number of patients with a negative correlation. As threshold we considered only miRNAs with a difference of at least 30% between positively and negatively correlated patients. We thereby identified 16 miRNAs including 10 positively and 6 negatively correlated miRNA. Although the overall tendency of certain miRNA levels to either increase or decrease over time is in agreement with the results obtained with the expression levels for all patients combined, the data for the single patients show strong variability. These miRNAs indicate that although a general trend exists single patients substantially deviate from the general trend (see Supplemental Figure 1 and Supplemental Table 1).

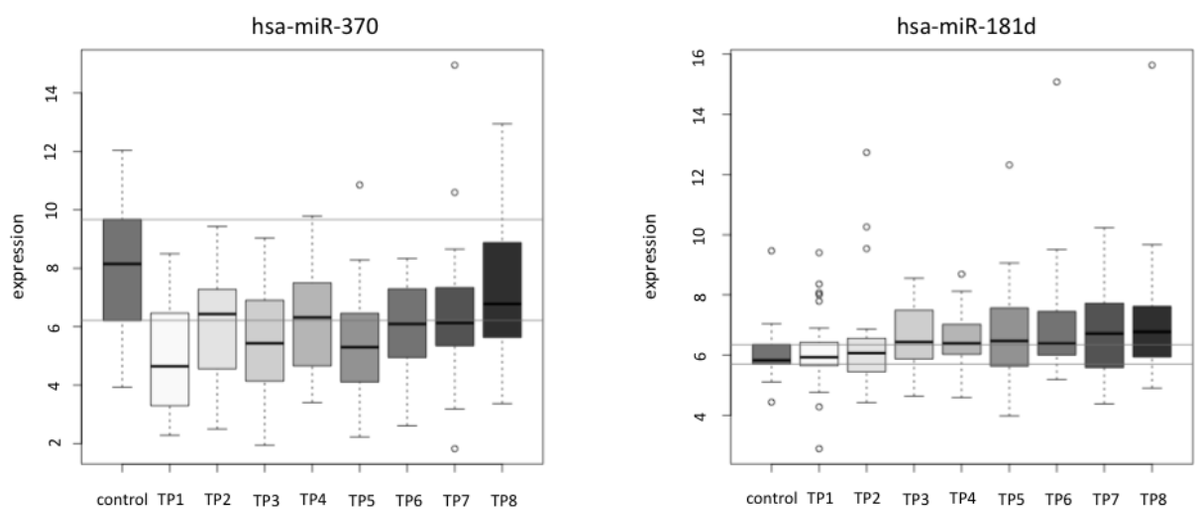


Figure 2: Examples of the correlation analysis of miRNA pattern over time for single miRNAs shown for each patient separately. hsa-miR-370 is an example for a positive correlated miRNA and has-miR-181d is an example for a negative correlated miRNA. In both figure panels the y-axis shows the normalized expression values (in log scale) and the x-axis indicates the time points 1 to 8.

Identification of plasma miRNAs influenced by the development of metastases

To understand the changes of miRNA levels over time we related the changes to a clinical endpoint. This was also in keeping with the goal to discover prognostic miRNAs. Thus, we choose the development of metastases as endpoint and asked whether patients with metastases show different plasma miRNA levels as compared to patients without clinically identified metastases. To this end we calculated significance values for each time point with respect to the two groups of patients, i.e., patients developing metastases (n = 8) versus patients not developing metastases (n = 18). At the time point directly before cancer resection (TP1) we found 25 plasma miRNAs that showed significantly different plasma levels between patients with and those without metastases (non-adjusted p-value < 0.05). At TP2, i.e., shortly after resection four of the 25 miRNAs were still significant, but in total 18 miRNAs showed significantly different abundance (non-adjusted p-value < 0.05). The highest number of 40 miRNAs with significantly different plasma levels between patients that developed lung cancer metastases and patients that did not develop metastases was obtained at TP3 around three moths after resection. At TP4 9 miRNAs were significant, at TP5 33 miRNAs, at TP6 13 miRNAs, at TP7 18 miRNAs, and at TP8 23 miRNAs. However, for the comparisons of the single time points no miRNAs remained significant after multiple testing. This fact is not necessarily due to decreased effect sizes for single time points but may reflect the comparably small cohort size.

We also performed a more general comparison of all expression values independent of the time point and compared all lung cancer samples to the non-cancer controls. To evaluate the patterns we considered both, raw and adjusted p-values. Of the 485 analyzed miRNAs, 139 were significantly altered between cancer patients and non-cancer controls, of which 56 remained significant following adjustment. Lowest p-values of below 10^{-10} were found for hsa-miR-3647-5p and hsa-miR-144. In the comparison of non-cancer controls versus lung cancer patients that did not develop metastases 138 miRNAs were significant (55 following adjustment) and 125 miRNAs for the comparison of controls versus metastases developing patients (41 following adjustment). Importantly, we also discovered 131 miRNAs that were significantly altered between patients that developed metastases and those that did not (38 following adjustment). Here, the highest significance was reached for hsa-miR-197 ($p = 3x10^{-7}$). This miRNA was also significant in the previously mentioned comparison of controls compared to lung cancer patients that did not develop metastases (p = 0.004) while it was not significantly differentially regulated for controls versus patients that developed metastases (p = 1). The most significant changes (p < 0.05) for this miRNA were found at TP2, TP3, and TP5. Another miRNA, hsa-miR-630 was even significant in four time points, i.e., TP1, TP2, TP4, and TP6. Hsa-miR-130b was the most significant miRNA that showed larger deviation of lung cancer patients that developed metastases from controls (p = 0.0004) than patients that did not develop metastases (p = 0.083).

The full list of the 485 miRNAs with the expression data and the non-adjusted p-values is provided in Supplemental Table 2.

To compare the metastases and non-metastases group directly to non-cancer controls, we calculated for each miRNA the p-values for the comparison of its expression value in plasma samples collected from lung cancer patients that developed metastases and those that did not at each time point versus its expression value in plasma samples from non-cancer controls. In total, 139 miRNAs were significant in the comparison of the samples obtained before resection (TP1) from lung cancer patients that did not develop metastases with the non-cancer controls, but only 98 miRNAs in the comparison of the samples obtained before resection (TP1) from lung cancer patients that developed metastases with the non-cancer controls. We observed the same trend in the comparison of the non-cancer controls and the lung cancer samples obtained shortly after resection (TP2). Here 92 miRNAs were significant in the group of patients that did not develop metastases and only 72 in the group of patients that developed metastases. Figure 3 shows the above mentioned comparions for selected miRNAs as pie charts and all data are provided in Supplemental Table 3. The miRNA hsa-miR-197 was significantly up-regulated at 7 time points (non-adjusted p-values) for the group of patients that did not develop metastases while not in the group of patients that developed metastases. Similarly, hsa-miR-1227 was constantly up-regulated, however, again just the patients without metastases were significant. In contrast, hsa-miR-4292 was more significantly down-regulated in the group of patients that developed metastases as compared to the group of patients that did not develop metastases.

We next focused only on the samples obtained from lung cancer patients and compared the samples collected before surgery at TP1 with samples from each other time point after surgery (TP2 to TP8) resulting in 7 comparison. The calculated t-test p-values for the respective comparisons are listed in Supplemental Table 4. This analysis was done separately for patients with and without metastases. For the patients without metastases the comparison of the sample drawn before cancer resection (TP1) and the sample obtained shortly after resection (TP2) revealed 103 significant miRNAs, while we found for the same comparison only 44 significant miRNAs in samples obtained from patients that developed metastases during follow-up and this trend was observed for all of the 7 comparisons. This indicates a trend to a more profound

change in the miRNA pattern for samples of patients that did not develop lung cancer metastases.

We found 2 miRNAs including hsa-miR-454 and hsa-miR-3152 that were significantly deregulated in all seven comparisons, and 2 miRNAs including hsa-miR-181b and hsa-miR-98 that were significantly deregulated in 6 out of the 7 comparisons. Of those miRNAs deregulated in patients without metastases hsa-miR-454 was also significantly deregulated in two comparisons of patients with metastases and hsa-miR-98 in only one. In contrast, hsa-miR-3152 and hsa-miR-181b that were significantly deregulated in patients without metastases were not significantly deregulated in patients with metastases. Hsa-miR-454, hsa-miR-181b, and hsa-miR-98 were down-regulated at TP 2-8 compared to TP 1 in patients without metastases and hsa-miR-3152 showed significantly increased plasma levels at TP 2-8 compared to TP 1.

We also found one miRNA, namely hsa-miR-101, that showed significantly decreased plasma abundance in all seven comparisons of patients with metastases but was not significantly deregulated in the comparisons of patients without metastases. Has-miR-186 was still significant in 6 of 7 comparisons of patients without metastases, but also in two comparisons of patients with metastases. Both miRNAs were down-regulated at time points 2-8 compared to time point 1 in patients with metastases.

In sum, the data demonstrate that miRNA changes over time can be related to clinical end points like the development of metastases and that effects are largest 3 months following surgery.

qRT-PCR validation of selected miRNAs

In the previous section we described miRNAs identified by microarray that are correlated to lung cancer and that have a potential prognostic impact. Using qRT-PCR we exemplarily measured the time courses consisting of the up to 8 time points for 4 patients, including 2 patients did not develop metastases (patients J and P) and two patients that that later on developed metastases (patients V and Z) and three miRNAs (hsa-miR-197, hsa-miR-130b, hsa-miR-762). Additionally, the 12 samples from non-cancer control patients were analyzed using qRT-PCR. One very interesting and potentially prognostic miRNA was hsa-miR-197 as this miRNA was significantly up-regulated in 7 of 8 time points (TP1 to TP7) in plasma of patients that did not develop metastases compared to plasma of non-cancer control patients but it was similarly abundant in plasma from lung cancer patients that developed metastases and in plasma of non-cancer control patients. Investigating the miRNA abundance using qRT-PCR at the different time points for the four patients and 12 controls we were able to reproduce these results. Although the considered cohorts were comparably small, the difference between cases and controls was significant (0.004). While considering all measurements without respect to the time points slightly missed the alpha level of 0.05 ($p = 0.059$), the paired analysis of the time course for both lung cancer patient groups (with metastases and without metastases) was significant ($p = 0.025$). In detail, the time course of all patients matched in general well between microarray and qRT-PCR. The most significant miRNA where the mean expression value of all samples from patients of the metastases group was lower than the mean expression value of all samples from patients of the non-metastases group and all samples from non-cancer controls showed the highest mean value was hsa-miR-130b. Although the time courses of the analyzed patients generally showed a high concordance with a median correlation value of 0.75 for all patients and the controls we were not able to reproduce the lower expression of this miRNA in patients that developed metastases. Especially the time course of patient Z for hsa-miR-130b plasma levels that was measured by microarray could not be validated completely by qRT-PCR. However, the higher plasma levels in non-cancer control samples were indeed validated. As third candidate we picked hsa-miR-762, which shows a similar behavior in the mean expression values according to microarray as hsa-miR-130b. Here, we observed for two patients deviations in the time course as compared to array measurements (patients P and V).

In sum, for patient J all three miRNAs were validated, while for the other patients two of three miRNAs were reproduced. For patients P and V hsa-miR-762 diverged and for patient Z hsa-miR-130b.

In Supplemental Figure 2A-2L a comparison of the microarray data and the qRT-PCR data for the up to 8 samples for the four different lung cancer patients and the three miRNAs is shown.

As there is no endogenous smallRNA or miRNA that can reliably serve as “housekeeping gene” that is stably detected/abundant in serum or plasma [19] we used as normalizer the miRNA mimic syn-cel-miR-39, that was spiked into the plasma sample before RNA isolation. Interestingly, this synthetic miRNA cannot only serve as normalizer but can also be used to control the extraction process. In the present study the mean Ct value was 21±3.13.

DISCUSSION

There is an undisputable requirement for molecular tests to assist in the diagnosis, prognosis and prediction of cancers including lung cancer. Although histological evaluation of tumor tissues from biopsies will at least for the near future remain the ‘gold standard’ of diagnosis, these samples necessarily represent only a single time point in the overall tumor development. Blood based tests open the possibility to monitor the course of tumor development. Currently, there are, however, only few blood based markers in clinical use including CA125

for ovarian cancer, CA19-9 for pancreatic cancer, CEA for colon cancer, and PSA for prostate cancer [20]. These established markers have, however, rather limited accuracy, which can be improved by longitudinal measurements as shown for PSA where continuously increasing levels strongly indicate a carcinoma [21]. As of now, there is no biomarker established for lung cancer in a screening setting.

Beside the need to have measuring from different time points of tumor development, there is a need to have biomarkers that do not rely on the measuring of a single kind of molecule like the aforementioned markers. Since combinations of different molecules can be more accurate and are likely to be more robust than single-molecule markers, an increasing number of studies aimed at identifying marker signatures. Notably miRNA signatures appear of especial interest due to their rather high stability in body fluids. Since the first description of miRNAs in serum of patients with diffuse large B cell lymphoma, blood born miRNAs have been related to tumor diagnosis and prognosis [19, 22]. The majority of these studies, however, analyzes miRNA pattern at one time point only. In addition, the analysis of circulating miRNAs has some methodological challenges. As these

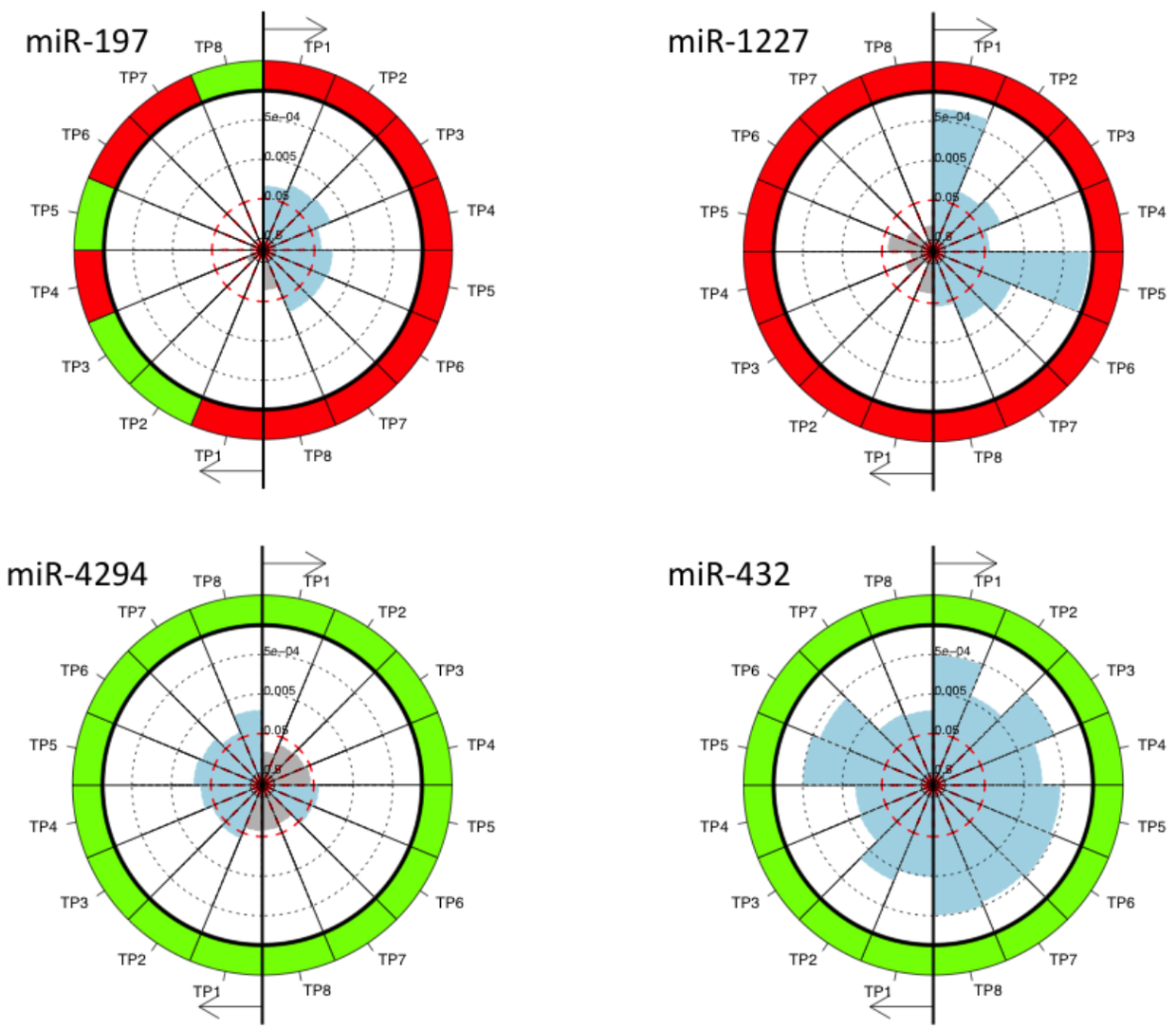


Figure 3: The pie charts for miRNAs significant in the comparison of non-cancer controls and the lung cancer samples collected at the different time points and for patients with and without metastases separately. MiRNAs were measured at eight different time points. The time points are numbered TP1 to TP8 and each time point TP1 to TP8 is compared to the non-cancer controls. The right part of each pie chart represents the comparison between non-cancer controls and lung cancer patients without metastases and the left part of the pie chart represents the comparison between non-cancer controls and lung cancer patients with metastases. Each sector represents one comparison with the color of the outer ring indicating down-regulation (green) or up-regulation (red) at the respective time point compared to non-cancer controls. The inner part of the circle indicates the significance values with blue shaded sectors representing significant differences and the grey sectors not significant differences.

challenges are exhaustively summarized in a recent review article by Moldovan et al. [23] we do not want to further discuss them here in more detail. Nevertheless, Moldovan et al. [23] found out that there are many studies comparing different biological fluids side-by-side and find little or no difference in extracellular miRNA quantification. Interestingly, higher concentrations were consistently found in sera and a possible explanation for that might be that platelets, that contain a wide spectrum of miRNAs, may release their content into the serum during coagulation. This is one argument for the use of plasma samples. But, we are aware of the disadvantages of heparinized plasma samples in terms of the effect of heparin on downstream applications. However, as for the current study only heparinized plasma was available we established a protocol that includes a heparin digestion step to isolate RNA that could be used for downstream analyses like microarray and qRT-PCR. We also checked for RNA extraction efficiency by using a synthetic miRNA mimic (syn-cel-miR-39).

Our study on 1,205 different miRNAs in 26 lung cancer patients over a period of 18 months measured at up to 8 time points is the most comprehensive longitudinal analysis of miRNA signatures in cancer patients. This is the follow-up of a proof-of-principle study that we published previously [17]. However, our previous study focused only on the changes of the plasma miRNA profile over time after surgery without the comparison with non-cancer control samples. In addition, we compared our microarray data with circulating miRNAs that were previously described in literature as deregulated in lung cancer and found 11 of 35 published miRNAs detected in all samples prior to surgery. In the present study these 11 miRNAs were also detected in all analyzed plasma samples obtained from lung cancer patients at TP1, i.e., prior to surgery. However, these 11 miRNAs were also detectable in all of the analyzed samples from non-cancer controls and there was no difference in expression level after adjustment between both groups. These findings indicate that the respective miRNAs are not well suited as reliable diagnostic biomarkers for lung cancer.

For the correlation analysis of the single patients and time points, we identified in the present study 6 negative correlated miRNAs and 10 positive correlated miRNAs. The comparison of the correlated miRNAs for each lung cancer patient between our former study and the present study is complicated by the different analysis methods. In the former study, we considered the miRNAs with positive or negative correlation for each patient, respectively. In the present study, we also calculated the correlation of each miRNA for each patient but excluded those miRNAs that do not show a general trend to positive or negative correlation. Thus the list of miRNAs is smaller and we find only an overlap of two miRNAs. The miRNA hsa-miR-24 was negatively correlated in patient B in the former study and is also negatively correlated in most of the 26 patients, including patient B analyzed in the present study. The miRNA hsa-miR-1202 was negatively correlated in patient D in the former study but in the present study it is positively correlated in the majority of patients. Interestingly, when only considering patient D it shows a negative correlation.

A correlation analysis of the plasma miRNAs identified in samples of all lung cancer patients combined and the non-cancer control patients revealed 6 negative correlated miRNAs that showed no deregulation of the lung cancer samples at the beginning but increasing difference from the non-cancer control samples in expression over time and 28 positive correlated miRNAs that were deregulated in lung cancer samples at the beginning but levels to the non-cancer control expression level over time. As control samples were not included in our former study, a comparison for this analysis was not possible.

Overall, our data show that miRNA levels are changing over time after tumor surgery and that these changes are not necessarily fluctuating around a median value but can have a clear tendency to either increase or decrease. Since circulating miRNA profiles in healthy individuals seem to be rather stable over time, the observed changes in our study are likely to be disease related [13]. This idea of miRNA pattern changing in the course of a disease under treatment is consistent with previous reports on changes in the abundance of circulating miRNAs between samples collected prior and after radiochemotherapy of head and neck cancer patients [24]. A study on 4 miRNAs in 82 lung carcinoma patients identified altered serum levels in samples obtained before surgery and samples obtained 10 days after surgery [16]. Likewise, 90 miRNAs were analyzed in plasma obtained before and after tumor removal in 32 squamous cell lung cancer patients [15].

It remains the question of the biological meaning of the increasing or decreasing miRNA levels. In a longitudinal expression analysis of 3 miRNAs on serum samples of 15 patients with colorectal cancer over a period of three years post surgery or after chemotherapy, the authors found that serum levels of miRNAs returned to normal levels after cancer resection or chemotherapy in the samples from patients with good prognosis [14]. However, our data for single patients show a strong fluctuation between the different time points making a biological interpretation difficult. The specific variations of miRNA levels over time in single patients may be due to a combination of factors that are related to the physiological state and the specific treatment response of each patient and it will be highly demanding to define the specific influence of any of these factors on a specific miRNA plasma level.

Nevertheless, variations of miRNA levels over time might be related to clinical endpoints such as the development of metastases. For example, we found a

higher number of miRNAs that were significantly changed in plasma levels between the time point TP1 before and the time points TP2 to TP8 after surgery in patients that did not develop metastases during the follow-up compared to the patients that developed metastases. These relationships between miRNA plasma levels would have gone undetected by an analysis at a single time point only. Besides a potential diagnostic value of altered miRNA levels, the changes observed in the present study might help to contribute to the understanding of systemic aspects associated with metastases. Overall, our data indicate more changes of miRNA levels in patients without metastases as compared to patients with metastases. This is not only true for the comparisons between the time point before surgery with all seven time points after surgery but also for the comparison between the time point before surgery with the first time point directly after surgery and also for the comparison of TP1 samples with the non-cancer control samples. As described above, the latter comparison identified more significantly altered miRNAs in patients without metastases as compared to patients with metastases, possibly indicating that the miRNome of patients that developed metastases is more similar to the miRNome of non-cancer controls than the miRNome of patients that do not develop metastases during the follow-up.

Nevertheless, we are aware of the limitations of the present study and thus do not intend to over-interprete our findings. For example we want to point out that we analyzed groups of different sizes, i.e., the group of patients that did not develop metastases encompassed 18 patients while we obtained only blood of eight patients that later on developed metastases. In addition, as discussed above, the choice of the right blood collections system is very crucial for downstream analyses. Furthermore, the here presented results have to be confirmed in larger patient cohorts in future studies.

Although highly hypothetical, our data may indicate that the overall miRNome of a patient that later develops metastases is less affected by surgery than the miRNome of a patient that is not prone to develop metastases. An overall stability of the miRNome has previously been reported for healthy adults by MacLellan et al. [13]. Possibly, such an overall stability can also be found for a pathological status and changes of the miRNA pattern would indicate either a treatment success or a significant deterioration of the patients' health.

MATERIALS AND METHODS

Study population

We obtained blood from 26 different NSCLC patients. Blood of lung cancer patients was drawn directly before tumor resection (TP1), around two weeks after tumor resection (TP2) and then around three months (TP3), six months (TP4), nine months (TP5), 12 months (TP6), 15 months (TP7) and 18 months (TP8) after tumor resection. From 3 patients we obtained only blood from 7 time points and from one patient we obtained blood only from 6 time points. In a follow-up of 4 years, 18 patients were free of metastases or recurrences. In addition, we obtained blood from 12 patients from the same clinic that did not suffered from lung cancer but from other non-tumor lung diseases. Blood of all patients was drawn in Lithium-Heparin monovettes (Sarstedt). Plasma was isolated by centrifugation at 3000 rpm for 10 min and stored at -80°C until use. Samples were collected with patient informed consent. The local Ethics Committee approved the study (Ärztekammer des Saarlandes, 01/08). Patient details are provided in Supplemental Table 5.

Isolation of total RNA including miRNA

As it is well known that heparin is co-purified with RNA and can interfere with downstream applications the RNA was isolated using an optimized protocol for Lithium-Heparin plasma samples as previously described [17]. We first treated 100µl plasma with 10µg Heparinase I (Sigma) and 100U RNaseOUT™ (Life Technologies) and incubated the mixture at 25°C for 1 hour. Nuclease free water (Life Technologies) was added to a final volume of 250µl. A total of 750µl TRIzol®LS (Life Technologies) was added and incubated at RT for 5 min. Then, 20µg glycogen, 5µl spike-in miRNA (miRNA mimic syn-cel-miR-39, 5nM, Qiagen) and 200µl chloroform were added, vigorously vortexed, and incubated for 3 min at RT. After centrifugation at 14000rpm and 4°C, the aquaeous phase was transferred into a new tube and RNA was precipitated with 1,5 volumes of 100% ethanol. RNA was then isolated using the miRNeasy Mini Kit (Qiagen) according to manufacturers instructions but with the use of the RNeasy Mini Elute column to allow for a reduced elution volume of 15µl. RNA concentration was measured using the Nanodrop2000 (ThermoScientific) and RNA quality was checked using the Bioanalyzer2100 and the Small RNA Kit (Agilent).

Quantitative real time PCR (qRT PCR)

Using quantitative Real Time-Polymerase Chain Reaction (qRT-PCR) with the miScript PCR System (Qiagen) we validated the microarray data for three exemplarily chosen miRNAs (hsa-miR-130b, hsa-miR-762, hsa-miR-197) and the follow-up samples from two patients that developed metastases and two patients that did not. In brief, 2 µl RNA was converted into cDNA using the miScript II Reverse Transcription Kit and the HiSpec Buffer according to the manufacturers´ protocol.

The PCR was performed with the miScript SYBR® Green PCR Kit in a total volume of 20µl per reaction containing 2µl (1:5 diluted) cDNA according to the manufacturers´ protocol on a StepOne Plus Real Time Analyzer (Life Technologies). Data were normalized using the spike-in miRNA mimic syn-cel-miR-39 (Qiagen).

miRNA microarray

Microarray analysis has been performed according to manufacturer's instructions and as previously described using SurePrint G3 8x60K miRNA microarrays (Agilent) [20]. In brief, a total of 100 ng total RNA was processed using the miRNA Complete Labeling and Hyb Kit (Agilent) to generate fluorescently (cyanine-3) labeled miRNA. The microarrays, that contain 40 replicates of each of the 1,205 miRNAs of miRBase v16 (http://www.mirbase.org/ [26]) were hybridized with the labeled miRNA for 20 hours at 55°C and 20rpm. Microarray scan data were further processed using Feature Extraction software (Agilent). The Feature Extraction software removes outlier pixels, does statistics on inlier pixels of features and backgrounds. It further flags outlier features and backgrounds and subtracts the background from features. The output of the Feature Extraction Software provides the raw background corrected miRNA data (gTotalGeneSignal) and the present calls (IsGeneDetected). The results of the microarray analyses are freely available in the GEO database under accession number GSE68951 (http://www.ncbi.nlm.nih.gov/geo/).

Biostatistics

All downstream biostatistics calculations have been carried out using the freely available statistical programming environment R. Two analysis strategies were carried out. First, we focused on the present calls, i.e. the information whether a miRNA *m* in patient *p* is expressed significantly above the background. This information was obtained from the Agilent feature extraction software according to manufacturers instruction and as sketched above. For all samples and miRNAs a binary matrix was build, where entries (*m,p*) equaled 1 if miRNA *m* was present in patient *p* and 0 otherwise. To minimize the noise contributed by low expressed markers we focused for all analyses on the miRNAs that were expressed above background in at least 5% of all tested samples. Using this definition, we performed all further analyses using 485 miRNAs.

In addition to the present call analysis, we likewise carried out a quantitative analysis of the expression level for the detected miRNAs. Since microarrays frequently show batch effects we tested and corrected for such technological bias. In detail, the identification and visualization of the batch effects was performed using the R-package "pvca". The ComBat function of the R-package "sva" was then applied in order to account for the found batch effects in the data. Quantil normalization has been carried out using the Bioconductor "preprocessCore" package. Pairwise two-tailed *t*-tests have been carried out. Here, each time point following resection has been compared to the time point prior to resection. The results have been displayed as circular diagrams, specifically, time points are ordered clockwise such that each time point has an own sector. The shading of the sector denotes the significance, the further the shading, the more significant the respective time point is for this miRNA. Moreover, correlation between time-points and expression or significance values have been calculated using Pearson Correlation coefficient and a significance value for each correlation has been calculated using the "cor.test" function. For assessing the significance of correlations we calculated a statistic based on Pearson's product moment correlation coefficient, which follows a t-distribution. Additionally, 90% Confidence Intervals for the correlation are provided, which are calculated based on Fishers Z Transform. If not mentioned explicitly, *p*-values have been adjusted for multiple testing using the Benjamini-Hochberg approach.

FINANCIAL SUPPORT

This study was funded by Deutsche Forschungsgemeinschaft (DFG LE2783/1-1) and in part by Siemens.

CONFLICTS OF INTEREST

CS is employee of Siemens.

REFERENCES

1. Siegel R, Naishadham D and Jemal A. Cancer statistics, 2013. CA Cancer J Clin. 2013; 63:11-30.
2. Padda SK, Burt BM, Trakul N and Wakelee HA. Early-stage non-small cell lung cancer: surgery, stereotactic radiosurgery, and individualized adjuvant therapy. Seminars in oncology. 2014; 41:40-56.
3. Hung JJ, Hsu WH, Hsieh CC, Huang BS, Huang MH, Liu JS and Wu YC. Post-recurrence survival in completely resected stage I non-small cell lung cancer with local recurrence. Thorax. 2009; 64:192-196.
4. Dai CH, Li J, Yu LC, Li XQ, Shi SB and Wu JR. Molecular diagnosis and prognostic significance of lymph node micrometastasis in patients with histologically node-negative non-small cell lung cancer. Tumour Biol. 2013; 34:1245-1253.
5. Martini N, Bains MS, Burt ME, Zakowski MF, McCormack P, Rusch VW and Ginsberg RJ. Incidence of local recurrence and second primary tumors in resected stage I

lung cancer. J Thorac Cardiovasc Surg. 1995; 109:120-129.

6. Martin J, Ginsberg RJ, Venkatraman ES, Bains MS, Downey RJ, Korst RJ, Kris MG and Rusch VW. Long-term results of combined-modality therapy in resectable non-small-cell lung cancer. J Clin Oncol. 2002; 20:1989-1995.
7. al-Kattan K, Sepsas E, Fountain SW and Townsend ER. Disease recurrence after resection for stage I lung cancer. Eur J Cardiothorac Surg. 1997; 12:380-384.
8. Westeel V, Choma D, Clement F, Woronoff-Lemsi MC, Pugin JF, Dubiez A and Depierre A. Relevance of an intensive postoperative follow-up after surgery for non-small cell lung cancer. Ann Thorac Surg. 2000; 70:1185-1190.
9. Saikumar J, Ramachandran K and Vaidya VS. Noninvasive Micromarkers. Clin Chem. 2014; 60:1158-1173.
10. Dogini DB, Pascoal VD, Avansini SH, Vieira AS, Pereira TC and Lopes-Cendes I. The new world of RNAs. Genetics and molecular biology. 2014; 37:285-293.
11. Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, Haas J, Ruprecht K, Paul F, Stahler C, Lang CJ, Meder B, Bartfai T, et al. A blood based 12-miRNA signature of Alzheimer disease patients. Genome Biol. 2013; 14:R78.
12. Keller A, Leidinger P, Steinmeyer F, Stahler C, Franke A, Hemmrich-Stanisak G, Kappel A, Wright I, Dorr J, Paul F, Diem R, Tocariu-Krick B, Meder B, et al. Comprehensive analysis of microRNA profiles in multiple sclerosis including next-generation sequencing. Mult Scler. 2014; 20:295-303.
13. MacLellan SA, MacAulay C, Lam S and Garnis C. Pre-profiling factors influencing serum microRNA levels. BMC clinical pathology. 2014; 14:27.
14. Chen J, Wang W, Zhang Y, Chen Y and Hu T. Predicting distant metastasis and chemoresistance using plasma miRNAs. Medical oncology. 2014; 31:799.
15. Aushev VN, Zborovskaya IB, Laktionov KK, Girard N, Cros MP, Herceg Z and Krutovskikh V. Comparisons of microRNA patterns in plasma before and after tumor removal reveal new biomarkers of lung squamous cell carcinoma. PLoS One. 2013; 8:e78649.
16. Le HB, Zhu WY, Chen DD, He JY, Huang YY, Liu XG and Zhang YK. Evaluation of dynamic change of serum miR-21 and miR-24 in pre- and post-operative lung carcinoma patients. Medical oncology. 2012; 29:3190-3197.
17. Leidinger P, Keller A, Backes C, Huwer H and Meese E. MicroRNA expression changes after lung cancer resection: a follow-up study. RNA biology. 2012; 9:900-910.
18. Mestdagh P, Hartmann N, Baeriswyl L, Andreasen D, Bernard N, Chen C, Cheo D, D'Andrade P, DeMayo M, Dennis L, Derveaux S, Feng Y, Fulmer-Smentek S, et al. Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. Nature methods. 2014; 11:809-815.
19. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, Peterson A, Noteboom J, O'Briant KC, Allen A, Lin DW, Urban N, Drescher CW, et al. Circulating microRNAs as stable blood-based markers for cancer detection. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105:10513-10518.
20. Kulasingam V and Diamandis EP. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. Nature clinical practice Oncology. 2008; 5(10):588-599.
21. Stephan C, Jung K, Lein M, Sinha P, Schnorr D and Loening SA. Molecular forms of prostate-specific antigen and human kallikrein 2 as promising tools for early diagnosis of prostate cancer. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2000; 9:1133-1147.
22. Lawrie CH, Gal S, Dunlop HM, Pushkaran B, Liggins AP, Pulford K, Banham AH, Pezzella F, Boultwood J, Wainscoat JS, Hatton CS and Harris AL. Detection of elevated levels of tumour-associated microRNAs in serum of patients with diffuse large B-cell lymphoma. Br J Haematol. 2008; 141:672-675.
23. Moldovan L, Batte KE, Trgovcich J, Wisler J, Marsh CB and Piper M. Methodological challenges in utilizing miRNAs as circulating biomarkers. Journal of cellular and molecular medicine. 2014; 18:371-390.
24. Summerer I, Niyazi M, Unger K, Pitea A, Zangen V, Hess J, Atkinson MJ, Belka C, Moertl S and Zitzelsberger H. Changes in circulating microRNAs after radiochemotherapy in head and neck cancer patients. Radiation oncology. 2013; 8:296.

analytical chemistry

Article

pubs.acs.org/ac

Influence of Next-Generation Sequencing and Storage Conditions on miRNA Patterns Generated from PAXgene Blood

Christina Backes,[†] Petra Leidinger,[‡] Gabriela Altmann,[§] Maximilian Wuerstle,[§] Benjamin Meder,[∥] Valentina Galata,[†] Sabine C. Mueller,[†] Daniel Sickert,[§] Cord Stähler,[§] Eckart Meese,[‡] and Andreas Keller*[,†]

[†]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany

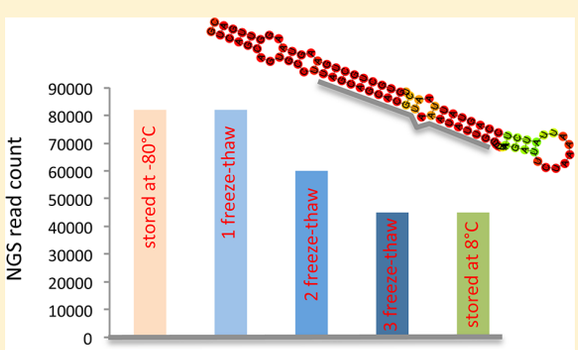
[‡]Institute of Human Genetics, Saarland University, D-66424 Homburg, Germany

[§]Siemens AG, 80333 Munich, Germany

[∥]Internal Medicine II, University Hospital Heidelberg, 69120 Heidelberg, Germany

S *Supporting Information*

ABSTRACT: Whole blood derived miRNA signatures determined by Next-Generation Sequencing (NGS) offer themselves as future minimally invasive biomarkers for various human diseases. The PAXgene system is a commonly used blood storage system for miRNA analysis. Central to all miRNA analyses that aim to identify disease specific miRNA signatures, is the question of stability and variability of the miRNA profiles that are generated by NGS. We characterized the influence of five different conditions on the genome wide miRNA expression pattern of human blood isolated in PAXgene RNA tubes. In detail, we analyzed 15 miRNomes from three individuals. The blood was subjected to different numbers of freeze/thaw cycles and analyzed for the influence of storage at −80 or 8 °C. We also determined the influence of blood collection and NGS preparations on the miRNA pattern isolated from a single individual, which has been sequenced 10 times. Here, five PAXGene tubes were consecutively collected that have been split in two replicates, representing two experimental batches. All samples were analyzed by Illumina NGS. For each sample, approximately 20 million NGS reads have been generated. Hierarchical clustering and Principal Component Analysis (PCA) showed an influence of the different conditions on the miRNA patterns. The effects of the different conditions on miRNA abundance are, however, smaller than the differences that are due to interindividual variability. We also found evidence for an influence of the NGS measurement on the miRNA pattern. Specifically, hsa-miR-1271-5p and hsa-miR-182-5p showed coefficients of variation above 100% indicating a strong influence of the NGS protocol on the abundance of these miRNAs.



For the identification of biomarkers and even more for the translation from basic research to clinical routine, it is crucial to understand how markers vary depending on different storage conditions and technical analysis. Especially for complex marker profiles like miRNA signatures, a systematic bias will compromise their diagnostic and prognostic values. While tissue based miRNA profiles have first been in the focus of research, there are increasing efforts to identify miRNA signatures as non- or minimally invasive markers in body fluids, such as blood, serum, or urine. Besides Heparin and EDTA blood tubes, PAXgene blood RNA tubes have frequently been used to collect patients' blood. Examples of PAXgene blood RNA pattern include biomarkers for myocardial infarction,[1] lung cancer,[2,3] multiple sclerosis,[4,5] melanoma,[6] ovarian cancer,[7] chronic obstructive pulmonary disease,[8] glioblastoma,[9] and Alzheimer's disease.[10] More recently, miRNA profiles of single blood cell types or exosomes have been accomplished.[11,12]

To obtain profiles of miRNAs, different technologies have been applied. In the early stages of miRNA profiling, microarrays have been widely used to generate miRNA patterns. High-throughput qRT-PCR platforms also enable the parallel measurement of hundreds of miRNAs. As the most recent technology, Next-Generation Sequencing (NGS) generates millions of short reads that can be aligned to known miRNAs annotated in the miRBase.[13] Likewise, new miRNA candidates can be predicted by aligning the fragments to the target genome. To facilitate clinical applications, other methods such as immunoassays are currently developed.[14] Technical stability of the profiles for reliable biomarker discovery is of high impact, independent of the applied screening technique. Since NGS is increasingly applied to

Received: June 1, 2015
Accepted: July 24, 2015

DOI: 10.1021/acs.analchem.5b02043
Anal. Chem. XXXX, XXX, XXX−XXX

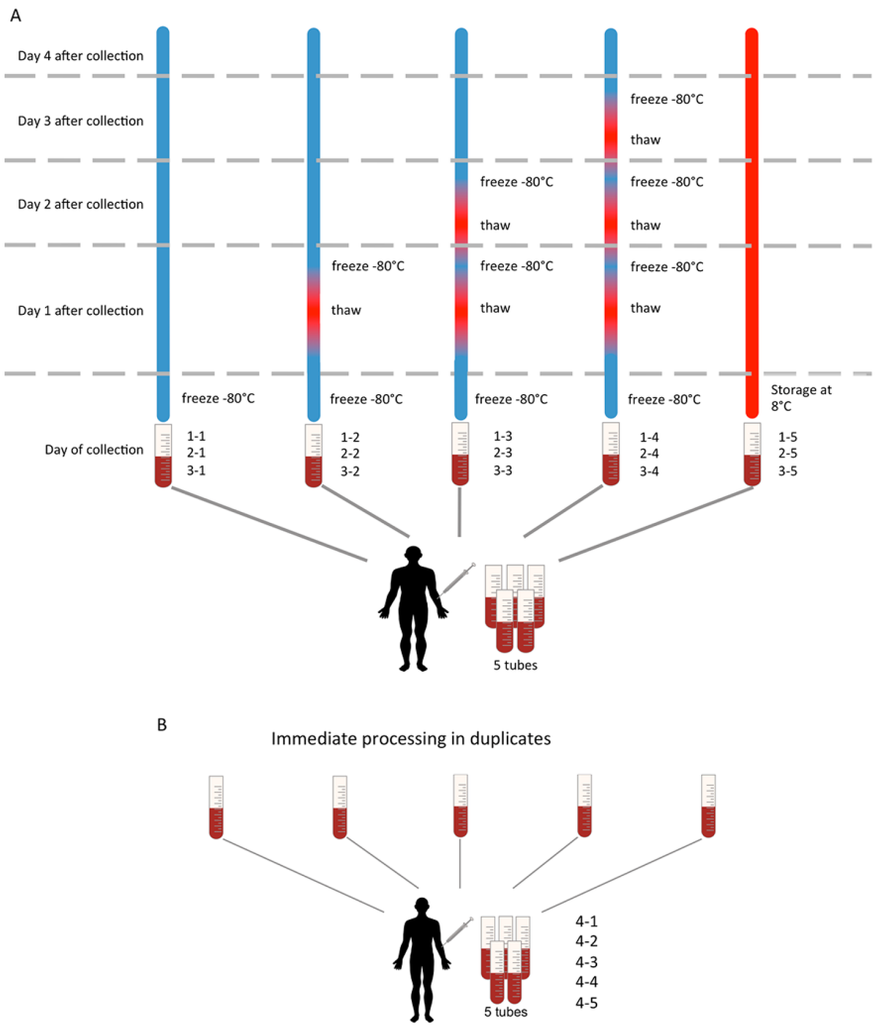


Figure 1. Study set up: 4 donors without known disease affection were included. For the first three donors, 5 blood tubes were extracted and processed over the next 4 days (panel A). For the fourth donor, 5 blood tubes were extracted and handled in duplicates (panel B). All patients are labeled by X-Y, where X is the patient number and Y the condition.

generate patient based miRNA signatures we investigated NGS-related variability and stability of miRNA pattern that are derived from PAXgene samples. In detail, we investigate three main question: First, markers are frequently discovered in retroperspective studies. Often, samples that are stored in biobanks are thawed and a part of the sample is used for measurement. We asked whether additional freeze/thaw cylces have a significant influence of miRNAs. Second, we asked how storage at 8 °C relates to the samples stored at −80 °C. The time was thereby restricted to 4 days. Third, we also investigated the influence of NGS on the variability/stability of miRNA patterns by performing two NGS batches echa containing technical replicates of five PAXgene blood tubes, which were all taken from the same individual.

MATERIALS AND METHODS

Study Setup and miRNA Profiling. In this study, we focused on the influence of freeze/thaw cycles and short time storage of PAXGene blood samples. We performed 25 miRNA measurements from 4 individuals. To minimize the influence of pathogenic processes healthy individuals without known diseases affection were investigated. All blood donors participating in this study gave their informed consent. For each of the first three individuals we collected 5 PAXgene Blood RNA tubes. The first tube has been stored at −80 °C for 4 days, while the second tube has been frozen at −80 °C and was subsequently subjected to one additional freeze/thaw cycle on the first day and finally stored again at −80 °C for the remaining days. The third tube has been subjected to an additional freeze/thaw cycle on the first day, frozen again at −80 °C, subjected to a second additional freeze/thaw cycle on the second day and stored at −80 °C for the remaining days. The fourth tube has been subjected to one freeze/thaw cycle on the first day, a second freeze/thaw cycle on the second day, a third additional freeze/thaw cycle on the third day and stored at −80 °C for the fourth day. The fifth tube has been stored at 8 °C for 4 days. The study setup is sketched in Figure 1A. For all samples independent miRNA isolation using the PAXgene Blood miRNA Kit and individual library preps using the Illumina TruSeq small RNA Library Prep Kit have been generated according to manufacturer's instruction.

To determine the influence of NGS on the miRNA pattern, NGS was performed on miRNAs isolated from a single

DOI: 10.1021/acs.analchem.5b02043
Anal. Chem. XXXX, XXX, XXX−XXX

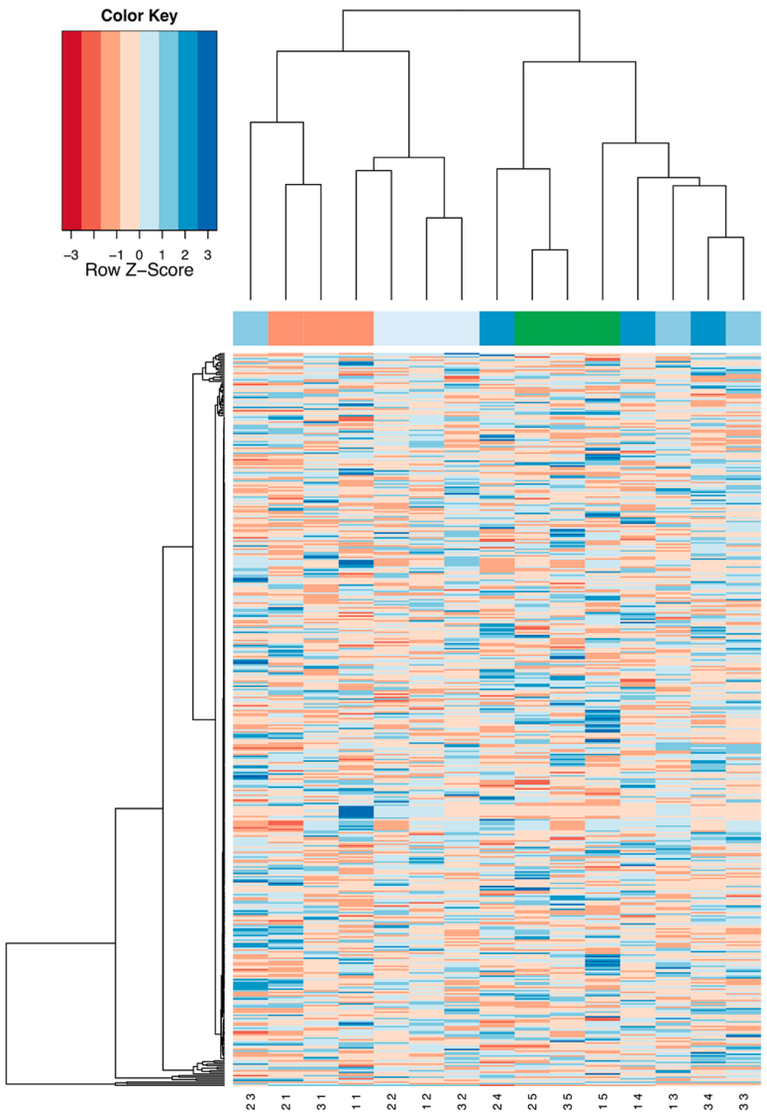


Figure 2. Cluster heat map with dendrogram on top and left. Color scale for the clustering heat map is provided in the upper left corner. The samples are colored as follows: The three different cycles are shown in three different blue shadings above the heat map, orange samples have been directly handled and the four-day refrigerator samples are colored in green. The three different cycles that are shown in three different blue shadings above the heat map mix between the orange samples that have been directly handled (clustering on the left side) and the four-day fridge control samples that are colored in green (clustering on the right side). This left cluster also contains the samples that have been frozen and thawed once while the right cluster contains the samples with three freeze/thaw cycles. PCA, which has been used to generate a 2-dimensional mapping of the high-dimensional miRNA profiles, confirmed these clustering results.

individual. We have taken five PAXgene tubes from this donor (storage at −80 °C), isolated the RNA and performed two batches of library preps and NGS runs on these five RNA eluates (10 miRNomes). An overview of the study setup is illustrated in Figure 1B.

For each of the 25 libraries, Illumina HiSeq2500 runs have been carried out according to manufacturer's instruction. For all samples around 20 Million raw sequencing reads have been generated. All miRNA extractions and sequencing runs have been carried out by CeGaT GmbH (Tübingen, Germany).

Bioinformatics Analysis. We preprocessed the raw sequencing data as described previously.[10] In brief, the reads were mapped against the current miRBase v21 sequences by using the miRDeep2 pipeline.[15] The raw miRBase counts for all samples were summarized in an expression matrix.

In order to carry out hierarchical clustering and calculate heat maps source code from the heat map.2 function, provided as part of the "gplots" CRAN package (version 2.12.1) has been used. In more detail, hierarchical clustering relying on the Euclidian distance has been carried out on quantile normalized data (normalization has been done by the "preprocessCore" package using standard parameters). As alpha level, 0.05 has been used through the manuscript. If not stated explicitly, p-values have been adjusted for multiple testing using the Benjamini−Hochberg approach.[16] Analysis of variance has been calculated by using the "anova" package.

RESULTS AND DISCUSSION

In total, we generated 25 miRNomes from four individuals by NGS. Fifteen miRNomes have been derived from three

DOI: 10.1021/acs.analchem.5b02043
Anal. Chem. XXXX, XXX, XXX−XXX

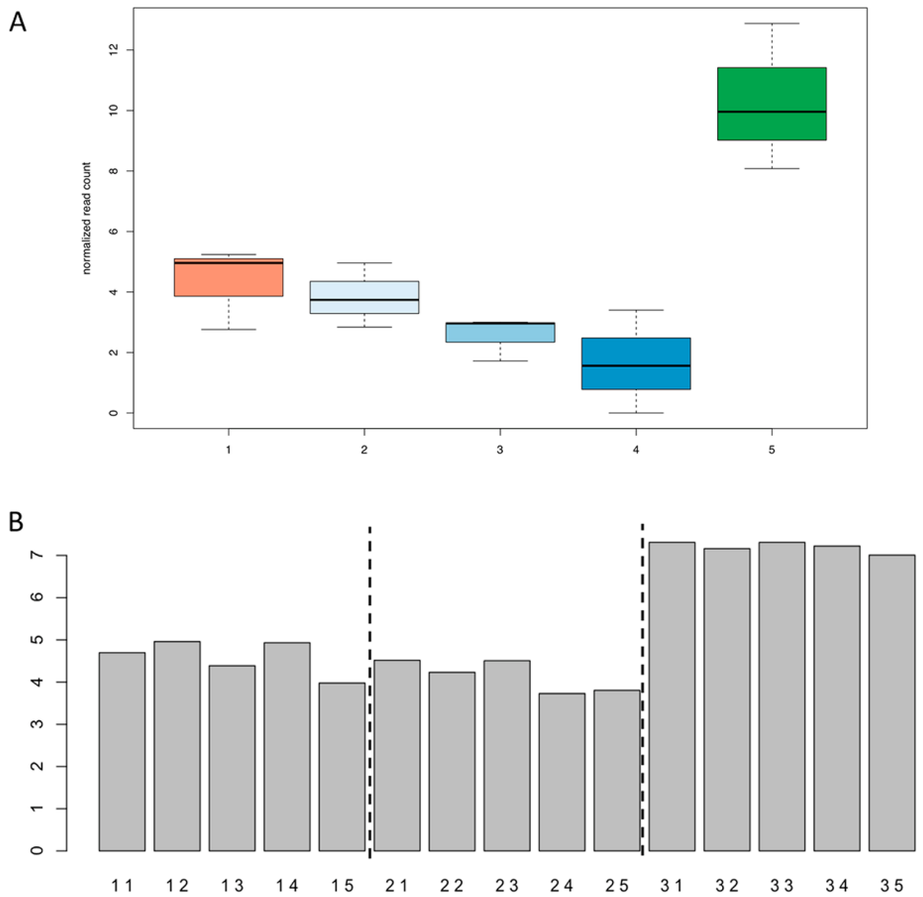


Figure 3. (A) Boxplot resulting from the ANOVA for conditions 1−5 and donors 1−3. For miR-375 signals increase for samples stored for 4 days at the refrigerator. The color scheme corresponds to Figure 2. (B) log of normalized read counts for conditions 1−5 and donors 1−3 for miR-99a-5p. The overall high variability is due to the overall higher expression of that miRNA in donor 3 as compared to the first two donors.

individuals to determine the influence of different storage conditions and freeze/thaw cycles (see Figure 1A) and from the fourth individual 10 NGS runs have been performed (5 consecutive blood drawings, two technical replicates that have been processed in batches, see Figure 1B).

The expression values of blood miRNAs showed a high dynamic range of approximately 7 orders of magnitude. Out of 2588 human miRNAs annotated in miRBase 21, we found 1252 different miRNAs in at least one of the 25 samples analyzed. 1060 miRNAs remained after markers covered by just a single read were removed. Regarding the distribution of NGS reads to miRNAs we observed significant variations. 90.3% of all reads matched to hsa-miR-486-5p and 5% to hsa-miR-92a-3p. The remaining 4.7% reads (approximately 20 million reads), matched to 1250 miRNAs. To minimize a potential bias introduced by very low abundant miRNAs, we empirically determined a threshold of 50 counts and continued the analysis with a remaining set of 455 miRNAs. All miRNAs with absolute read count and percentage of all reads mapping to this miRNA are summarized in Supporting Table 1.

Influence of Storage Conditions and Additional Freeze/Thaw Cycles on miRNA Patterns from Human Blood. To determine the effect of different storage conditions and additional freeze/thaw cycles on the miRNA pattern, we first employed cluster analysis and Principal Component Analysis (PCA) as two commonly applied statistical approaches, on the set of 455 miRNAs. In detail, we performed a complete linkage hierarchical clustering using the Euclidian distance as distance measure. The results for all miRNAs measured under five different conditions for three individuals, are summarized in Figure 2 as heat map with dendrograms on top for the storage conditions and on the left side for the miRNAs. The heat map indicates clustering of samples that have been stored throughout the experiment at −80 °C without additional freeze/thaw cycles (indicated by an orange color in the heat map). Likewise, the samples that have been stored at 8 °C for 4 days without changing the storage condition (indicated with green color in the heat map), cluster together as well. In addition, samples that have been stored at −80 °C without thawing cluster together with samples that have been stored at −80 °C with only one additional freeze/thaw cycle (indicated by an light blue color in the heat map). We also observed a clustering of samples that are stored at −80 °C with three additional freeze/thaw cycles (indicated by a dark blue color in the heat map) with samples that have been stored solely at 8 °C. The results may be indicative of an overall influence of the time period during which a sample is stored either at −80 °C or at 8 °C. The PCA, which has been used to generate a 2-dimensional mapping of the high-dimensional miRNA profiles, largely confirmed the clustering results.

After having investigated the systematic effects on the miRNA patterns, we analyzed whether single miRNAs show

DOI: 10.1021/acs.analchem.5b02043
Anal. Chem. XXXX, XXX, XXX−XXX

differences under the tested storage conditions. Therefor an analysis of variance (ANOVA) as well as the coefficient of variation (CV) have been applied. ANOVA identified 41 markers that were significant according to raw ANOVA p-values and an alpha level of 0.05. Since multiple markers were measured, the p-values had to be adjusted for multiple testing, resulting in 5 miRNAs being still significant, including hsa-miR-320b (p = 0.0002), hsa-miR-320a (p = 0.001), hsa-miR-16-5p (0.018), hsa-miR-18b-5p (0.037), and hsa-miR-375 (0.0375). As one example of a significant miRNA, has-miR-375 is presented as boxplot in Figure 3A. For miRNA-375, we found a significant difference between samples that have been stored at 8 °C for 4 days and samples that have been stored at −80 °C. We did not find a significant influence of the freeze/thaw cycles. All miRNAs along with the raw and adjusted p-values are presented in Supporting Table 2.

Finally, we addressed the question of the importance of the miRNA changes observed under the different storage conditions. To this end, we compared the differences between the three donors to the differences between the different storage conditions. An analysis of the coefficient of variation highlighted 37 miRNAs with the standard deviations exceeding the mean value (CV > 1). The mean value, standard deviation and CV for all miRNAs are presented in Supporting Table 3. Largest CV was calculated for miR-1291 (mean of 24.6, standard deviation of 40.7, CV of 1.7). One example is provided in Figure 3B, where the log of normalized read counts of the five different conditions for the three individuals is presented for miR-99a-5p. For this miRNA, the mean value is 508, the standard deviation 624 and the CV 1.23. The variation of the five measurements for each of the individuals is substantially smaller than the deviation between individuals 1 and 2 compared to individual 3, which had high read counts for this miRNA. In general, we observed that the variability of miRNA abundance between different donors was higher than the miRNA variability under different storage conditions.

We conclude that there is a general systematic influence of the storage conditions on miRNA patterns. Largest variability was observed between storage at 8 °C and samples stored at −80 °C without additional freeze/thaw cycles. The specific effect of the storage conditions has to be verified for each single miRNAs separately. Importantly, the effects of storage conditions on miRNA abundance are generally smaller than the differences due to inter donor variability.

While we investigated a short time period, the long time storage of samples has also been investigated. Seelenfreund and co-workers reported that miRNA from PAXGene tubes can be recovered even after periods of up to 4 years, if samples are frozen at −80 °C.[17] In this study, a subset of all known miRNAs analyzed by qRT-PCR has been included. Viprey et al. considered even longer time periods of up to 5 years.[18] In detail they evaluated the reliability of expression for a subset of 377 miRNAs by qRT-PCR. The authors did not observe a correlation of miRNA abundance with storage time. As most stable reference miRNAs, miR-26a, miR-28−5p, and miR-24 were identified. These miRNAs were also not affected significantly in our study with respect to freeze/thaw cycles, indeed rendering them as reasonable and stable reference markers.

Influence of NGS Preparation on miRNA Patterns from Human Blood. As mentioned above, two NGS runs with different library prep were performed with each on miRNAs isolated from five PAXgene tubes from a single individual and thus resulting in 10 miRNomes. The five samples of the first NGS run have been processed together, and likewise the samples of the second NGS run with a slightly changed preprocessing step as mentioned in material and methods. The results of the analysis are summarized as heat map in Figure 4. The dendrogram on top of the heat map show

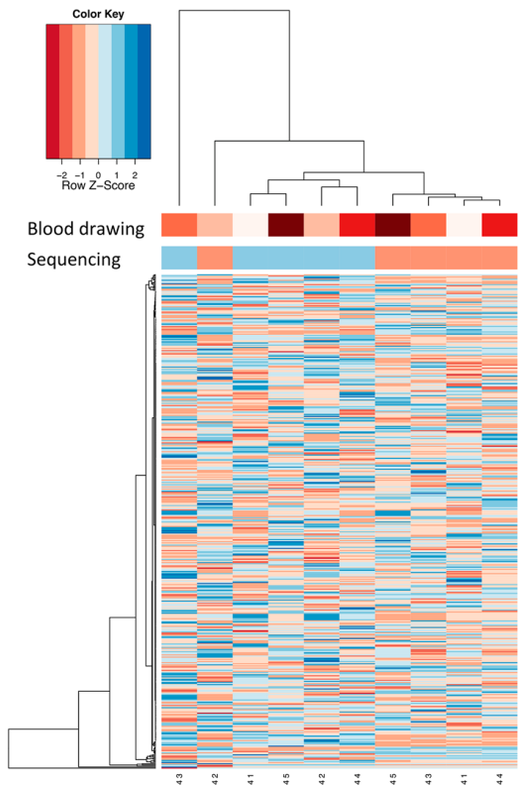


Figure 4. Cluster heat map with dendrogram on top. The samples are colored with respect to the blood drawing (first row on top of the heat map) and according to the analytical NGS batch (second row on top of the heat map).

the clustering of different blood tubes and sequencing procedures, while the dendrogram on the left side for the miRNAs. The heat map indicates no clustering between the five different blood sample tubes (indicated by five shadings of red). However, a strong clustering between the two different NGS preparations is shown (indicated by an orange color for the first NGS run and a blue color for the second NGS run). These results show a significant influence of the NGS preparation on the miRNA pattern identified in human blood. The findings from this cluster analysis were confirmed by a PCA.

We next addressed the question of importance miRNA changes observed for the NGS preparations by analyzing the coefficient of variation. The analysis showed lower CV values for the 10 miRNomes measured by two NGS preparations than for the miRNome obtained under different storage conditions. The standard deviation exceeded the mean for only 21 miRNAs the majority of which were low abundant. Only two miRNAs with expression levels (normalized read counts >5) showed average CV values >1 including hsa-miR-182−5p with average value of 2,824 and standard deviation of 2,960 resulting in a CV of 1.05 and hsa-miR-1271−5p with average read count of 5.7 and standard deviation of 6.2 resulting in a CV of 1.07. As

DOI: 10.1021/acs.analchem.5b02043
Anal. Chem. XXXX, XXX, XXX−XXX

indicated by the barplot of logarithmized normalized read counts in Figure 5A, the second batch of NGS library

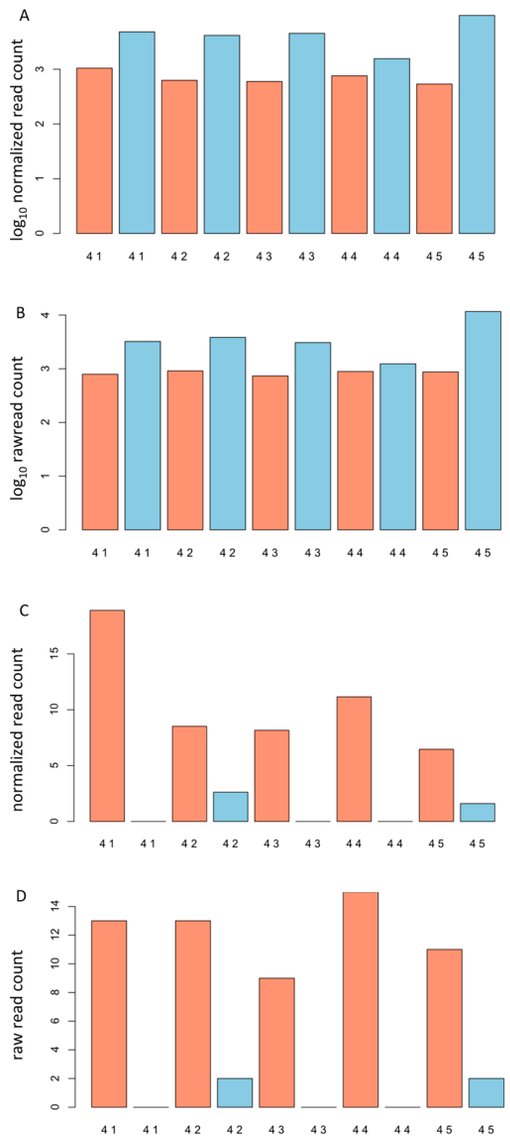


Figure 5. Normalized (Panel A) and raw (panel B) read counts for hsa-miR-182−5p. Shown are logarithmized reads for five replicated blood drawings (denoted 41, 42, 43, 44, and 45) in two NGS batches. The first batch is highlighted in orange the second in blue. Panels C and D present the same analysis for the second variable miRNA, has-miR-1271−5p (please note the absolute scale in contrast to the logarithmic for panels A and B).

preparation (indicated in blue) showed for hsa-miR-182−5p significantly higher expression level as the first NGS library preparation (indicated in orange). Notably, an analysis of raw read counts revealed the same behavior indicating that the differences between the two NGS batches are not due to the normalization process (Figure 5B). The same values are presented in Figure 5C and 5D for miR-1271−5p. This miRNAs was almost not present in the second NGS batch while substantially expressed in the first batch. The CV values, mean and standard deviations for all miRNAs are presented in Supporting Table 4.

In summary, we found evidence for an influence of the NGS measurement on specific miRNAs' profiles including the library preparation.

CONCLUSION

In this study, we systematically explored the influence of different conditions and freeze/thaw cycles on miRNA profiles generated by using NGS. Furthermore, we investigated the stability and reproducibility of the respective miRNA patterns by carrying out 10 replicated measurements of the same individual.

For selected miRNAs, we found an influence with respect to up to three additional freeze thaw cycles. Directly processed samples showed overall closest proximity to samples undergoing one freeze/thaw cycle. Interpreting the replicated measurements of the same donor also revealed a certain degree of variability. Specifically, we observed that the influence of the NGS procedure of miRNAs seems to be partially exceed the variability of the blood drawing and miRNA extraction. These results were determined by considering all miRNAs for the analysis, with techniques such as clustering and PCA as well as considering the coefficient of variation. However, the variability was only observed for a subset of all miRNAs. It is therefore essential to be aware of the potential pitfalls of sample storage and NGS preparation that can contribute to the variability of miRNAs. If specific case control studies present these miRNAs as candidates to detect a disease, further in-depth validation is required.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.5b02043.

All miRNAs with absolute read count and percentage of all reads mapping to this miRNA (XLXS)

All miRNAs along with the raw and adjusted *p*-values (XLXS)

Mean value, standard deviation, and CV for all miRNAs (XLXS)

CV values, mean, and standard deviations for all miRNAs (XLXS)

AUTHOR INFORMATION

Corresponding Author

*E-mail: andreas.keller@ccb.uni-saarland.de. Phone: +49 681 302 68611.

Author Contributions

C.B., P.L., V.G., S.M. contributed in data analysis, G.A. contributed in study set up, coordinated experiments, evaluated data, M.W. contributed in manuscript preparation and writing, D.S. contributed in study design, C.S. contributed in study design and set up, contributed in manuscript preparation, E.M. wrote the manuscript, and A.K. wrote the manuscript and contributed in data analysis.

Notes

The authors declare no competing financial interest.

DOI: 10.1021/acs.analchem.5b02043
Anal. Chem. XXXX, XXX, XXX−XXX

■ ACKNOWLEDGMENTS

This work has been funded in parts by internal funds of Saarland University, by Siemens Healthcare and by the EU FP7 Project BestAgeing.

■ REFERENCES

(1) Meder, B.; Keller, A.; Vogel, B.; Haas, J.; Sedaghat-Hamedani, F.; Kayvanpour, E.; Just, S.; Borries, A.; Rudloff, J.; Leidinger, P.; Meese, E.; Katus, H. A.; Rottbauer, W. *Basic Res. Cardiol.* **2011**, *106*, 13−23.
(2) Keller, A.; Leidinger, P.; Borries, A.; Wendschlag, A.; Wucherpfennig, F.; Scheffler, M.; Huwer, H.; Lenhof, H. P.; Meese, E. *BMC Cancer* **2009**, *9*, 353.
(3) Leidinger, P.; Backes, C.; Blatt, M.; Keller, A.; Huwer, H.; Lepper, P.; Bals, R.; Meese, E. *Mol. Cancer* **2014**, *13*, 202.
(4) Keller, A.; Leidinger, P.; Lange, J.; Borries, A.; Schroers, H.; Scheffler, M.; Lenhof, H. P.; Ruprecht, K.; Meese, E. *PLoS One* **2009**, *4*, e7440.
(5) Keller, A.; Leidinger, P.; Steinmeyer, F.; Stahler, C.; Franke, A.; Hemmrich-Stanisak, G.; Kappel, A.; Wright, I.; Dorr, J.; Paul, F.; Diem, R.; Tocariu-Krick, B.; Meder, B.; Backes, C.; Meese, E.; Ruprecht, K. *Multiple Sclerosis Journal* **2014**, *20*, 295.
(6) Leidinger, P.; Keller, A.; Borries, A.; Reichrath, J.; Rass, K.; Jager, S. U.; Lenhof, H. P.; Meese, E. *BMC Cancer* **2010**, *10*, 262.
(7) Hausler, S. F.; Keller, A.; Chandran, P. A.; Ziegler, K.; Zipp, K.; Heuer, S.; Krockenberger, M.; Engel, J. B.; Honig, A.; Scheffler, M.; Dietl, J.; Wischhusen, J. *Br. J. Cancer* **2010**, *103*, 693−700.
(8) Leidinger, P.; Keller, A.; Borries, A.; Huwer, H.; Rohling, M.; Huebers, J.; Lenhof, H. P.; Meese, E. *Lung cancer* **2011**, *74*, 41−47.
(9) Roth, P.; Wischhusen, J.; Happold, C.; Chandran, P. A.; Hofer, S.; Eisele, G.; Weller, M.; Keller, A. *J. Neurochem.* **2011**, *118*, 449−457.
(10) Leidinger, P.; Backes, C.; Deutscher, S.; Schmitt, K.; Mueller, S. C.; Frese, K.; Haas, J.; Ruprecht, K.; Paul, F.; Stahler, C.; Lang, C. J.; Meder, B.; Bartfai, T.; Meese, E.; Keller, A. *Genome biology* **2013**, *14*, R78.
(11) Leidinger, P.; Backes, C.; Dahmke, I. N.; Galata, V.; Huwer, H.; Stehle, I.; Bals, R.; Keller, A.; Meese, E. *Oncotarget* **2014**, *5*, 9484−9487.
(12) Leidinger, P.; Backes, C.; Meder, B.; Meese, E.; Keller, A. *BMC Genomics* **2014**, *15*, 474.
(13) Kozomara, A.; Griffiths-Jones, S. *Nucleic Acids Res.* **2011**, *39*, D152−157.
(14) Kappel, A.; Backes, C.; Huang, Y.; Zafari, S.; Leidinger, P.; Meder, B.; Schwarz, H.; Gumbrecht, W.; Meese, E.; Staehler, C. F.; Keller, A. *Clin. Chem.* **2015**, *61*, 600−607.
(15) Friedländer, M. R.; Mackowiak, S. D.; Li, N.; Chen, W.; Rajewsky, N. *Nucleic Acids Res.* **2012**, *40*, 37.
(16) Benjamini, Y.; Hochberg, Y. *Journal of the Royal Statistical Society. Series B. Methodological* **1995**, *57*, 289−300.
(17) Seelenfreund, E.; Robinson, S. E.; Amato, C. M.; Bemis, L. T.; Robinson, W. A. *Biopreserv. Biobanking* **2011**, *9*, 29−33.
(18) Viprey, V. F.; Corrias, M. V.; Burchill, S. A. *Anal. Biochem.* **2012**, *421*, 566−572.

DOI: 10.1021/acs.analchem.5b02043
Anal. Chem. XXXX, XXX, XXX−XXX

analytical chemistry

Article

pubs.acs.org/ac

Double-Stranded Ligation Assay for the Rapid Multiplex Quantification of MicroRNAs

Stefan Hofmann,[†] Yiwei Huang,[‡] Peter Paulicka,[‡] Andreas Kappel,[‡] Hugo A. Katus,[§] Andreas Keller,[∥] Benjamin Meder,[§] Cord Friedrich Stähler,[⊥] and Walter Gumbrecht*[,‡]

[†]Department of Bioinformatics, University of Würzburg, Würzburg, 97074, Germany

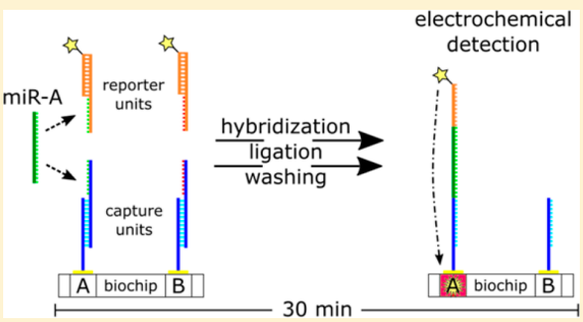
[‡]Technology Center, Siemens Healthcare, Erlangen, 91058, Germany

[§]Internal Medicine III, University Hospital Heidelberg, Heidelberg, 69120, Germany

[∥]Clinical Bioinformatics, Medical Faculty, Saarland University, Saarbrücken, 66123, Germany

[⊥]Strategy, Siemens Healthcare, Erlangen, 91052, Germany

ABSTRACT: MicroRNAs are auspicious candidates for a new generation of biomarkers. The detection of microRNA panels in body fluids promises early diagnosis of many diseases, including cancer or acute coronary syndrome. For a fast, sensitive, and specific detection of microRNA panels on the bedside, medical point-of-care systems that measure those biomarkers are required. As microchips are promising technical tools for a robust signal measurement at biochemical interfaces, we developed an assay for the electrochemical multiplex quantification of microRNAs on a CMOS chip with interdigitated gold electrode sensor positions. The method is based on the formation of a tripartite hybridization complex and subsequent both-sided ligation of the target nucleic acid to a reporter and capture strand. With a time to results of 30 min, the reported assay achieves a limit of detection below 1 pM, at a specificity down to single mismatch discrimination. It also offers very good signal dynamics between 1 pM and 1 nM, thus, allowing reliable quantification of the detected microRNAs and easy implementation into automated devices due to a simple workflow.



MicroRNAs (miRNAs) are short, noncoding transcripts of 18−24 bases that play an important role in the regulation of gene expression.[1,2] Especially the utilization of blood-borne miRNAs as noninvasive biomarkers is a promising field for new medical applications.[3−5] As miRNA levels within the body fluids of patients can be used to diagnose diseases like different types of cancer or heart disease,[6−9] diagnostic techniques for fast, cheap and unbiased quantification of miRNAs need to be developed.

Current detection technologies fail to meet all of the criteria needed for the broad application of miRNA-based diagnostic assays in medicine. In particular, most methods that are able to deliver valuable data about clinically relevant miRNAs, like next-generation sequencing, quantitative real-time PCR (qRT-PCR) or microarray, are time-consuming, require amplification or are costly.[10]

Electrochemical detection approaches promise a robust and cost-effective alternative to optical techniques and are therefore designated to be used in integrated medical devices for point-of-care (POC) diagnostics.[11] Microchip modules offering an array of sensor spots with interdigitated gold electrodes have been combined with gold−thiol coupled capture probes and an electrochemically active reporter enzyme product, to form a powerful measurement system for the detection of viral DNA, bacterial RNA and PCR products, as reported previously.[12−14] In this article we present a new miRNA quantification assay format, which leverages this detection mechanism. The assay format is based on hybridization and subsequent ligation of the target miRNA to an immobilized capture-component and a label enzyme-reporter conjugate. The reported method is well suited for application in POC diagnostics, as it is very fast and free of target amplification or prelabeling.

MATERIALS AND METHODS

Chip Module. The CMOS microchips used in this study were supplied by Siemens Corporate Technology (CT), Erlangen, Germany. These monolithically integrated silicon chips presented an array of 16 × 8 sensor positions (spots) on the surface (Figure 1A), with each spot encircled by a polymeric ring structure (Figure 1B). The microchips were implemented into a sealing compound forming a cavity to serve as an interface between the sensor array and the reaction solutions. The sensors comprised two interdigitated gold electrodes for the generation and detection of a *p*-aminophenol

Received: July 28, 2015
Accepted: November 17, 2015
Published: November 17, 2015

DOI: 10.1021/acs.analchem.5b02850
Anal. Chem. 2015, 87, 12104−12111

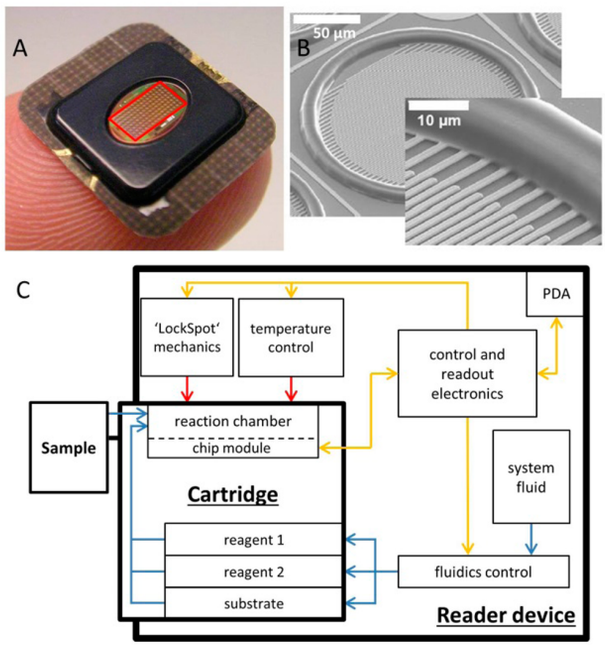


Figure 1. CMOS chip module and schematic overview of experimental setup. (A) Picture of a chip module in size comparison with a fingertip. The red frame marks the sensor array with 16 × 8 positions, which is encircled by the black sealing compound to form a reaction chamber when the module is covered by a flat gasket. (B) Detailed view of a single sensor position showing the interdigitated electrodes surrounded by a polymeric ring. (C) Schematic drawing of the experimental setup with the cartridge and the reader device as the main components comprising several functional units. Blue arrows represent fluidic connections, red arrows mechanical and thermal effects, yellow arrows indicate data transfer and electrical control. PDA: personal digital assistant.

(pAP) redox cycling process as has been described previously.[15,16]

Experimental Setup for CMOS Chip Measurements. A fully integrated reader device prototype and suitable cartridges with fluidic channels and reagent reservoirs were provided by Siemens CT, Erlangen. The cavity of the chip module was covered with a polydimethylsiloxane (PDMS; SYLGARD 184) gasket to form a reaction chamber which was connected to the fluidic channels of the cartridge after assembly. Following the introduction of the sample the cartridge was inserted into the reader device, which included a pump and valves to control the cartridge fluidics, a Peltier element for thermal regulation of the reaction chamber, contacts for communication with the microchip, corresponding electronics and a personal digital assistant with control software offering a graphical user interface and measurement data storage (Figure 1C). Additionally, a mechanism was embedded that could press the PDMS gasket onto the chip surface with a defined pressure in order to lock the spots of the sensor array during signal acquisition ("LockSpot"). This procedure increases the measurement signal by reducing the volume of the redox cycling reaction chamber above the interdigitated electrodes and prevents cross-talk between the sensor spots.[17,18]

Nucleic Acids. All synthetic oligonucleotides used in this study are listed in Tables 1–4. Capital letters represent DNA-, lowercased letters RNA-bases. The Esterase 2 reporter conjugate was synthesized as described by Wang et al.[19] The spike-in miRNA has an artificial alien sequence, which has no BLAST match in the *Homo sapiens* RefSeq RNA database,[20] to be applicable in measurements of endogenous RNA samples.

Endogenous total RNA including miRNAs was extracted from blood donor samples collected in PAXgene tubes as described by Keller et al.[21] The collection and use of human samples was approved by the Institutional Ethics Committee of the University Erlangen-Nuremberg, Germany.

Table 1. Synthetic miRNA Targets (RNA)

name	sequence (5′ → 3′)
miR-191	phosphate-caa cgg aau ccc aaa agc agc ug
miR-145	phosphate-guc cag uuu ucc cag gaa ucc cu
miR-181a	phosphate-aac auu caa cgc ugu cgg uga gu
miR-425	phosphate-aau gac acg auc acu ccc guu ga
miR-636	phosphate-ugu gcu ugc ucg ucc cgc ccg ca
miR-15a	phosphate-uag cag cac aua aug guu ugu g
miR-30c	phosphate-ugu aaa cau ccu aca cuc uca gc
miR-362	phosphate-aau ccu ugg aac cua ggu gug agu
spike-in	phosphate-aga ucg cca uac ccu gga gau a
let-7a	phosphate-uga ggu agu agg uug uau agu u
let-7b	phosphate-uga ggu agu agg uug ugu ggu u
let-7c	phosphate-uga ggu agu agg uug uau ggu u
let-7f	phosphate-uga ggu agu aga uug uau agu u

Table 2. Immobilization Strands (DNA/RNA Chimeras)

name	sequence (5′ → 3′)[a]
IS 1	thiol-T_6-CAG GAC GAT GAT GGc acg
IS 2	thiol-T_6-GAC CCA GCT CGT AGa ccg
IS 3	thiol-T_6-CGA CGA TAG CTT GGu acg
IS 4	thiol-T_6-TCA ACT TGT GCA GCc agc
IS 5	thiol-T_6-CAC GTC AGA CAG CTc cag
IS 6	thiol-T_6-CTT CTC GGT GTC CAc agg
IS 7	thiol-T_6-ACG TGT CTT CCG ctc g
IS 8	thiol-T_6-TAG GCT GAT GCC gca a
IS 9	thiol-T_6-GAG TCA CCT GCG CTg aac
IS 10	thiol-T_6-GCT AGA GCT GCG guc g

[a] T_6: T base spacer.

Table 3. Specific Capture Strands (DNA)

name	sequence (5′ → 3′)	complement
SCS-miR-191	GGA TTC CGT TGC GTG CCA TCA TCG TCC TG	IS 1
SCS-miR-181a	GCG TTG AAT GTT CGG TCT ACG AGC TGG GTC	IS 2
SCS-miR-15a	ATG TGC TGC TAC GTA CCA AGC TAT CGT CG	IS 3
SCS-miR-425	GAT CGT GTC ATT GCT GGC TGC ACA AGT TGA	IS 4
SCS-miR-145	GGA AAA CTG GAC CTG GAG CTG TCT GAC GTG	IS 5
SCS-miR-30c	AGG ATG TTT ACA CCT GTG GAC ACC GAG AAG	IS 6
SCS-miR-636	GAC GAG CAA GCA CAC GAG CGG AAG ACA CGT	IS 7
SCS-miR-362	GTT CCA AGG ATT TTG CGG CAT CAG CCT A	IS 8
SCS-spike-in	TAT GGC GAT CTG TTC AGC GCA GGT GAC TC	IS 9
SCS-negative	GTA CCG ATC CTA CGA CCG CAG CTC TAG C	IS 10
SCS-let-7a	TAC TAC CTC AGC TGG CTG CAC AAG TTG A	IS 4

DOI: 10.1021/acs.analchem.5b02850
Anal. Chem. 2015, 87, 12104–12111

Table 4. Reporter Conjugate and Specific Reporter Strands (DNA)

name	sequence (5′ → 3′)[a]	complement
RC-Est2	phosphate-GCA ACG AGC GC-T_4-Esterase2	
SRS-miR-191	GGT TGC GCT CGT TGC CAG CTG CTT TTG	RC-Est2
SRS-miR-181a	GGT TGC GCT CGT TGC ACT CAC CGA CA	RC-Est2
SRS-miR-15a	GGT TGC GCT CGT TGC CAC AAA CCA TT	RC-Est2
SRS-miR-425	GGT TGC GCT CGT TGC AGG GAT TCC TG	RC-Est2
SRS-miR-145	GGT TGC GCT CGT TGC AGG GAT TCC TG	RC-Est2
SRS-miR-30c	GGT TGC GCT CGT TGC GCT GAG AGT GT	RC-Est2
SRS-miR-636	GGT TGC GCT CGT TGC TGC GGG CGG	RC-Est2
SRS-miR-362	GGT TGC GCT CGT TGC ACT CAC ACC TAG	RC-Est2
SRS-spike-in	GGT TGC GCT CGT TGC TAT CTC CAG GG	RC-Est2
SRS-let-7a	GGT TGC GCT CGT TGC AAC TAT ACA ACC	RC-Est2

[a] T_4: T-base spacer.

Oligonucleotide Annealing. For the capture units, 20 μM each of the immobilization strand (IS) and the complement specific capture strand (SCS) were added to 300 mM NaCl, 10 mM $MgCl_2$, and 0.02% Tween 20 in 50 mM Tris/HCl at pH 7.6, incubated at 60 °C for 2 min and slowly cooled down to room temperature.

In case of the reporter units, for each nucleic acid target 0.7 μM of the Esterase 2 reporter conjugate (RC-Est2) and 1 μM of the target specific reporter strand (SRS) were annealed using the same procedure.

Chip Spotting. Immobilization of the capture units on the chip surface was done with a SciFlexArrayer S5 (Scienion) spotting system. Spotting solutions were made of 10 μM capture double-strand diluted in 3× SSC buffer pH 8.0, 1.5 M betaine, and 100 μM TCEP. The cleaned gold electrodes of each sensor position were covered with 1.2 nL of the respective spotting solution, with each of the 16 columns of the sensor array forming an eight spot detection cluster for one miRNA target. The capture units for let-7a, spike-in, and negative control covered three clusters each that were distributed on the chip surface. After an incubation time of 2 h at room temperature and 50% humidity the chip modules were washed with ultrapure water, blocked with The Blocking Solution (Candor Bioscience) for 15 min in a humidity chamber, dried and stored in a N_2 atmosphere until further use.

MiRNA Detection Assay. A quick ligation buffer was prepared for the ligation reaction containing 50 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT, 1 mM ATP, and 7.5% PEG-6000 in 66 mM Tris/HCl buffer at pH 7.6.[22] The esterase 2 substrate *p*-aminophenyl butyrate (pAPB) was synthesized as described by Wang et al.[19] The utilized cartridge offered reagent reservoirs of 70 (ligation solution), 60 (low salt buffer), and 240 μL (substrate reagent), respectively. Unless otherwise stated, 1× SSC buffer was used as system fluid, which was also utilized to rinse the reaction chamber between assay steps. The applied volume of sample solution was 60 μL (exception: 10 ng/μL measurement in Figure 7A: 48 μL). The assay protocol is described in Figure 2 and the corresponding figure legend.

Data Recording and Analysis. During measurement, the reader device digitally recorded a single data point per sensor spot every 0.5 s. For that purpose the electrical currents received from both interdigitated electrodes were automatically summarized by addition of the absolute values. We analyzed these raw data using Labview 2011 software (National Instruments). The slope (ΔI) was calculated as described in

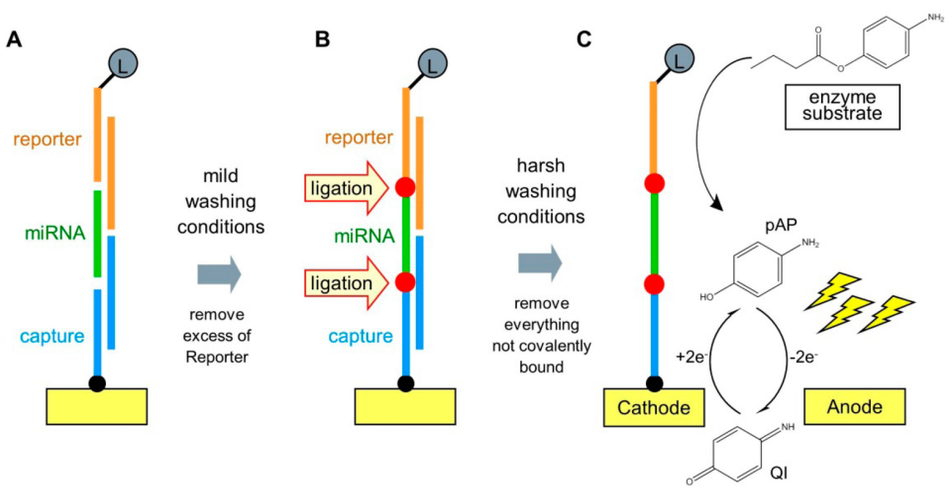


Figure 2. Schematic assay principle and protocol. Hybridization (A): The sample solution containing the target nucleic acids, 5.8 nM of each Esterase 2 reporter unit and 0.05% Tween 20 in 5x SSC buffer was drawn over the chip surface in three portions and incubated for 5 min each to form a tripartite complex with the immobilized capture units. The reaction chamber was washed with system fluid to remove any excess of reporter. Ligation (B): T4 DNA Ligase (100 u/ml, Thermo Scientific) was applied to the array in quick ligation buffer for 5 min. Then the reaction chamber was washed with low salt buffer (2 mM NaCl in 2 mM Tris/HCl pH 7.6) for 5 min to get rid of all assay components not covalently bound by ligation. Measurement (C): 1 mM pAPB enzyme substrate in 20 mM NaCl, 20 mM Tris/HCl pH 7.6 was pumped on the array surface and the sensor spots were locked for data acquisition. pAPB was converted to pAP by surface bound Esterase 2 enzymes and subsequent redox cycling at the interdigitated gold electrodes was measured. All assay steps were executed at 37 °C: L, label enzyme (esterase 2); pAP, *p*-aminophenol; QI, quinoneimine.

DOI: 10.1021/acs.analchem.5b02850
Anal. Chem. 2015, 87, 12104−12111

the General Principle. For each target, the ΔI values gained from all corresponding spots in three consecutive measurements at the end of one assay run (technical replicates) were used to calculate the median and the median absolute deviation (MAD). The median was preferred over the arithmetic mean to eliminate outliers caused by irregularities in the employed materials (sensor electrodes, surface of PDMS gasket). Correction of calculated values by negative control was done by subtraction of the median of the negative signals from the respective median and addition of the MAD of the negative signals to the corresponding MAD value.

miRNA qRT-PCR Measurement. qRT-PCR quantification of miRNA was performed using assays and accompanying reagents from Life Technologies. Life Technologies reagents: Taqman microRNA RT kit (Cat. No. 4366597), Taqman universal MMIX II with UNG (Cat. No. 4440045), Taqman microRNA assays INV (Cat. No. 4427975, INV 002299 for miR-191-5p, INV000480 for miR-181a-5p, INV 000389 for miR-15a-5p, INV001516 for miR-425-5p, and INV 000419 for miR-30c-5p). Measurements were performed on the Stratagene MX-3005p Real-Time PCR System (Agilent Technologies) according to the manufacturers' instructions. Standard curves with concentrations from 1 pM to 10 nM were generated for each target miRNA to calculate the molar concentrations of the endogenous miRNAs in the analyzed total RNA sample.

Calculation of Endogenous miRNA Concentrations. Standard curves for the CMOS chip assay were measured with synthetic miRNAs from 1 to 100 pM total concentration in the sample solution. The equation of the relationship between the signal ΔI and the target concentration was calculated using a linear trend line. The endogenous miRNA was measured employing chip modules from the same immobilization batch. The target concentrations in the total sample solution were calculated using the respective equation. The volume fraction of the total RNA sample in the sample solution was incorporated in the calculation to obtain the original target concentration.

RESULTS AND DISCUSSION

General Principle. The presented method makes use of a generic esterase 2-coupled reporter oligonucleotide (reporter conjugate, RC-Est2)[19] and a set of thiol-modified chimeric capture immobilization strands (IS). Additionally complementary counter-strands with overhangs specific for the target miRNAs (specific reporter/capture strands, SRS/SCS) are required to form preannealed double-stranded reporter and capture units. The sequences of all assay components used in this study can be found in Tables 2−4. Prior to the assay procedure, the capture double-strands had been immobilized on the interdigitated electrodes of their respective sensor positions by gold−thiol coupling.

In the first step of the assay run, the reporter double-strands and the miRNA targets were incubated with the surface-bound capture units to form a tripartite hybridization complex, which is stabilized by base-stacking effects at the three emerging nick-sites (Figure 2A).[23−26] After a short washing step performed to remove the excess of reporter units, ligase was added to covalently connect the two ends of the target miRNA to the adjacent reporter and capture strand (Figure 2B). As phosphorylated 5′-RNA-ends cannot be efficiently ligated to 3′-OH-DNA ends by T4 Ligases,[27] a chimeric 5′-thiol-modified oligonucleotide strand with a major DNA part and four RNA-bases at the 3′-end was used. After the ligation step, the reaction chamber was washed with a low salt buffer to remove all assay components that were not covalently bound to the immobilized capture strands (Figure 2C). To read out the esterase 2 reporter signal, p-APB substrate was added and the electrical current caused by pAP redox cycling at the interdigitated gold electrodes was measured. The whole process was fully automized with a time to results (TTR) of only 30 min.

Figure 3 shows an example of raw electrical current data collected during a measurement at the end of a 100 pM target

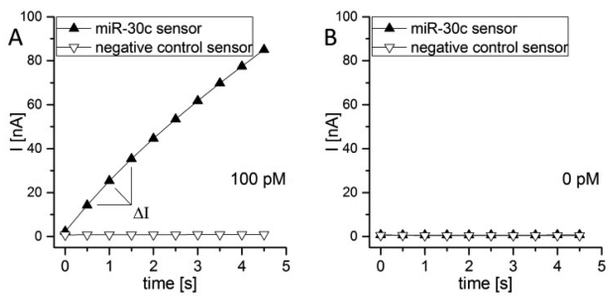


Figure 3. Electrochemical signal course. The digitally recorded raw data of the electrical currents (I) acquired from a miR-30c specific sensor position and a negative control sensor during assay runs with miR-30c target concentrations of 100 (A) and 0 pM (B) are plotted over measurement time. Reporter units for nine different miRNAs (cardiac panel) and at least one additional miRNA target (showing positive signals) were present in the sample solution of both assay runs. The sensor positions were locked at time point 0 s. Points 0.5, 1.0, and 1.5 were used to calculate the slope of the electrical current (ΔI) in downstream data analysis.

miRNA and a 0 pM control assay. The current shows an increase over time only when the sensor target was present in the hybridization sample, proving a successful practical execution of the described procedure. In downstream analysis the raw data of the experiments were processed by using three time points (0.5 to 1.5 s) to calculate the slope of the current for each sensor position with the least-squares linear fit technique.

Multiplex. To investigate the ability of the here proposed quantification method to measure several miRNAs in parallel in a multiplex setup, a panel of measurement components for eight miRNAs related to cardiovascular disease[3,28−33] (cardiac panel), a spike-in, and a negative control were designed, respectively (see Tables 1−4). Using this set of assay components, four members of the cardiac panel, miR-191, -15a, -145, and -636, were multiplex measured at a concentration of 100 pM each. In a second assay run, the same concentrations of the other four target nucleic acids of the panel, miR-181a, -425, -30c, and -362, were detected in the same manner. The spike-in control was used in both runs to secure comparability of the results. As the data confirms, the reported method provides a very good multiplexing capacity being insensitive to cross-talk effects (Figure 4).

Sensitivity. Calibration curves for all miRNA members of the cardiac panel were determined by repeating the assays described in Multiplex and applying several different concentration levels of the target miRNAs. A double logarithmic overlay of the acquired data points illustrates very good signal dynamics between 1 pM and 1 nM target concentration. This demonstrates the suitability of the presented method for miRNA quantification at low concentration levels (Figure 5A).

DOI: 10.1021/acs.analchem.5b02850
Anal. Chem. 2015, 87, 12104−12111

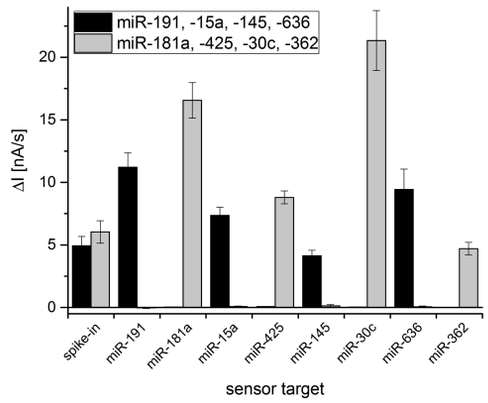


Figure 4. Multiplex measurements with an eight miRNA target panel. The results of two assay runs with reporter units for miR-191, miR-181a, miR-15a, miR-425, miR-145, miR-30c, miR-636, miR-362 (cardiac panel), and spike-in are shown. The sample solutions also contained 200 pM spike-in and 100 pM of each panel miRNA divided on the two assays. Medians and median absolute deviations (MAD) were plotted after correction by negative control.

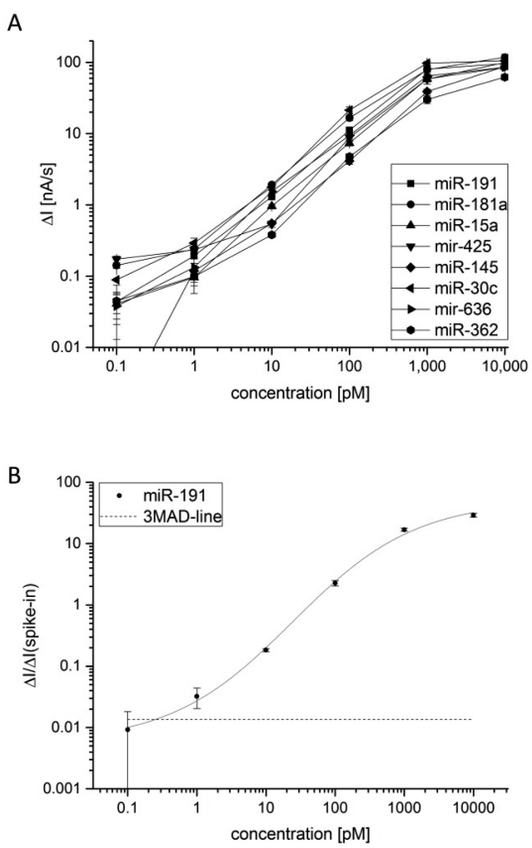


Figure 5. Standard curves and analytical sensitivity. (A) The multiplex assays from Figure 4 were repeated with different target concentrations ranging from 100 fM to 10 nM. Medians and MADs were plotted to obtain standard curves for all miRNAs of the cardiac panel. (B) Data of miR-191 were corrected by negative control and normalized to the spike-in signal. The 3MAD-line (median+3*MAD) was gained from a control experiment with only spike-in control as target nucleic acid.

Figure 5B illustrates a sigmoidal calibration curve obtained from the normalized measurement data for miR-191. The corresponding zero line representing the median plus three median absolute deviations (MADs) was gained from a zero concentration control assay. The shown data indicate a high analytical sensitivity considering the absence of any target or reporter amplification in the applied method.

Specificity. The biggest challenge for the specificity of nucleic acid detection assays is the differentiation of targets that differ only in a single or few nucleotides. When working with miRNAs the let-7 family is often used as a benchmark for specificity testing, as several of its members exhibit only one or two nucleotide differences. Four candidates of this family (let-7a, let-7b, let-7c, and let-7f) were selected to evaluate the specificity of the assay format (Figure 6A). The assay

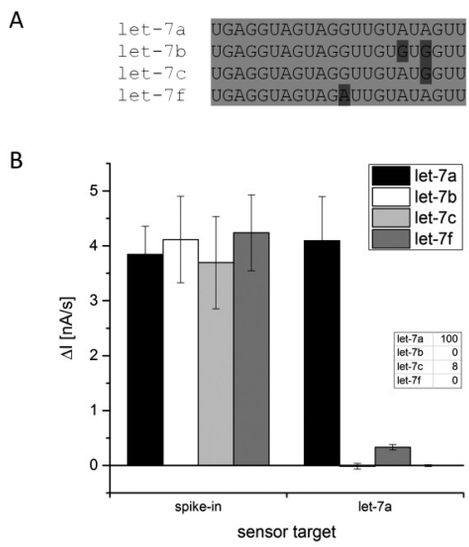


Figure 6. Discrimination of let-7 family members. (A) Let-7b, -7c, and -7f sequences differ from the assay target let-7a only by one or two mismatches (shaded dark gray). (B) 1 nM of each synthetic miRNA target was measured in separate assay runs. The sample solutions contained reporter units for let-7a and spike-in as well as 1 nM spike-in target in 2× SSC buffer with 0.05% Tween 20. A 0.5× SSC buffer was used as system fluid. Medians and MADs were plotted after correction by negative control. The table shows the relative signal for each miRNA after normalization to the spike-in control in percent.

components were designed for the quantification of let-7a. One nM of each target was measured separately in the presence of the spike-in control. The acquired data point out that let-7b with two mismatches and let-7f with a single mismatch near the central nick site of the tripartite hybridization construct can be distinguished from let-7a very well, leading to no false-positive signal at the let-7a specific sensor positions (Figure 6B). Let-7c shows a cross hybridization of 8% caused by an unfavorable position of the single mismatch. Presumably an optimized oligonucleotide design and fine-tuning of the hybridization conditions could improve this result if required. Overall the presented miRNA quantification method exhibits high specificity, which is comparable to the performance of commercially available miRNA detection assays (Affymetrix QuantiGene 2.0 miRNA Assay; Exiqon miRCURY LNA microRNA Array).

Endogenous miRNA. The reported miRNA detection method was used to quantify endogenous miRNAs of the cardiac panel in purified total RNA samples from donor blood.

DOI: 10.1021/acs.analchem.5b02850
Anal. Chem. 2015, 87, 12104−12111

Different amounts of total RNA were taken from a single sample and were multiplex measured utilizing the validated assay components for the eight miRNAs. For miR-191, miR-181a, miR-15a, miR-425, and miR-30c the detected signal corresponded with the concentration of total RNA in the hybridization solution (Figure 7A). Two miRNAs (miR-636, miR-362) did not show this correlation due to very low abundance in the measured sample. The miR-145 signal levels gained from the measurements of the two lower RNA amounts were inverted hinting to a miRNA concentration near the limit of the analytical sensitivity in the investigated range of sample material. The measured signal for the spike-in control was reciprocal to the applied relative sample volume suggesting a sensitivity of the synthetic miRNA to remaining impurities in the endogenous RNA fraction, whereas the detection system was not affected. This has to be taken into account when comparing the spike-in signals acquired from measurements of samples from different purification runs.

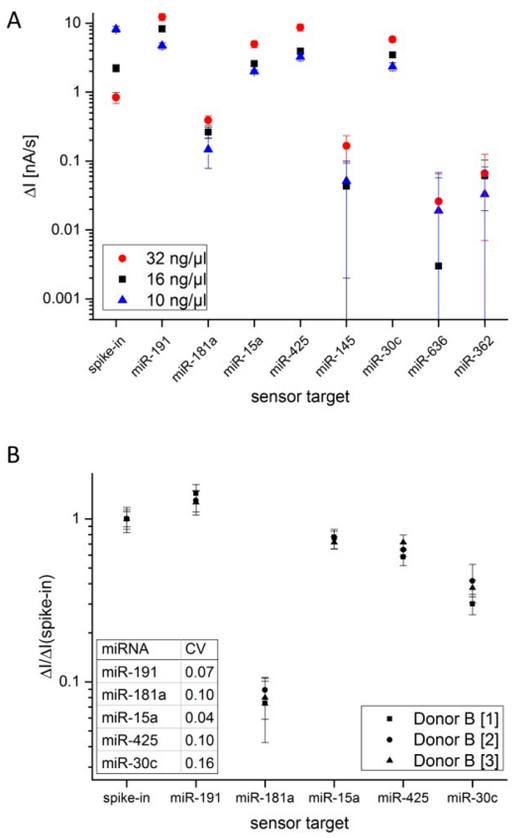


Figure 7. Quantification of endogenous miRNAs in total RNA samples from donor whole blood. (A) Different amounts of total RNA from a single purification sample were measured with the reported detection method applying the cardiac panel assay components and 200 pM spike-in. Medians and MADs were plotted after correction by negative control. (B) Three assay runs with 1 μg total RNA from a second sample and 200 pM spike-in target were performed using the cardiac assay panel to investigate reproducibility. Medians and MADs of the five higher abundant miRNAs were plotted after negative control correction and normalization to the spike-in control. The table shows calculated CV values for the diagramed miRNA candidates.

Figure 7B shows the normalized signals for miR-191, miR-181a, miR-15a, miR-425, and miR-30c determined from three equal assay runs. The same amount of material was taken from a single total RNA sample for all three experiments. The results reveal a good reproducibility (CV values 0.04−0.16) of the reported quantification method for the multiplex measurement of endogenous miRNAs.

Finally, we compared the measured quantities for the five higher abundant miRNAs of the cardiac panel with values obtained by qRT-PCR. For this purpose, a pool of extracted total RNA from several purifications was analyzed applying both quantification methods. The molar concentration of each target miRNA in the total RNA sample was calculated from the measurement data using corresponding standard curves. The resulting concentration values reveal a weak correlation between the two quantification methods. Details are provided in the respective scatter plot (Figure 8). This is, however, in-

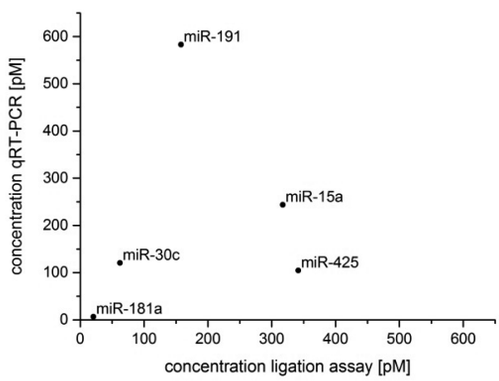


Figure 8. Comparison with qRT-PCR results. A pool of extracted total RNA was measured with the reported detection method as described in Figure 7 (2 μg total RNA per run) and with qRT-PCR. The concentrations of the five higher abundant miRNAs of the cardiac panel were calculated for two ligation assay runs and three qRT-PCR experiments, as described in Materials and Methods. The mean values were plotted in a scattergraph.

line with known results. Intraplatform comparability of different miRNA detection systems is known as consisting challenge.[34] Specifically, several studies have already demonstrated limited correlations between hybridization-based methods and qPCR when analyzing miRNA expression profiles.[35−37] Stated reasons include lack of standardized normalization, differences in miRNA processing, and difficulties with the distinction of precursors and mature miRNAs.[34] The efficiency of the detection of frequently occurring variants of miRNA targets, so-called isomiRs, has even been shown to vary considerably between different qPCR platforms.[38] Therefore, a case-related validation of the suitability of a detection system for the analysis of a specific miRNA expression pattern will be necessary prior to clinical application.

CONCLUSIONS

The miRNA detection method reported in this paper is sensitive, specific and very fast. The presented data show an analytical sensitivity below 1 pM target concentration, and though amplification-based methods like qRT-PCR might be more sensitive, they are additionally prone to errors caused by contaminations and amplification bias.[39] Furthermore, the specificity of this hybridization-based measurement technique was demonstrated by successful discrimination of down to single nucleotide mismatch candidates of the let-7 family.

DOI: 10.1021/acs.analchem.5b02850
Anal. Chem. 2015, 87, 12104−12111

An extraordinary characteristic of the reported assay is the excellent quantifiable range over 3 orders of magnitude with a TTR of only 30 min. Therefore, a high resolution for the quantification of target nucleic acids is ensured. The keys to this feature are the use of a huge excess of reporter compounds and its effective removal prior to the signal measurement through a harsh washing step. Thus, both fast formation of the hybridization construct and low background signal are combined. The covalent attachment of the label enzyme to the immobilized capture molecule via ligation and the utilization of a very stable label enzyme are crucial to maintain the positively labeled capture sites during the low salt washing conditions.

A combination of the reported approach with a CMOS array microchip for electrochemical redox cycling signal acquisition allowed for an 8-plex (plus controls) measurement of a predefined miRNA panel. As the microchip provides 128 sensor positions there is still room for extension of the number of simultaneously quantified miRNAs. The assay procedure itself is simple and thus was easily integrated into a portable reader device prototype offering a fluidic system and an electronic signal read-out (Siemens CT), whereby fully automated assay runs were enabled.

This portable quantification system can be a great tool to evaluate or utilize diagnostically relevant miRNA biomarker panels that are currently investigated for a multitude of different diseases by numerous research groups.

AUTHOR INFORMATION

Corresponding Author

*Phone: +49-9131-732871. Fax: +49-9131-732164. E-mail: walter.gumbrecht@siemens.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank the AG Sprinzl (University of Bayreuth) for supporting the preparation of Est2-ODN-conjugate. S.H. thanks his doctoral adviser Prof. Dandekar (University of Würzburg). This work was in part supported by funds from the EU-FP7-HEALTH-2012 Project BestAgeing (Grant No. 306031).

REFERENCES

(1) Bartel, D. P. *Cell* **2004**, *116*, 281−297.

(2) Lee, R. C.; Feinbaum, R. L.; Ambros, V. *Cell* **1993**, *75*, 843−854.

(3) Meder, B.; Keller, A.; Vogel, B.; Haas, J.; Sedaghat-Hamedani, F.; Kayvanpour, E.; Just, S.; Borries, A.; Rudloff, J.; Leidinger, P.; Meese, E.; Katus, H. A.; Rottbauer, W. *Basic Res. Cardiol.* **2011**, *106*, 13−23.

(4) Kondkar, A. A.; Abu-Amero, K. K. *BioMed Res. Int.* **2015**, *2015*, 821823.

(5) Keller, A.; Leidinger, P.; Bauer, A.; Elsharawy, A.; Haas, J.; Backes, C.; Wendschlag, A.; Giese, N.; Tjaden, C.; Ott, K.; Werner, J.; Hackert, T.; Ruprecht, K.; Huwer, H.; Huebers, J.; Jacobs, G.; Rosenstiel, P.; Dommisch, H.; Schaefer, A.; Muller-Quernheim, J.; Wullich, B.; Keck, B.; Graf, N.; Reichrath, J.; Vogel, B.; Nebel, A.; Jager, S. U.; Staehler, P.; Amarantos, I.; Boisguerin, V.; Staehler, C.; Beier, M.; Scheffler, M.; Buchler, M. W.; Wischhusen, J.; Haeusler, S. F.; Dietl, J.; Hofmann, S.; Lenhof, H. P.; Schreiber, S.; Katus, H. A.; Rottbauer, W.; Meder, B.; Hoheisel, J. D.; Franke, A.; Meese, E. *Nat. Methods* **2011**, *8*, 841−843.

(6) Bianchi, F.; Nicassio, F.; Marzi, M.; Belloni, E.; Dall'olio, V.; Bernard, L.; Pelosi, G.; Maisonneuve, P.; Veronesi, G.; Di Fiore, P. P. *EMBO Mol. Med.* **2011**, *3*, 495−503.

(7) Li, C.; Fang, Z.; Jiang, T.; Zhang, Q.; Liu, C.; Zhang, C.; Xiang, Y. *BMC Med. Genomics* **2013**, *6*, 16.

(8) Schultz, N. A.; Dehlendorff, C.; Jensen, B. V.; Bjerregaard, J. K.; Nielsen, K. R.; Bojesen, S. E.; Calatayud, D.; Nielsen, S. E.; Yilmaz, M.; Hollander, N. H.; Andersen, K. K.; Johansen, J. S. *JAMA* **2014**, *311*, 392−404.

(9) Vogel, B.; Keller, A.; Frese, K. S.; Leidinger, P.; Sedaghat-Hamedani, F.; Kayvanpour, E.; Kloos, W.; Backe, C.; Thanaraj, A.; Brefort, T.; Beier, M.; Hardt, S.; Meese, E.; Katus, H. A.; Meder, B. *Eur. Heart J.* **2013**, *34*, 2812−2822.

(10) de Planell-Saguer, M.; Rodicio, M. C. *Anal. Chim. Acta* **2011**, *699*, 134−152.

(11) Chin, C. D.; Linder, V.; Sia, S. K. *Lab Chip* **2012**, *12*, 2118−2134.

(12) Elsholz, B.; Nitsche, A.; Achenbach, J.; Ellerbrok, H.; Blohm, L.; Albers, J.; Pauli, G.; Hintsche, R.; Worl, R. *Biosens. Bioelectron.* **2009**, *24*, 1737−1743.

(13) Elsholz, B.; Worl, R.; Blohm, L.; Albers, J.; Feucht, H.; Grunwald, T.; Jurgen, B.; Schweder, T.; Hintsche, R. *Anal. Chem.* **2006**, *78*, 4794−4802.

(14) Nebling, E.; Grunwald, T.; Albers, J.; Schafer, P.; Hintsche, R. *Anal. Chem.* **2004**, *76*, 689−696.

(15) Niwa, O.; Xu, Y.; Halsall, H. B.; Heineman, W. R. *Anal. Chem.* **1993**, *65*, 1559−1563.

(16) Albers, J.; Grunwald, T.; Nebling, E.; Piechotta, G.; Hintsche, R. *Anal. Bioanal. Chem.* **2003**, *377*, 521−527.

(17) Gumbrecht, W.; Hintsche, R.; Mund, K.; Stanzel, M. Method for Preventing Chemical Crosstalk in Enzyme-linked Reactions, and Associated System. U.S. Patent 7,838,261 B2, Nov 23, 2010.

(18) Gumbrecht, W.; Paulicka P.; Stanzel, M. Apparatus and Method Comprising a Sensor Array and a Porous Plunger and Use Thereof. U.S. Patent 8,753,582 B2, June 17, 2014.

(19) Wang, Y. R.; Stanzel, M.; Gumbrecht, W.; Humenik, M.; Sprinzl, M. *Biosens. Bioelectron.* **2007**, *22*, 1798−1806.

(20) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.

(21) Keller, A.; Leidinger, P.; Borries, A.; Wendschlag, A.; Wucherpfennig, F.; Scheffler, M.; Huwer, H.; Lenhof, H. P.; Meese, E. *BMC Cancer* **2009**, *9*, 353.

(22) Hayashi, K.; Nakazawa, M.; Ishizaki, Y.; Hiraoka, N.; Obayashi, A. *Nucleic Acids Res.* **1986**, *14*, 7617−7631.

(23) Lane, M. J.; Paner, T.; Kashin, I.; Faldasz, B. D.; Li, B.; Gallo, F. J.; Benight, A. S. *Nucleic Acids Res.* **1997**, *25*, 611−617.

(24) Yakovchuk, P.; Protozanova, E.; Frank-Kamenetskii, M. D. *Nucleic Acids Res.* **2006**, *34*, 564−574.

(25) Yuan, B. F.; Zhuang, X. Y.; Hao, Y. H.; Tan, Z. *Chem. Commun.* **2008**, 6600−6602.

(26) Pöhlmann, C.; Sprinzl, M. *Anal. Chem.* **2010**, *82*, 4434−4440.

(27) Bullard, D. R.; Bowater, R. P. *Biochem. J.* **2006**, *398*, 135−144.

(28) Arora, P.; Wu, C.; Khan, A. M.; Bloch, D. B.; Davis-Dusenbery, B. N.; Ghorbani, A.; Spagnolli, E.; Martinez, A.; Ryan, A.; Tainsh, L. T.; Kim, S.; Rong, J.; Huan, T.; Freedman, J. E.; Levy, D.; Miller, K. K.; Hata, A.; Del Monte, F.; Vandenwijngaert, S.; Swinnen, M.; Janssens, S.; Holmes, T. M.; Buys, E. S.; Bloch, K. D.; Newton-Cheh, C.; Wang, T. J. *J. Clin. Invest.* **2013**, *123*, 3378−3382.

(29) Li, J.; Xu, J.; Cheng, Y.; Wang, F.; Song, Y.; Xiao, J. *J. Cell. Mol. Med.* **2013**, *17*, 1363−1370.

(30) Liu, X.; Wang, L.; Li, H.; Lu, X.; Hu, Y.; Yang, X.; Huang, C.; Gu, D. *Atherosclerosis* **2014**, *233*, 349−356.

(31) van Rooij, E.; Olson, E. N. *Nat. Rev. Drug Discovery* **2012**, *11*, 860−872.

(32) Vogel, B.; Keller, A.; Frese, K. S.; Kloos, W.; Kayvanpour, E.; Sedaghat-Hamedani, F.; Hassel, S.; Marquart, S.; Beier, M.; Giannitis, E.; Hardt, S.; Katus, H. A.; Meder, B. *Clin. Chem.* **2013**, *59*, 410−418.

(33) Zhou, L.; Zang, G.; Zhang, G.; Wang, H.; Zhang, X.; Johnston, N.; Min, W.; Luke, P.; Jevnikar, A.; Haig, A.; Zheng, X. *PLoS One* **2013**, *8*, e79805.

(34) Moldovan, L.; Batte, K. E.; Trgovcich, J.; Wisler, J.; Marsh, C. B.; Piper, M. *J. Cell. Mol. Med.* **2014**, *18*, 371−390.

DOI: 10.1021/acs.analchem.5b02850
Anal. Chem. 2015, 87, 12104−12111

(35) Chen, Y.; Gelfond, J. A.; McManus, L. M.; Shireman, P. K. *BMC Genomics* **2009**, *10*, 407.
(36) Camarillo, C.; Swerdel, M.; Hart, R. P. *Methods Mol. Biol.* **2011**, *698*, 419−429.
(37) Wang, B.; Howel, P.; Bruheim, S.; Ju, J.; Owen, L. B.; Fodstad, O.; Xi, Y. *PLoS One* **2011**, *6*, e17167.
(38) Lee, L. W.; Zhang, S.; Etheridge, A.; Ma, L.; Martin, D.; Galas, D.; Wang, K. *RNA* **2010**, *16*, 2170−2180.
(39) Bustin, S. A.; Nolan, T. *J. Biomol. Technol.* **2004**, *15*, 155−166.

DOI: 10.1021/acs.analchem.5b02850
Anal. Chem. 2015, 87, 12104−12111

Alzheimer's & Dementia 12 (2016) 565-576

Featured Article

Validating Alzheimer's disease micro RNAs using next-generation sequencing

Andreas Keller[a,*], Christina Backes[a], Jan Haas[b], Petra Leidinger[c], Walter Maetzler[d], Christian Deuschle[d], Daniela Berg[d], Christoph Ruschil[d], Valentina Galata[a], Klemens Ruprecht[e], Cord Stähler[f], Maximilian Würstle[f], Daniel Sickert[f], Manfred Gogol[g], Benjamin Meder[b], Eckart Meese[c]

[a]*Clinical Bioinformatics, Saarland University, Saarbrücken, Germany*
[b]*Internal Medicine, Heidelberg University, Heidelberg, Germany*
[c]*Department for Human Genetics, Saarland University Hospital, Homburg, Germany*
[d]*Department of Neurodegeneration and Hertie-Institute of Clinical Brain Research of the Eberhard-Karls-University, German Center for Neurodegenerative Diseases, Tübingen, Germany*
[e]*Department of Neurology, Charité - Universitätsmedizin Berlin, Berlin, Germany*
[f]*Siemens Healthcare, Erlangen, Germany*
[g]*Krankenhaus Lindenbrunn, Lindenbrunn, Germany*

Abstract

Introduction: Molecular biomarkers for Alzheimer's disease (AD) can support detection and improved care for patients, but novel candidates require verification. We previously reported a 12-micro RNA (miRNA) blood-based signature using next-generation sequencing (NGS) of 54 AD cases and 22 controls.

Methods: We performed validation of 49 AD cases and 55 controls using NGS and also included 20 mild cognitive impairment and 90 multiple sclerosis samples to identify nonspecific markers. Thus, 103 AD cases, 77 unaffected controls, and 110 diseased controls were sequenced. Although the initial cohort came predominantly from the United States, the validation samples were collected in Germany.

Results: Five hundred eighty miRNAs were detected in the blood. In the initial cohort, we observed 203, in the validation cohort, 146 dysregulated miRNAs at a significance level of 0.05. With 68 miRNAs, the overlap was significant ($P = .0003$). Likewise, the area under the receiver operator characteristic curve values of the miRNAs correlated (correlation of 0.93; 95% confidence interval 0.89–0.96; $P < 10^{-16}$).

Discussion: MiRNAs have the potential to support AD diagnosis and patient care.

Keywords: Alzheimer's disease; miRNA; Biomarker; Validation

1. Introduction

Alzheimer's disease (AD) care represents one of the grand challenges in health care systems worldwide. It is the most common form of dementia affecting already in 2009 more than 27 million individuals. Given demographic changes, it is expected that by 2050, the worldwide number of AD patients continuously will rise to 86 million [1]. The identification of peripheral biomarkers for an early, at best presymptomatic, detection of AD has the potential to improve AD patient care. Currently, β-amyloid (Aβ) and tau protein levels in the cerebrospinal fluid (CSF) are applied to distinguish between patients with AD and elderly individuals without AD [2]. In addition to these tests and imaging-based approaches that are applied in clinical routine

A part of the study has been funded by Siemens Healthcare. Siemens had no influence in study design, study set up, cohort selection, or data analysis.

*Corresponding author. Tel.: +49-681-302-68611; Fax: +49-681-302-68610.

E-mail address: andreas.keller@ccb.uni-saarland.de

http://dx.doi.org/10.1016/j.jalz.2015.12.012

(e.g., positron emission tomography, structural magnetic resonance imaging [MRI] or functional-connectivity MRI), many molecular biomarker panels have been proposed for improved diagnosis. An overview of respective novel early test candidates is provided here [3]. Such tests are frequently not validated or just to a limited amount. Among the most promising candidates are multiplexed protein panels, as described by Doecke et al. [4], or lipidomic panels as described by Mapstone et al. [5]. Another class of markers are small non-coding micro RNAs (miRNAs). These have been described as circulating markers in many human pathologies [6]. Like other biomarker panels, blood-borne miRNAs were usually validated just to a restricted degree. Few studies reveal the full potential of respective test on large cohorts. Among the most promising studies are validated biomarker signatures in pancreatic cancer, as recently described by Schultz et al. [7].

For AD, over a dozen studies in blood cells, plasma, and serum have been carried out. The heterogeneity in study setup, underlying technology, number of miRNAs profiled, cohort sizes, and biostatistics impedes a comparison or meta-analysis of the studies. Among the studies, we presented a case-control study on a US cohort of AD patients and controls that indicated a certain potential of miRNAs as AD markers [8]. After an initial screening using next-generation sequencing (NGS) of 54 AD cases and 22 unaffected controls, we performed technical and biological validation of 12 markers, including 10 known miRNAs and two novel miRNA candidates, using real-time quantitative reverse transcription PCR (RT-qPCR). Toward a clinical application, we recently established a novel assay that allows for quantifying respective miRNA on immunoassay analyzers that are used for routine diagnosis in central laboratories worldwide [9]. This assay allows for quantifying miRNAs with performance metrics comparable with standard enzyme-linked immunosorbent assay tests.

Whether the initially proposed signature measured predominantly from US samples can be replicated in an independent cohort remained, however, unclear. To facilitate clinical application, respective independent validation is, however, urgently required. Thereby, it is essential to use the same technologies (miRNA extraction, miRNA profiling, and biostatistics) to prevent falsified results introduced by bias. We, thus, set out to understand whether the miRNAs that have been discovered in the screening can be replicated in a German cohort by NGS. Although one alternative would have been to measure only the 12 miRNAs evaluated by RT-PCR in the initial study, we profiled the full portfolio of miRNAs by NGS again to understand how well the miRNAs overall can be replicated in a group of patients with different ethnical background.

Altogether, we screened 290 individuals by NGS, including the initial 54 AD cases and 22 unaffected controls that have been previously published, a replication cohort consisting of individuals collected in Germany of 49 AD cases and 55 controls. Beyond these samples, we also included 20 mild cognitive impairment (MCI) patients and 90 multiple sclerosis (MS) patients to understand whether the discovered miRNAs are specific for AD. Thereby, we generated almost 4 billion small RNA reads that were evaluated by computer-aided approaches.

2. Methods

2.1. Patients and miRNA profiling

We collected 2.5-mL blood from AD patients, controls, and MCI and MS patients in PAXgene Blood RNA tubes (PreAnalytiX) tubes. Patient characteristics (age, gender, age of onset, mini-mental state examination, Montreal cognitive assessment score, Aβ 42, tau and phospho-tau values, antidementive drugs, beta-blocker, and antihypertensive drugs) of the novel AD (n = 49) and unaffected control (n = 55) cohort are presented in Table 1.

The analytical procedure was performed as described previously [6,8]. In brief, from the tubes, total RNA was isolated using the PAXgene Blood miRNA Kit (Qiagen) following the manufacturer's instruction. For sequencing library preparation, 200 ng of total RNA was used (quantified by RNA 6000 Nano Chip using Bioanalyzer 2100 [Agilent]). Preparation was performed according to the protocol of the TruSeq Small RNA Sample Prep Kit (Illumina). Concentration of the ready prepped libraries was measured by using the Bioanalyzer (DNA 1000 Chip). Libraries were then clustered with a concentration of 9 pmol with six samples in one lane. Sequencing of 50 cycles was performed on a HiSeq 2000 instrument (Illumina) and demultiplexing of the raw sequencing data was done using CASAVA version 1.8.2.

Table 1
Information on newly measured AD samples and controls in the validation study

Variable	Alzheimer's disease	Controls	*P* value
Age	70.7 ± 8.2	67.3 ± 7.8	.034
Age of onset	68.7 ± 8.1	NA	.39
Gender (f/m)	22/27	29/26	.44
MMSE (0–30)	21.6 ± 3.8	29.5 ± 0.86	$<10^{-16}$
MoCA (0–30)	15.9 ± 4.7	28.8 ± 2.1	$<10^{-16}$
Abeta42 (pg/mL)	453 ± 209	NA	NA
Tau (pg/mL)	629 ± 334	NA	NA
p-tau (pg/mL)	83 ± 43	NA	NA
Antidementive drugs (yes/no)	14/34	NA	NA
Beta-blockers (yes/no)	7/31	NA	NA
Other antihypertensive drugs (yes/no)	17/21	NA	NA

Abbreviations: AD, Alzheimer's disease; NA, not applicable; MMSE, mini-mental state examination; MoCA, Montreal cognitive assessment; p-tau, phosphorylated tau.

NOTE. For age, age of onset, MMSE, and MoCA, two-tailed unpaired *t* tests were calculated. For gender, Fisher's exact test has been applied. With respect to the age of onset, the age distribution was compared with the age distribution of controls.

2.2. Statistical analysis

All 290 samples were processed by miRDeep2 as described previously [8,10] before downstream analysis in R (version 3.0.2) had been carried out. For all samples together, quantile normalization was performed and all miRNAs with <5 reads in less than five samples were excluded to minimize noise. This procedure resulted in a set of 580 miRNAs that were further investigated. Where applicable, *P* values were adjusted for multiple testing using Benjamini-Hochberg correction. For hypothesis testing, we calculated unpaired two-tailed *t* tests. Because not all miRNAs were normally distributed, we also calculated nonparametric Wilcoxon Mann-Whitney (WMW) tests (unpaired, two tailed). Beyond the hypothesis tests, the area under the receiver operator characteristic curves (AUC) was calculated for each miRNA. For correlating AUCs in both cohorts, AUCs were provided in an interval between 0 and 1. miRNAs with higher expression in AD have AUC <0.5 and miRNAs with higher expression in controls >0.5, miRNAs that are equally abundant have AUCs of around 0.5. To calculate confidence intervals (CIs) for the AUC, 1000 bootstrap samples have been performed using the pROC package. As further statistical approaches, we performed hierarchical clustering as implemented in the Heatplus R package (read counts were transformed to z-scores and complete linkage clustering relying on the Euclidian distance was done). We also carried out principal component analysis (PCA) as implemented in the prcomp R package and showed the first versus second principal component as scatter plot. Finally, analysis of variance (ANOVA) has been applied to the three groups: AD, unaffected controls, and diseased controls (MCI/MS).

To combine the predictive power of multiple miRNAs, machine learning has been performed similar to the approach described previously for lung cancer [11]. In detail, support vector machines using a radial basis function as kernel were trained and evaluated using fivefold cross validation on the complete data set. To account for variations between different cross-validation runs, the procedure has been repeated with 20 random partitions in test and training data. To select most informative miRNAs with respect to AD, a stepwise forward feature selection based on the *P* values has been carried out. Here, in each iteration, the k features (k was varied between two and 500 features) with lowest *P* values in the training part of the cross validation were selected and subsequently evaluated on the test sample part. To check for potential over training, 20 repetitions of permutation tests have been performed. Here, the complete subset selection step as well as the classification was carried out with randomly permuted class labels.

2.3. MiRNA enrichment and targetome analysis

We applied the miEAA tool (http://www.ccb.uni-saarland.de/mieaa_tool), which builds up on GeneTrail [12], which is tailored for gene set enrichment analysis, to find categories that are enriched with the 68 miRNAs significant in both studies and compared them to the background of all 580 miRNAs that were expressed in this study. All results with adjusted *P* values <.05 in an overrepresentation analysis after adjustment for multiple testing were considered significant.

To investigate putative downstream effects, we focused only on validated targets that have been extracted from the most recent build of the miRTarBase database (release 6, September 2015) [13]. We excluded the targets with weak interactions and include only those with functional interactions from that database leaving us with 6862 pairs of miRNAs targeting genes. Of these, 1638 have been duplicated entries, which were also removed, leaving us with 5224 pairs of miRNAs and validated targets. For the 68 miRNAs that overlapped in both studies, we built the full target network and also considered hubs, i.e., genes that are targeted by at least five miRNAs. Because these results may be biased toward more frequently analyzed miRNAs or genes, we also carried out random permutation tests. From all 580 miRNAs that were expressed but not among the 68 miRNAs overlapping in both studies (512 miRNAs) as background distribution, we randomly picked 68 miRNAs and performed the same analysis as mentioned previously. Specifically, we counted how many miRNAs target the hubs that are discovered for the original data. This random procedure has been repeated 10,000 times.

3. Results

For each of the 290 individuals (54 AD cases and 22 unaffected controls that have been previously published, novel 49 AD cases and 55 controls, 20 MCI and 90 MS patients), about 14 million reads were generated, totaling 3.85 billion NGS reads that have been statistically evaluated. The main goal of the present study is to compare the results on the previously published screening cohort (54 AD patients and 22 unaffected controls with similar age/gender distribution) and the newly measured German validation cohort (49 AD patients and 55 unaffected controls with similar age/gender distribution).

Beyond the validation of the initial results comparing AD to unaffected controls, we asked whether the signatures found by NGS are specific for AD. We, thus, sequenced 90 MS and 20 MCI samples that were used as non-AD controls. After excluding miRNAs that are expressed close to the background and contribute substantially to the noise in the signatures, 580 markers remained in our final data set (Supplementary Table 1).

In the following, we first compare the overall signatures in the screening and replication cohort and then focus specifically on the initially published signature. Second, we compare the miRNA abundance to clinical information such as therapy. Third, we derive in silico downstream information on the targets and target networks of the reported

Alzheimer miRNAs, and fourth, we compare the Alzheimer patients to patients with other diseases (MS and MCI).

3.1. Comparing AD samples to controls

First, we compared the dysregulation of all miRNAs between AD patients and unaffected controls in the screening and replication, not including the MCI and MS patients. Because for all miRNAs the abundances were not normally distributed, we performed WMW tests for calculating significance values in addition to *t* tests (the *t* test *P* values are provided in Supplementary Table 1). In the US cohort, we observed 203 dysregulated miRNAs at a significance level of 0.05 before adjustment and 127 dysregulated miRNAs after adjustment for multiple testing using WMW tests. In the validation cohort, we observed lower effect sizes and generally higher *P* values. Here, 146 miRNAs were dysregulated at a significance level of 0.05 before adjustment, 49 remaining after adjustment for multiple testing. In both cohorts, we found slightly more miRNAs with lower expression in AD patients. Of the 203 and 146 miRNAs, 68 overlapped. Given the total number of 580 expressed miRNAs, 203 miRNAs in the screening, and 146 in the validation cohort and an overlap of 68 miRNAs, we asked whether this overlap is statistically significant. Using the hypergeometric distribution, we calculated a statistically significant overlap ($P = .0003$). Details on the significance values (raw and adjusted *t* test and WMW test, *P* values, and AUC values) are provided in Supplementary Table 1. To provide further evidence for the high degree of concordance, we correlated the AUC values of the 68 miRNAs in the screening and validation cohort (Fig. 1). The correlation was as high as 0.93 (95% CI, 0.89–0.96) with a significance value $<10^{-16}$. Importantly, all 68 miRNAs match in the direction of regulation in the screening and replication cohort. As a graphical representation, we illustrate the expression of the 68 miRNAs as heat map after hierarchical clustering in Fig. 2. This heat map, which is based on z-scores of miRNAs in the screening and validation cohort, highlights a cluster with most controls on the right hand side, most AD patients in the middle, and a cluster containing AD and controls on the left hand side.

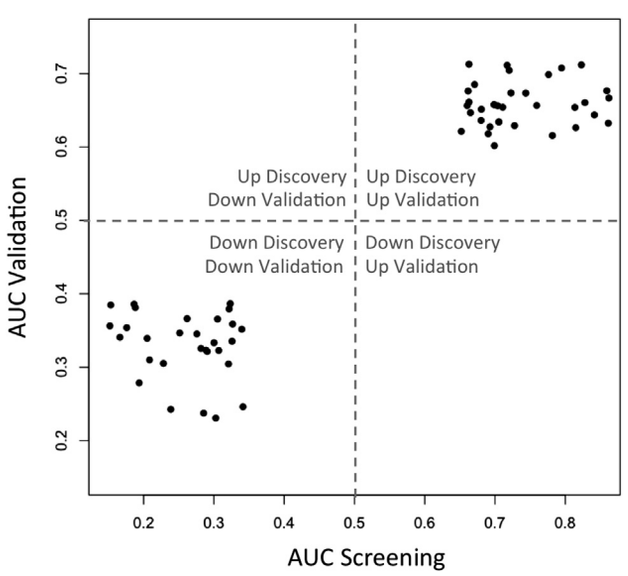


Fig. 1. AUC values for the comparison AD versus matched controls in the screening cohort (x-axis) and the validation cohort (y-axis). Abbreviations: AUC, area under the receiver operator characteristic curves; AD, Alzheimer's disease.

Initially, we published a 12-miRNA marker signature, containing 10 miRNAs known from miRBase and two novel miRNAs discovered in our screening cohort. In the replication, we focused only on known miRNAs as annotated in the reference database because novel miRNAs predicted by NGS may represent artifacts. Fig. 3 details all markers that have been dysregulated in the replication in the same direction as initially observed. However, not all miRNAs' *P* values (two-tailed WMW test adjusted for multiple testing) were below the alpha level of .05 in the replication. Especially miR-5010-3p and miR-26b-5p with significance values of 0.16 and 0.82 were not significantly dysregulated. Nonetheless, the correlation of AUC values of the screening and validation cohort was similar to the 68 markers overlapping in both studies (0.92), indicating that already the initial signature has been reasonably selected using only one cohort.

The marker with the lowest *P* value in the discovery and validation study combined was miR-151a-3p (adjusted *P* value of 10^{-7}) with an AUC of 0.74. On average, we measured 3758 reads in AD samples versus 2158 reads in control samples. Overall, largest AUCs were reached for hsa-miR-17-3p (AUC 0.77, adjusted *P* value of 10^{-5}). For miRNA 17-3p and 151a-3p, the receiver operator characteristic (ROC) curve is exemplarily shown in Figs. 4A and B, respectively. The blue shaded areas in the ROC curves correspond to the 95% CI that have been calculated by 1000 bootstrap samples. Altogether, the combined analysis of both cohorts yielded 192 significant miRNAs (two-tailed WMW test) before adjustment for multiple testing of which 127 remained significant after adjustment (details in Supplementary Table 1).

Although already single markers have a remarkable diagnostic potential, we performed a classification using Support Vector Machines (SVMs). The procedure has been carried out with a filter-based subset selection on the complete data set using 20 repetitions of fivefold cross validation (details are provided in the Methods section). In combining the predictive power of miRNAs using SVMs on 200 markers, the AUC increased significantly (z-score base *P* value of $<.05$) from 0.77 for best single marker to 0.84 on average for the 200 marker signatures. A representative example from the repeated cross-validation runs is presented in Fig. 4C. With the AUC, also the accuracy of the classification improved. For the best single marker accuracy, specificity and sensitivity were 73.3%, 75.3%, and 71.8%, respectively. By using signatures, the accuracy increases to 78.2%. Specificity and sensitivity were 68.9% and 87.6%. As the ROC curve in Fig. 4C demonstrates, specificity and sensitivity can, however, be well traded off against each other.

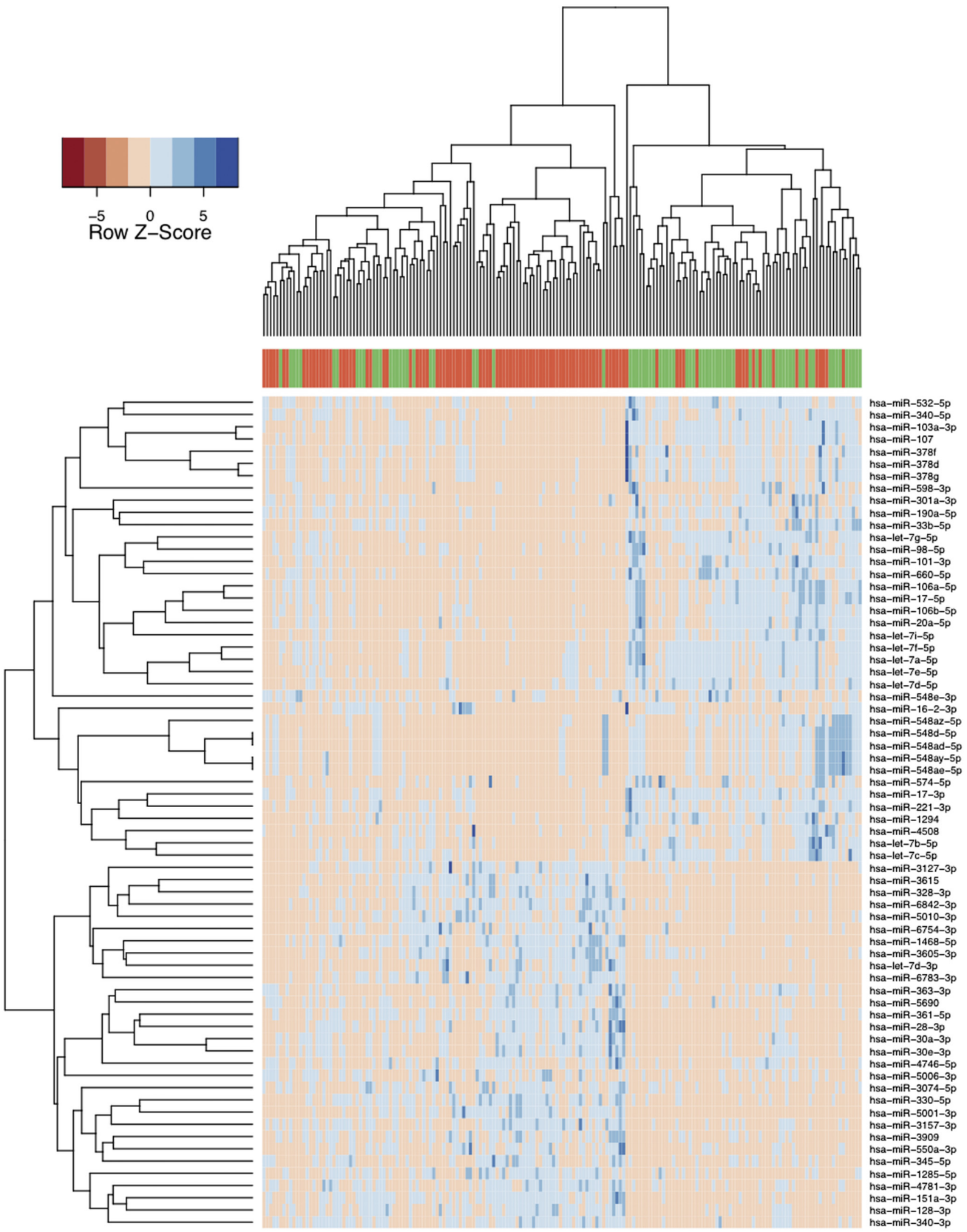


Fig. 2. Heat map after hierarchical clustering of the 68 miRNAs overlapping in both studies. Green individuals are controls, and red individuals AD cases. The color code for the cases and controls are projected between the dendrogram and the heat map. This figure contains all AD samples and all controls from the screening and validation cohort. Abbreviations: miRNA, micro RNA; AD, Alzheimer's disease.

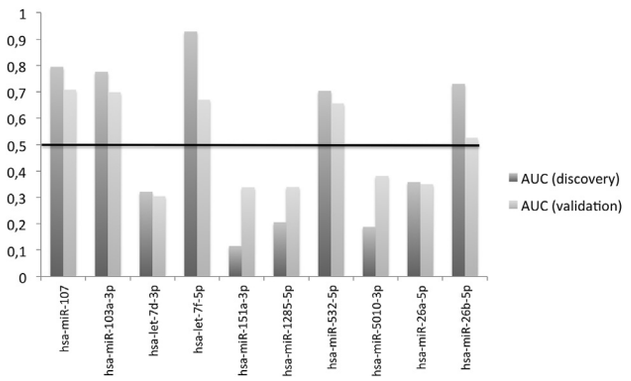


Fig. 3. AUC values in the screening and validation cohort for the 10 miRBase miRNAs from the initial signature. The dysregulation was concordantly observed for all miRNAs (upregulation in controls are above AUC values of 0.5—represented by the horizontal black line—and downregulation vice versa). The effects were, however, frequently lower in the replication cohort. Abbreviations: miRNA, micro RNA; AUC, area under the receiver operator characteristic curves.

3.2. Correlation of miRNome to therapy and other available clinical information

Because we had access to CSF markers of neurodegeneration, we correlated the available clinical and neurochemical information with the miRNA repertoires. Furthermore, the patients included in the study got different therapies such that we were also able to estimate variations in miRNA abundance correlated to the therapy.

First, we investigated a potential influence of antidementive and antihypertensive drugs. We did not observe any significant miRNA differences between AD patients with and without such treatment after adjustment for multiple testing. These data suggest that the influence of typical drug therapy on the miRNA pattern in the blood of the AD samples is negligible.

Disease duration may also influence miRNA pattern. In the present study, we included patients' blood samples close to the time of initial diagnosis but also samples of patients at more advanced stages. The mean lag time between diagnosis and blood collection was 2 years. We, therefore, compared samples of AD patients with disease duration <2 years to samples of patients with longer disease durations. In this comparison, we again calculated nonsignificant P values; none of the miRNAs remained significant after adjustment for multiple testing.

In a third comparison, we correlated values of cerebrospinal biomarkers including Abeta42, tau, and phosphorylated tau to all miRNAs separately. As for the drug analysis, we also did not observe any significant miRNA after adjustment for multiple testing.

3.3. MiRNA categories and the AD miRNAs' targetome

To understand common grounds of the respective miRNAs, we applied miEAA using the standard parameters. Specifically, we searched for categories that contain more of the 68 miRNAs overlapping in both cohorts as compared with the background of 580 miRNAs. With the lowest P value, we of course found our initial Alzheimer disease miRNA set. With respect to the organs category from miRWalk [14], we, e.g., observed overrepresentation of blood platelets and erythrocytes but also neurons. We also found a negative correlation of AD miRNAs with increasing age in individuals without known disease affection, meaning that the AD miRNAs per se were less expressed in controls older than 100 years [15]. All enriched categories at a significance level of 0.05 are summarized in Supplementary Table 2 along with the miRNAs contained in the respective categories.

We also investigated putative downstream effects and analyzed the targetome as described in the Materials and Methods section. Focusing on validated targets of the 68 miRNAs, we discovered a total of 563 interactions. The resulting network contains 33 miRNAs and 349 target genes. Of the 349 targeted genes, 14 are validated targets of at least five of the 33 miRNAs overlapping in both studies: VEGFA, DICER1, AGO1, PTEN, CDKN1A, APP, RB1, CCND1, CCND2, WEE1, IL13, HMGA2, TNFRSF10B, and MYC. The resulting subnetwork containing the respective hubs is presented in Fig. 5.

Because these analyses may be biased toward more frequently analyzed genes or miRNAs, we performed 10,000 permutation tests. For randomly selected 68 miRNAs from the background distribution, the same analyses as for the original 68 miRNAs were done. As compared to the 563 interactions in the original data, we observed an average of 390 miRNA-target interactions in the permutation tests, targeting on average 330 genes. Both numbers were lower compared with the original data; however, still 7.7% (overall number of interactions) and, respectively, 39.5% (overall number of genes) permutation runs exceeded the original results. Considering on the number of hubs, e.g., genes that are targeted by at least five miRNAs, we found an average of 1.3 genes across the 10,000 permutation test runs. In none of these runs, 14 genes were found to be targeted by at least five miRNAs as for the original data, indeed the maximal value was eight genes. Especially for the genes AGO1, APP, and IL13, not a single miRNA targeting these genes in the background distribution of 10,000 runs was observed. Calculating the P value for each gene as fraction of permutation tests with at least the same number of miRNAs targeting the respective gene (P value for those genes without any hit were set to 1/10,000), all 14 genes remained significant at an alpha level of 0.05 after adjustment for multiple testing.

3.4. Differentiation of AD from MCI and MS

In the previous section, we have described a successful validation of miRNAs that distinguish between AD samples and controls without known affection and similar age/gender distribution. We also observed potential relevance of the

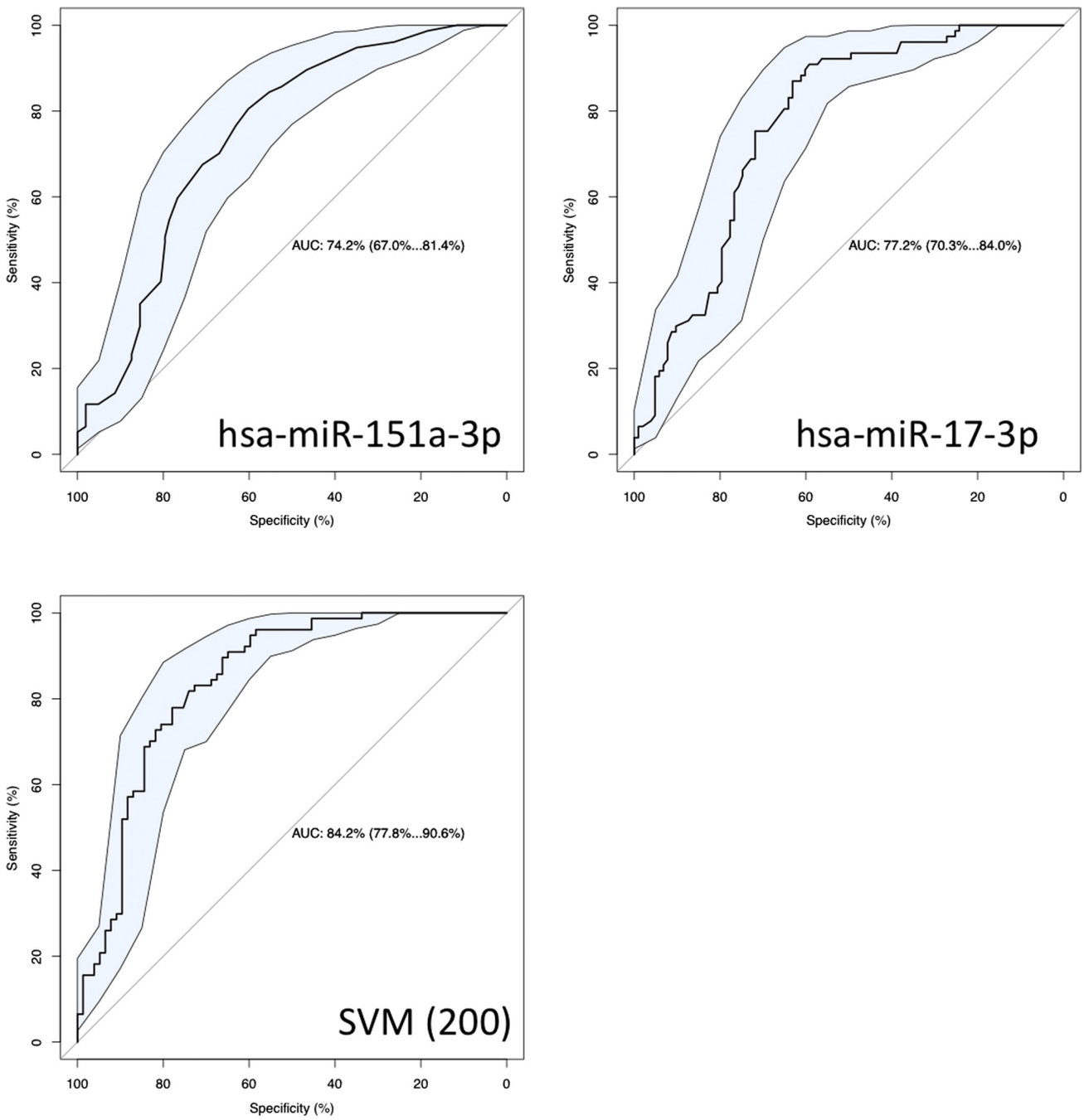


Fig. 4. ROC curves of the two miRNAs with lowest *P* values miR-151a-3p and miR-17-3p, and the 200-marker AD signature. The blue shaded areas represent the 95% confidence intervals computed by 1000 bootstrap samplings. Abbreviations: ROC, receiver operator characteristic; miRNA, micro RNA; AD, Alzheimer's disease.

miRNAs to changes in the metabolism of patients. The specificity of respective changes for AD remained, however, unanswered. We, thus, asked whether similar differences could be likewise detected in other diseases. In further computations, we first differentiated between AD and MCI patients. Again, we observed a substantial upregulation of miRNAs in AD patients. The lowest *P* value was discovered for miR-30c-5p. Here, 5836 reads mapped on average to AD samples, whereas 2158 mapped to MCI samples. Correspondingly, the adjusted *P* value was 4×10^{-13} and the AUC was 0.9. In sum, we found 148 significant miRNAs after adjustment for multiple testing remaining below the alpha level of 0.05. Of these, 119 were upregulated in AD and 29 were downregulated in AD samples as compared to MCI. The classification results for MCI versus AD again exceeded the single marker performance; however, from the limited

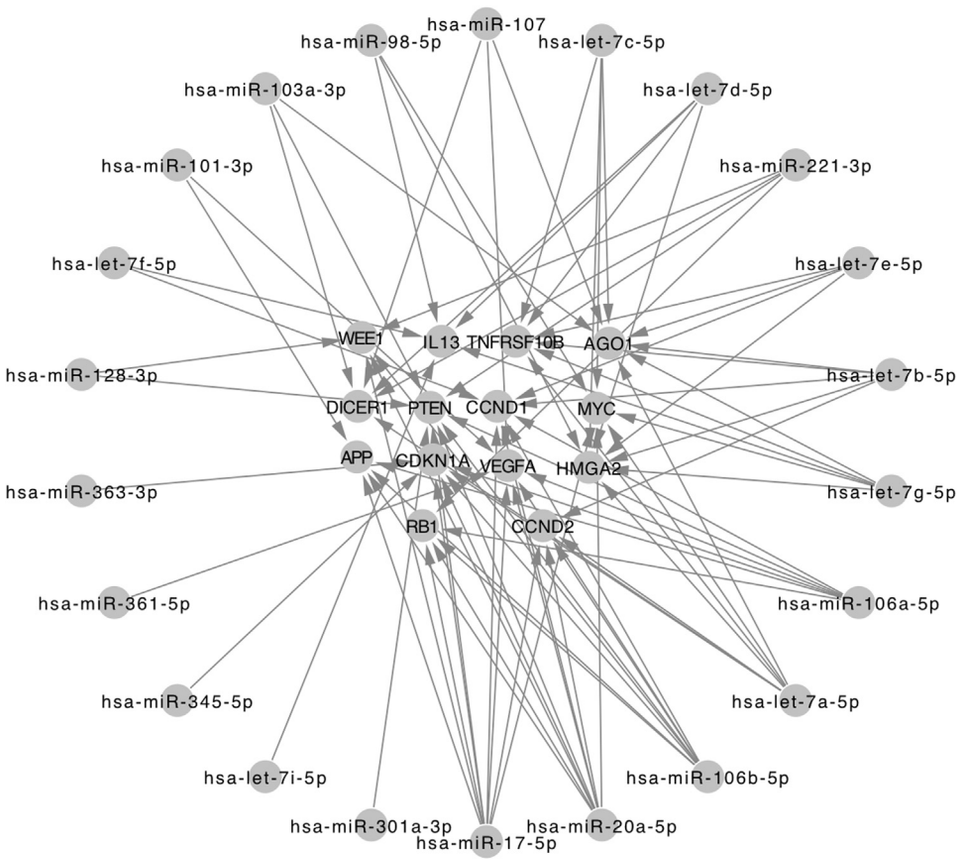


Fig. 5. Core regulatory network. The middle of the network contains the 14 genes targeted by at least five miRNAs, which are ordered as circle around the target genes. Abbreviation: miRNA, micro RNA.

MCI cohort, stable signatures can be derived just to a limited extent.

Besides MCI, we also compared the AD profiles to neuroinflammatory disorders. For MS (clinically isolated syndrome [CIS] as well as relapsing-remitting multiple sclerosis [RRMS]), we achieved the highest performance. Here, the AUC derived from the SVM model was 0.983 (95% CI, 0.969–0.997). Comparing AD to both MS subtypes CIS and RRMS, we did not observe significant differences in classification performance.

To provide further evidence that the results of AD, unaffected controls, and diseased control (MCI and MS) are different, we performed PCA and plotted the first versus second component as scatter plot (Fig. 6). Although the three cohorts show an overlap, the tendency of different patterns can be well observed, the unaffected controls are predominantly in the upper left part, the AD samples at the bottom, and the MS samples in the upper right part. Another advantage of our study is that for specific miRNAs, the patterns in those three cohorts can be directly compared to each other. Exemplarily, the two miRNAs differentiating between AD and controls presented in Fig. 4 are shown as box plots in Figs. 7A and B. For miR-151a-3p (adjusted ANOVA P value of 6×10^{-12}) and miR-17-3p (adjusted ANOVA P value of 3×10^{-11}), the differences between AD and unaffected controls can be observed. At the same time, diseased controls show a similar pattern as the unaffected controls, indicating that these miRNAs are specific for AD. On the other hand, miR-363-3p (adjusted ANOVA P value of 10^{-6}) presented in Fig. 7C is not only dysregulated in AD versus controls but also in MCI and MS against controls and, thus, not specific for AD. In sum, the results demonstrate that AD patients can be well separated from matched controls with similar age and gender distribution. Likewise, MCI patients show characteristic profiles that deviate from AD patients' profiles. Other neurologic disorders such as CIS and RRMS reveal even larger differences from AD and control profiles. Although a difference in the age of MS patients to the AD patients may contribute to the substantial differences in miRNA abundance, our results suggest that our signatures are rather specific for AD.

4. Discussion

To provide evidence that miRNAs measured from body fluids are reasonable disease markers, additional validation

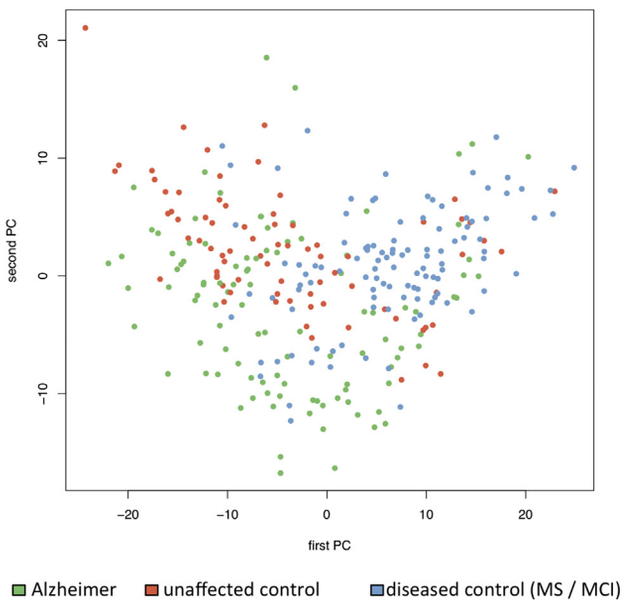


Fig. 6. PCA of miRNA expression in all 290 samples. The first and second principal component (PC) for all samples is visualized as scatter plot. Green dots are AD patients, orange dots are controls, and blue dots are MS/MCI patients. The profiles overall show an overlap but a clear tendency of samples of the different cohorts clustering together can be observed, although these two principal components contain only 25% of the overall data variance. Abbreviations: PCA, principal component analysis; miRNA, micro RNA; AD, Alzheimer's disease; MS, multiple sclerosis; MCI, mild cognitive impairment.

in independent cohorts has to be carried out. In the present study, we compared results of a miRNA marker discovery study on AD that has been performed on a US cohort with a German validation cohort. Between both studies, we observed a good concordance, 68 markers were significant in both studies.

Of the original marker signature, we focused on the 10 miRNAs from the reference database miRBase [16], leaving out the two novel candidates that deserve further investigation toward the question whether the molecules represent actual miRNAs or are artifacts from the NGS procedure.

Of the miRNAs, some were dysregulated significantly in both studies although the baseline level of miRNA between the US and German patients varied. One example is an miRNA from our original study: miR-1285-5p has average normalized read count of 8.9 and 3.6 in AD and controls samples in the United States cohort. In the German cohort, normalized read counts were 21.6 and 15.6 in AD and controls. This miRNA was downregulated in AD samples from United States and Germany; however, the absolute levels of that miRNA were higher in samples from Germany. Because these variations may reflect actual changes in miRNA levels but likewise sample handling may affect levels, especially for lower abundant miRNAs, technologies with improved quantification such as RT-PCR or immunoassay technology are likely more suitable for routine application. Likewise, different threshold values in miRNA abundance of individuals from different ethnics groups could be reasonable. With respect to our recently published immunoassay, we observed that around 25 of the 68 miRNAs are expressed in a sufficient manner to be above the detection limit of the immunoassay, whereas the remaining 43 would be too close to the detection limit of this amplification-free quantification approach. This together with the required degree of multiplexing makes RT-PCR a more reasonable platform for measuring the AD miRNAs in clinics as compared with our immunoassay.

Using machine learning techniques, we were able to distinguish well between AD patients and controls. Because of the previously described bias, we performed the whole classification procedure as cross-validation on the complete data set.

An enrichment analysis highlighted target genes that are controlled by the dysregulated miRNAs. Our analysis highlighted 14 genes that are targeted by at least five of the 68 miRNAs dysregulated in both cohorts: VEGFA, DICER1, AGO1, PTEN, CDKN1A, APP, RB1, CCND1, CCND2, WEE1, IL13, HMGA2, TNFRSF10B, and MYC. Many of those are key players for AD such as Aβ A4, or at least described in the context of AD. Vascular endothelial growth factor is known to be expressed in the brain of AD patients, e.g., in frontal and parahippocampal cortex [17]. Thomas et al. also report an increase of VEGF with disease severity. Recently, it has been reported that exogenous Aβs stimulate normal adult human astrocytes to produce and secrete even VEGF-A through calcium-sensing receptor–mediated mechanism [18]. Beyond the expression in the brain, low serum levels of VEGF are described to be associated with AD [19].

In addition, the tumor-suppressor PTEN has been reported to accumulate in Alzheimer neurofibrillary tangles [20]. Specifically, PTEN, alters tau phosphorylation [21,22].

miR-26b, which has been already included in our previously published signature on downstream targets, has been investigated. The known signaling cascades involve upregulation of Rb1/E2F leading to substantial downstream effects [23]. This miRNA was downregulated in blood of AD patients in the screening and validation cohort, the degree of downregulation was, however, marginal in the replication (Fig. 3). Interestingly, this miRNA is described to be upregulated in the brain of AD patients, showing the opposite behavior than blood-borne patterns. Inverse regulation of tissue and blood profiles has already been observed, e.g., in the case of cancer miRNAs [24]. A comprehensive PubMed analysis indicated several hundred hits for nine of the 14 genes related to AD. A less obvious example was Wee1, which is active in neurons of normal brain and is less active in AD patients. It is postulated that it promotes activation of Cdc2/cyclin B1 and, thus, represents a mitotic regulator, contributing to neurodegenerative processes [25].

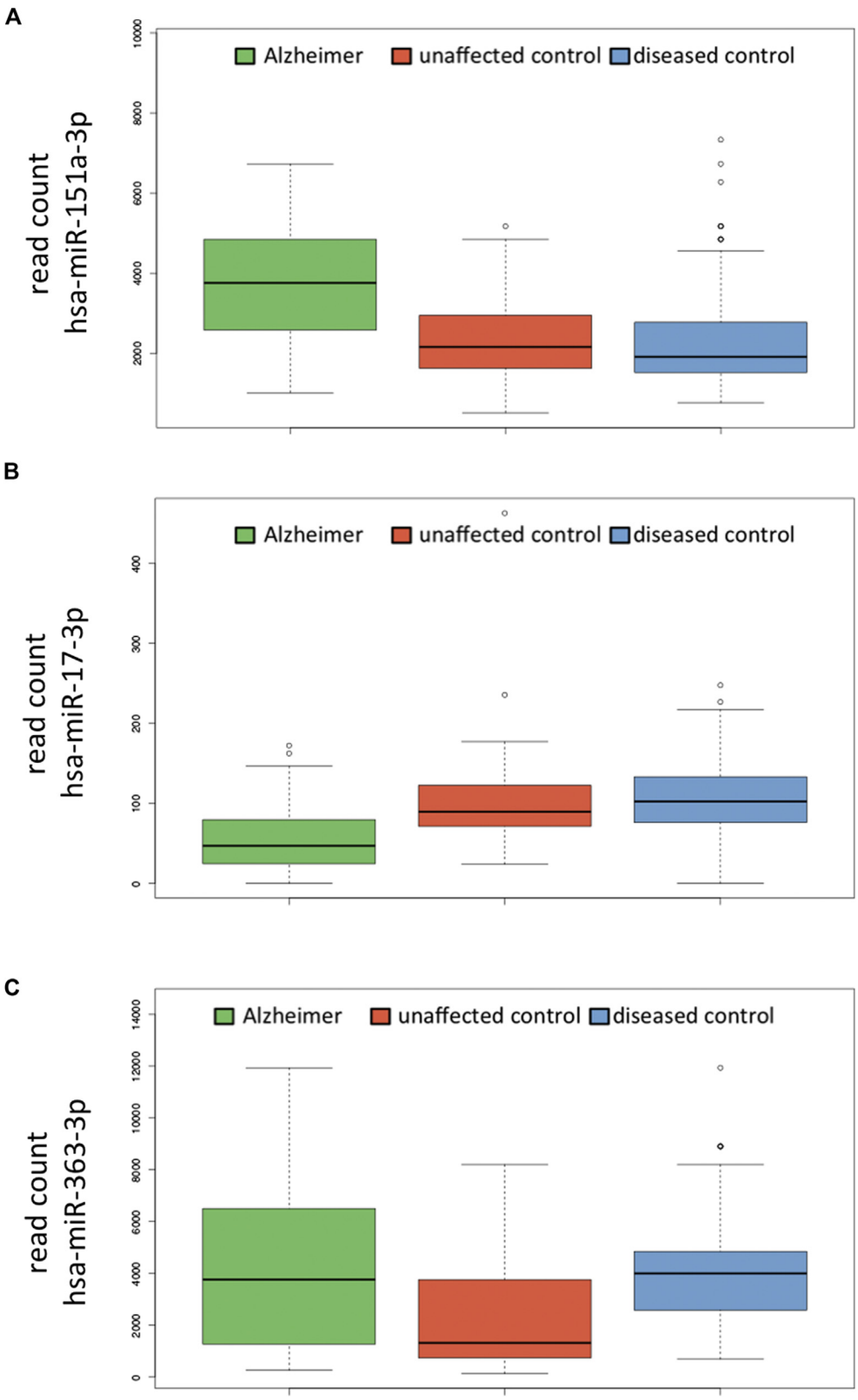


Fig. 7. Box plots for three miRNAs and the three cohorts AD (green), unaffected controls (blue), and MCI/MS (blue). The y-axis denotes the normalized NGS read count for the miRNAs in the three cohorts. Although the first two miRNAs in panels (A) and (B) are dysregulated in AD and match in MS and controls, the third example in panel (C) is upregulated in AD and MS samples. Abbreviations: miRNA, micro RNA; AD, Alzheimer's disease; MS, multiple sclerosis; MCI, mild cognitive impairment; NGS, next-generation sequencing.

The core of our study was, however, to test whether the initial miRNA signature from the screening cohort can be replicated and beyond this, to compare the signature to other diseases. To address the question whether the validated signature is specific for AD patients, we compared the profiles also to MCI and MS patients. In both comparisons, we observed significant miRNAs that let us distinguish between AD and the other two diseases. Although the proximity of AD patients to MCI patients was closer as compared with unaffected controls, we found larger differences from neuroinflammatory disorders. The different profiles between MCI and AD patients let us ask on significant alterations in miRNA abundance depending on the disease duration. In correlating the miRNA level to the disease duration of AD patients, we did, however, not observe a significant influence. There are three reasons, the observed time period may be too short, the observed cohort size is too small to discover small changes in the abundance of single miRNAs, or there is indeed no significant correlation between both variables. Similarly, we did not observe significant correlation of medication to miRNA abundance. Although we found a certain tendency for several miRNAs, no correlation remained significant at an alpha level of 0.05 after adjustment for multiple testing. As for the correlation to medication and disease duration, the correlation to other markers may become significant if larger cohorts are tested.

5. Conclusion

In this study, we performed a blinded validation of a US case-control study on AD with German patients and controls that show comparable age and gender distribution. In general, both cohorts showed a very substantial degree of concordance. In this study, the medication of patients and the duration of the disease had just a very limited influence on the AD patients' miRNA profiles. Increased cohorts are required, however, to provide further evidence that miRNA signatures are indeed not correlated to the disease duration or therapy. Beyond distinguishing between AD patients and unaffected controls, we also report differences in miRNA abundance between AD, MCI, and MS patients. Especially the comparison of AD and MCI patients may contribute to the in-time detection of patients. Because small sets of markers were sufficient to perform accurate diagnosis, a clinical application on established platforms such as RT-PCR seems to be feasible.

Acknowledgments

This study has been funded by Saarland University, Siemens Healthcare, and the BestAgeing grant from the FP7 program.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jalz.2015.12.012.

RESEARCH IN CONTEXT

1. Systematic review: We previously published an article on Alzheimer micro RNAs (miRNAs) and a systematic review on novel molecular Alzheimer biomarkers. The result was that almost all novel markers require additional validation.
2. Interpretation: Our findings suggest that miRNA signatures can be well validated and may contribute to in-time diagnosis of Alzheimer's disease and on the long term to improved patient care.
3. Future directions: The next reasonable step is the validation of a cohort of around 1000 individuals using the markers that were significant in both screening and validation study using another technology such as RT-PCR.

References

[1] Prince M, Jackson J. World Alzheimer Report 2009. M. Prince and J. Jackson. London, England: Alzheimer's Disease International; 2009.
[2] Mattsson N, Zetterberg H, Hansson O, Andreasen N, Parnetti L, Jonsson M, et al. CSF biomarkers and incipient Alzheimer disease in patients with mild cognitive impairment. JAMA 2009; 302:385–93.
[3] Zafari S, Backes C, Meese E, Keller A. Circulating biomarker panels in Alzheimer's disease. Gerontology 2015;61:497–503.
[4] Doecke JD, Laws SM, Faux NG, Wilson W, Burnham SC, Lam CP, et al. Blood-based protein biomarkers for diagnosis of Alzheimer disease. Arch Neurol 2012;69:1318–25.
[5] Mapstone M, Cheema AK, Fiandaca MS, Zhong X, Mhyre TR, MacArthur LH, et al. Plasma phospholipids identify antecedent memory impairment in older adults. Nat Med 2014;20:415–8.
[6] Keller A, Leidinger P, Bauer A, Elsharawy A, Haas J, Backes C, et al. Toward the blood-borne miRNome of human diseases. Nat Methods 2011;8:841–3.
[7] Schultz NA, Dehlendorff C, Jensen BV, Bjerregaard JK, Nielsen KR, Bojesen SE, et al. MicroRNA biomarkers in whole blood for detection of pancreatic cancer. JAMA 2014;311:392–404.
[8] Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, et al. A blood based 12-miRNA signature of Alzheimer disease patients. Genome Biol 2013;14:R78.
[9] Kappel A, Backes C, Huang Y, Zafari S, Leidinger P, Meder B, et al. MicroRNA in vitro diagnostics using immunoassay analyzers. Clin Chem 2015;61:600–7.
[10] Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Research 2012;40:37–52.
[11] Keller A, Leidinger P, Borries A, Wendschlag A, Wucherpfennig F, Scheffler M, et al. miRNAs in lung cancer—Studying complex fingerprints in patient's blood cells by microarray experiments. BMC Cancer 2009;9:353.
[12] Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, et al. GeneTrail—advanced gene set enrichment analysis. Nucleic Acids Res 2007;35(Web Server issue):W186–92.

[13] Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH, et al. miRTarBase update 2014: An information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res 2014; 42(Database issue):D78–85.
[14] Dweep H, Gretz N. miRWalk2.0: A comprehensive atlas of microRNA-target interactions. Nat Methods 2015;12:697.
[15] Meder B, Backes C, Haas J, Leidinger P, Stahler C, Grossmann T, et al. Influence of the confounding factors age and sex on microRNA profiles from peripheral blood. Clin Chem 2014;60:1200–8.
[16] Kozomara A, Griffiths-Jones S. miRBase: Annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 2014; 42(Database issue):D68–73.
[17] Thomas T, Miners S, Love S. Post-mortem assessment of hypoperfusion of cerebral cortex in Alzheimer's disease and vascular dementia. Brain 2015;138(Pt 4):1059–69.
[18] Dal Pra I, Armato U, Chioffi F, Pacchiana R, Whitfield JF, Chakravarthy B, et al. The Abeta peptides-activated calcium-sensing receptor stimulates the production and secretion of vascular endothelial growth factor-A by normoxic adult human cortical astrocytes. Neuromolecular Med 2014;16:645–57.
[19] Mateo I, Llorca J, Infante J, Rodriguez-Rodriguez E, Fernandez-Viadero C, Pena N, et al. Low serum VEGF levels are associated with Alzheimer's disease. Acta Neurol Scand 2007; 116:56–8.
[20] Sonoda Y, Mukai H, Matsuo K, Takahashi M, Ono Y, Maeda K, et al. Accumulation of tumor-suppressor PTEN in Alzheimer neurofibrillary tangles. Neurosci Lett 2010;471:20–4.
[21] Kerr F, Rickle A, Nayeem N, Brandner S, Cowburn RF, Lovestone S, et al. PTEN, a negative regulator of PI3 kinase signalling, alters tau phosphorylation in cells by mechanisms independent of GSK-3. FEBS Lett 2006;580(13):3121–8.
[22] Zhang X, Li F, Bulloj A, Zhang YW, Tong G, Zhang Z, et al. Tumor-suppressor PTEN affects tau phosphorylation, aggregation, and binding to microtubules. FASEB J 2006;20:1272–4.
[23] Absalon S, Kochanek DM, Raghavan V, Krichevsky AM. MiR-26b, upregulated in Alzheimer's disease, activates cell cycle entry, tau-phosphorylation, and apoptosis in postmitotic neurons. J Neurosci 2013;33:14645–59.
[24] Bauer AS, Keller A, Costello E, Greenhalf W, Bier M, Borries A, et al. Diagnosis of pancreatic ductal adenocarcinoma and chronic pancreatitis by measurement of microRNA abundance in blood and tissue. PLoS One 2012;7:e34151.
[25] Tomashevski A, Husseman J, Jin LW, Nochlin D, Vincent I. Constitutive Wee1 activity in adult brain neurons with M phase-type alterations in Alzheimer neurodegeneration. J Alzheimers Dis 2001;3:195–207.

Published online 25 February 2016

Nucleic Acids Research, 2016, Vol. 44, No. 8 ***3865–3877***
doi: 10.1093/nar/gkw116

Distribution of miRNA expression across human tissues

Nicole Ludwig[1], Petra Leidinger[1], Kurt Becker[2], Christina Backes[3], Tobias Fehlmann[3], Christian Pallasch[4,5], Steffi Rheinheimer[1], Benjamin Meder[6,7,8], Cord Stähler[9], Eckart Meese[1] and Andreas Keller[3,*]

[1]Institute of Human Genetics, Saarland University, Medical School, Homburg, Germany, [2]Institute of Anatomy and Cell Biology, Saarland University, Medical School, Homburg, Germany, [3]Chair for Clinical Bioinformatics, Saarland University, Saarbruecken, Germany, [4]Department I of Internal Medicine and Center of Integrated Oncology, University Hospital of Cologne, Cologne, Germany, [5]Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), Cologne, Germany, [6]Department of Internal Medicine III, University Hospital Heidelberg , 69120 Heidelberg, Germany, [7]German Center for Cardiovascular Research (DZHK), 69120 Heidelberg, Germany, [8]Klaus Tschira Institute for Integrative Computational Cardiology, D-69118 Heidelberg, Germany and [9]Siemens Healthcare, Hartmannstrasse 16, 91052 Erlangen, Germany

Received May 06, 2015; Revised February 17, 2016; Accepted February 17, 2016

ABSTRACT

We present a human miRNA tissue atlas by determining the abundance of 1997 miRNAs in 61 tissue biopsies of different organs from two individuals collected post-mortem. One thousand three hundred sixty-four miRNAs were discovered in at least one tissue, 143 were present in each tissue. To define the distribution of miRNAs, we utilized a tissue specificity index (TSI). The majority of miRNAs (82.9%) fell in a middle TSI range i.e. were neither specific for single tissues (TSI > 0.85) nor housekeeping miRNAs (TSI < 0.5). Nonetheless, we observed many different miRNAs and miRNA families that were predominantly expressed in certain tissues. Clustering of miRNA abundances revealed that tissues like several areas of the brain clustered together. Considering -3p and -5p mature forms we observed miR-150 with different tissue specificity. Analysis of additional lung and prostate biopsies indicated that inter-organism variability was significantly lower than inter-organ variability. Tissue-specific differences between the miRNA patterns appeared not to be significantly altered by storage as shown for heart and lung tissue. MiRNAs TSI values of human tissues were significantly ($P = 10^{-8}$) correlated with those of rats; miRNAs that were highly abundant in certain human tissues were likewise abundant in according rat tissues. We implemented a web-based repository enabling scientists to access and browse the data (https://ccb-web.cs.uni-saarland.de/tissueatlas).

INTRODUCTION

Knowing the expression and distribution of different molecule classes in tissues is essential for the understanding of both physiological and pathological mechanisms. The gene expression atlas (1), hosted at the European Bioinformatics Institute, collects gene expression patterns under different biological conditions in various organisms. Likewise, the Human Protein Atlas presents information on proteomes in various tissues (2). For the class of small non-coding nucleic acids, the so-called microRNAs or miRNAs, there is a lack of up-to-date databases showing their tissue-specific distribution. The first and as of now most comprehensive analysis of miRNA abundance in different tissues has been reported by Landgraf et al. in 2007 (3). This sequencing-based study reported 340 miRNAs in 26 organs. We recently investigated the miRNA repertoire of different blood cell types (4), already indicating a complex miRNA repertoire strongly dependent on the considered cell types. To improve the understanding of the miRNA abundance in human tissues, we now profiled 1997 different mature miRNAs for 61 tissues. In contrast to the previous catalogue of miRNAs in human tissues, we measured all miRNA profiles from only two different individuals to minimize inter-individual variability. We selected an array-based analysis to have a robust platform for determining the miRNA abundance. The applied Agilent microarray technology has been proven sensitive and, more important, reproducible in a recent comprehensive platform comparison (5). Using this technology, we achieved technical Pearson correlation co-

*To whom correspondence should be addressed. Tel: +49 681 302 68611; Fax: +49 681 302 58094; Email: andreas.keller@ccb.uni-saarland.de

efficients of between 0.97 and 1 for technical replicates in previous studies.

Here, we first characterize technical stability of our approach before we describe variations in the abundance of the miRNAs across tissues. To provide easy access to the tissue atlas, we implemented a web-based repository that also links results to important miRNA resources. This web service is freely available online at https://ccb-web.cs.uni-saarland.de/tissueatlas.

MATERIALS AND METHODS

Tissues and RNA extraction

Tissues analysed in this study originated from two male bodies. Both cadavers were obtained as anatomical gift to be dissected in a study of medicine under German law. The first body was from a 65-year-old male patient, who suffered from multiple myeloma, a cancer that forms in a type of white blood cells (plasma cells). The body was stored at 4°C upon arrival at the anatomical institute and tissue samples were collected the following day, i.e. 2 days post-mortem. In total, we analysed 24 different tissues, i.e. adipocytes, arachnoid mater, artery, colon, small intestine (ileum), dura mater, brain, urinary bladder, skin, myocardium, bone (rib), liver, lung, stomach, spleen, muscle, gall bladder, muscle fascia, epididymis, intercostal nerve, kidney, thyroid, testis and tunica albuginea of testis.

The second body was from a 59-year-old male individual, who died a natural death. The body was frozen at −20°C after arrival at the anatomical institute and dissected after 3 weeks of storage. Autopsy showed no signs of cancer. As we aimed at increasing the resolution our tissue atlas, we collected 37 samples including several sub-areas for different organs, i.e. nine brain areas (grey matter, white matter, frontal, temporal, occipital, nucleus caudatus, thalamus, pituitary gland and cerebellum), dura mater, spinal cord, nerve, artery, vein, myocard, muscle, lymph node, thyroid, esophagus, stomach, pancreas, duodenum, jejunum, colon, liver, three kidney areas (kidney unspecified, medulla and cortex), spleen, adrenal gland, prostate, testis, skin, adipocyte, lung, pleura and bone marrow.

To assess the influence of RNA degradation originating from different storage times of the tissue on the miRNA profile, we used normal lung and normal heart tissue that was stored in physiological salt solution at 4°C for 1, 2, 3, 7 and 14 days, before RNA isolation. To understand short-term effects on the miRNA pattern in a comprehensive manner, we analysed lung tissue from another individual. The following 16 time points were profiled: 0, 0.5, 1, 1.5, 2, 3, 4, 5, 6, 9, 12, 24, 36, 48, 72 and 96 h.

To estimate inter-individual variations, we exemplarily performed in-depth analysis for lung tissues. For 16 normal tissue biopsies from different individuals, the miRNA expression intensity was determined as for the two bodies and the samples from the degradation analysis.

RNA isolation and integrity

RNA was isolated using the miRNeasy Mini Kit (Qiagen) and the Qiagen tissue lyser using 7 mm stainless steel beads. Tissue samples were disrupted for 5 min 30 Hz (1800 oscillations/min) in Qiazol lysis reagent. Further purification was done according to manufacturer's instructions. Concentration and purity was measured using NanoDrop 2000 (Thermo Scientific). RNA integrity was measured using Bioanalyzer RNA Nano Chip (Agilent). As expected for autopsy samples, the RNA integrity values (RIN) ranged between 1.8 and 2.7.

miRNA profiling

Microarray analysis was performed using *SurePrint* 8 × 60K Human V19 *miRNA* microarrays (Agilent) that contain 2007 miRNAs of miRBase V19 (http://www.mirbase.org/), according to the manufacturer's instructions for the first corpse. For the second corpse, the most recent miRBAse v21 has been used and the analysis has been carried out on 1997 human miRNAs present in both versions. In brief, a total of 100 ng RNAs were processed using the miRNA Complete Labeling and Hyb Kit to generate fluorescently labelled miRNA. Microarrays were scanned with the Agilent Microarray Scanner at 3 μm in double path mode. Microarray scan data were further processed using Agilent Feature Extraction software. The raw expression intensity values are available for download at https://ccb-web.cs.uni-saarland.de/tissueatlas. Since the normalization may have an impact on the results, we performed all analyses on the raw data, normalized data by quantile normalization and by variance stabilizing normalization (6). For training the Variance Stabilized Normalization (VSN) model all samples and all miRNAs were used. The detailed results for the variance stabilizing normalization are provided in the supplementary material. To account for negative values (i.e. miRNAs that are not expressed, that may get a negative value due to background subtraction) a pseudo-count has been added. All calculations have been carried out in R version 3.0.2.

Tissue specificity index

To evaluate the variability of expression patterns, we calculated a tissue specificity index (TSI) for each miRNA analogously to the TSI 'tau' for mRNAs originally developed by Yanai et al. (7). This specificity index is a quantitative, graded scalar measure for the specificity of expression of a miRNA with respect to different organs. The values range from 0 to 1, with scores close to 0 represent miRNAs expressed in many or all tissues (i.e. housekeepers) and scores close to 1 miRNAs expressed in only one specific tissue (i.e. tissue-specific miRNAs). Specifically, the TSI for a miRNA j is calculated as

$$\mathrm{tsi}_j = \frac{\sum_{i=1}^{N}(1 - x_{j,i})}{N-1},$$

where N corresponds to the total number of tissues measured and $x_{j,\mathrm{i}}$ is the expression intensity of tissue i normalized by the maximal expression of any tissue for miRNA j.

Hierarchical clustering of tissues

To estimate the proximity of profiles from different tissues, hierarchical clustering analysis has been applied. To ac-

count for the high dynamic range of miRNAs, clustering has been performed on log expression intensities and miRNAs that are close to the background were removed. To extend the cluster analysis, the 100 most variable miRNAs have been selected. In each case, complete linkage hierarchical clustering using the Euclidian distance has been performed.

Expression of miRNA families

For estimating the tissue specificity of miRNA families, we extracted all miRNA families from the most recent miRBase version 21. For each miRNA precursor all mature forms have been considered as family members, duplicated mature miRNAs (e.g. coming from different precursors in the same family) have been counted once in order to minimize a potential bias introduced by multiple precursors. For discovering co-expressed miRNAs, Spearman correlation of intensity values between all pairs of miRNAs has been calculated. Network visualization has been performed in Cytoscape.

Conservation of tissue specificity

To compare conserved tissue specificity in humans and rats, we downloaded data from the Gene Expression Omnibus (GEO) series GSE52754, containing expression profiles for 55 different rat tissues that have been measured using Agilent microarrays (8). To match miRNAs we extracted all rat miRNA identifiers from the respective manuscript and matched them via a 100% sequence match. For matching miRNAs and matching tissues, we calculated and correlated the tissue specificity indices. To minimize artefacts introduced by normalization, we carried out all analyses on raw data. Since this analysis only addresses the question whether a miRNA is rather specific or a housekeeping miRNA, we also correlated the human and rat expression profiles using Spearman correlation.

Additional data from literature

In addition to the 44 tissue samples from the degradation and reproducibility analysis, the 16 individual lung cancer tissues and the 61 tissues from two bodies newly measured for this study, we searched the literature for other studies where normal tissues have been profiled. In the GEO (9), we found 1178 series related to miRNAs. Of these, 722 were from *Homo sapiens*. Excluding series with low sample count (below 20 samples), 302 series remained. After excluding studies from body fluids such as serum, plasma, blood or urine, we examined the remaining hits for availability of unaffected tissue measurements. The respective data tables were downloaded from GEO and all IDs were matched from the respective platform identifiers to miRBase Version 21 IDs. For the respective studies, raw and normalized data (VSN and quantile normalized) were added to our tissue atlas web repository. These include 43 samples from 9 tissues and 463 miRNAs from GSE11879, 40 samples measured for 709 miRNAs from normal gastric tissue from GSE23739, 48 benign prostate tissues measured for 480 miRNAs from series GSE54516 and 32 benign prostate tissues measured for 825 miRNAs from series GSE76260. The data have been used partially in the present manuscript, all data are included in the web-based tissue atlas resource.

RESULTS

In this work, we present the draft of a human tissue miRNome atlas. In the first part of the manuscript, we describe pre-analytics, investigating the general reproducibility of the miRNA profile measurements and also the effect of storage of tissues on miRNA profiles. In the pre-analytics consideration, we measured 44 tissue miRNomes. It is essential to understand respective variability to understand the biological variability of different tissue miRNomes.

In the second part, we describe the screening of all mature miRNAs from miRBase version 21 across different organs of two male bodies. We investigated miRNA expression in 24 different tissues from the first body and in 37 different tissues from the second body. To determine the miRNAs abundance in the different tissues, we utilized a TSI score, known from transcriptomics (7). Furthermore, we investigated the proximity of organs based on miRNA abundances by hierarchical clustering and co-expression analysis. To estimate inter-individual variations, we measured 16 additional miRNomes from control lung tissues and extracted further data sets from the GEO.

To provide researchers access to the first version of the miRNA tissue atlas, we implemented a web-based repository that is freely available at www.ccb.uni-saarland.de/tissueatlas.

Reproducibility of miRNA patterns

An important factor for estimating the biological variability is to understand the technical variability of the underlying profiling platform. Previously, we compared technical reproducibility of the two common platforms, microarrays (Agilent) and NGS (Illumina HiSeq) (10). In these experiments, we discovered an increased variability of miRNAs dependent of the sequencing library preparation. Similarly, we observed a strong bias based on the nucleotide composition of miRNAs (11). Of 10 replicated Agilent microarray measurements of the same individual, we calculated 10 * 9/2 pair-wise correlations of technical replicates. Minimal correlation was 0.998 and mean/median correlation 0.999, highlighting the high degree of technical reproducibility of the array platform. To translate these results on our tissue atlas and determine technical reproducibility of the array analysis, technical duplicates from nine randomly selected tissue samples from the second body were measured. The duplicates were processed at different days and have been measured on different arrays, each. Hierarchical cluster analysis shows that the technical replicates always clustered together showing that the applied technology was suited to provide reproducible results (Figure 1 shows the heat map for quantile-normalized data, Supplementary Figure S1 for VSN-normalized data). Altogether, we found high correlations between these technical replicates with the overall lowest correlations at 0.986 and 0.994 observed for liver tissue and pleura, respectively. Highest correlation of 0.999 was reached for the brain samples.

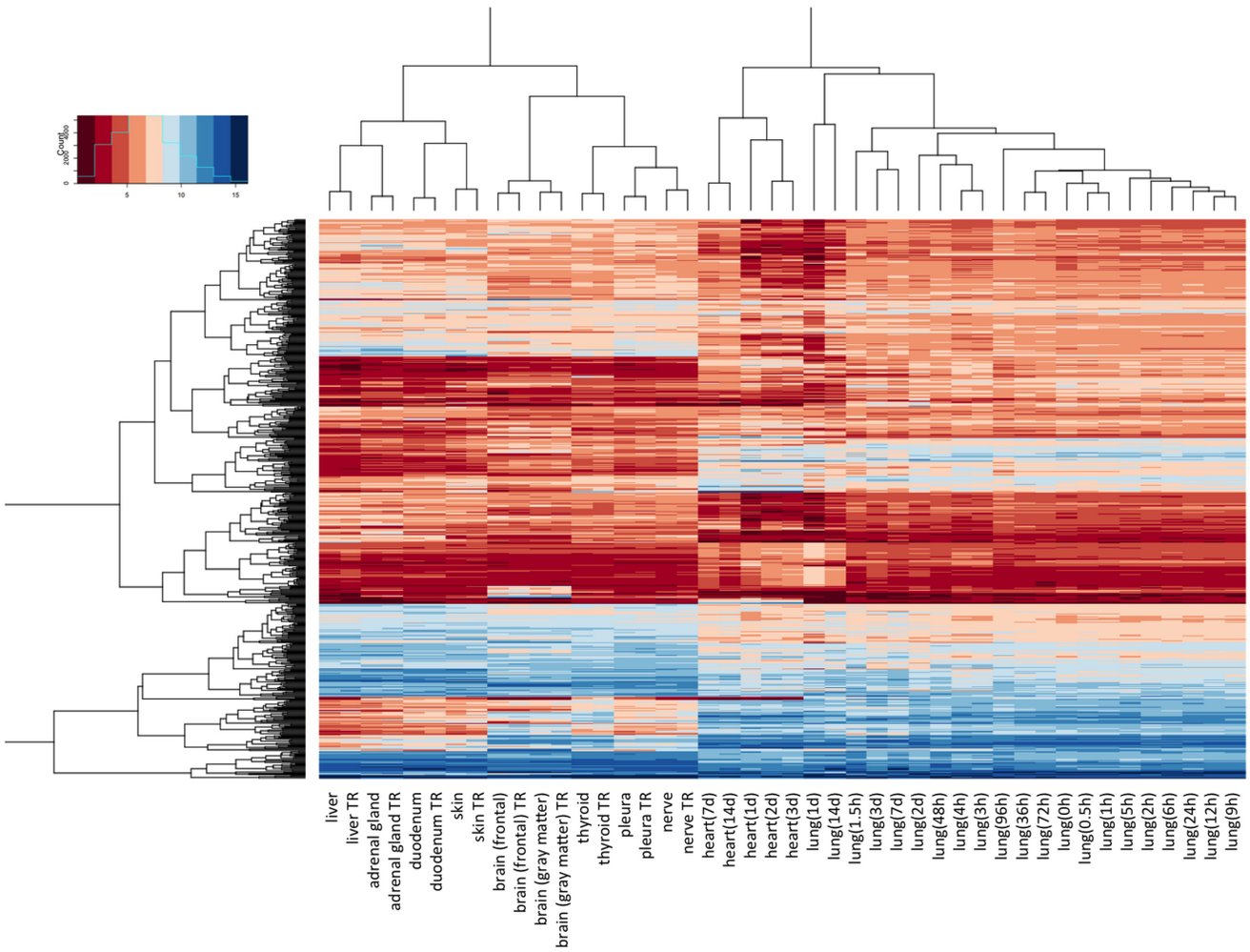


Figure 1. Hierarchical clustering of the 44 samples included in the stability and reproducibility study. Quantile normalized and $\log_2$ transformed expression intensity values were used for clustering. The intensity values and distribution are presented in the upper left corner. In the present heat map, heart and lung tissues cluster together on the right-hand side. Technical replicates (marked by 'TR' in the labels below the heat map) of other organs cluster together in each case in the left-hand side. For VSN-normalized data the same representation is provided in Supplementary Figure S1.

Stability of miRNA patterns in tissues

Measuring tissues of corpses the storage time prior to RNA extraction and a potential degradation of RNA may have an influence on the profiles. We exemplarily investigated the process for heart and lung tissue. Biopsies were taken from two individuals and have been stored for 1, 2, 3, 7 and 14 days at 4°C. Hierarchical cluster analysis shows that all lung and all heart samples each cluster together (Figure 1; Supplementary Figure S1). The duration of the storage was, however, not reflected in the clustering pattern indicating that a storage time between 1 and 14 days at 4°C has a limited influence on the overall miRNA tissue pattern.

We also performed the analysis with more dense time intervals within the first 3 days to understand short-term effects. For a lung tissue from a third individual 16 time points between 0 and 96 h were profiled. These biopsies clustered well together with the lung tissues from the second individual with storage time over 14 days. Again, no time curse could be recognized in the clustering pattern.

Remarkably, the results presented above describe the overall miRNA patterns. For single miRNAs still differences dependent on the storage could be observed. Thus, we calculated the TSI for all lung tissues and for all tissues in the pre-analytical study. With respect to lung tissues, large TSI values mean in this case not tissue specific but rather specific in one of the replicated measurements. We thus expect that TSI values of miRNAs from the lung tissue are low. Especially for five miRNAs we, however, calculated TSI values that are increased in lung tissue by at least 20%: hsa-miR-8069, hsa-miR-6821–5p, hsa-miR-4800–5p, hsa-miR-6775–5p, hsa-miR-5001–5p. For all miRNAs, TSI values from the pre-analytical step are summarized in Supplementary Table S1.

Frequency of miRNAs per tissue and tissue specificity of miRNAs

For each miRNA in each tissue, we determined its presence and frequency using the so-called present calls as determined by Agilent Feature Extraction software. Out of the 1997 different mature miRNAs, 633 (31.7%) were not detected in any of the tested tissues by the applied microarray technology. Out of the remaining 1364 miRNAs, 143 (10.5%) were found in all tissues. To present more comprehensive information on the tissue distribution of miRNAs, we utilized the miRNA TSI analogously to the mRNA TSI 'tau' that has successfully been employed by Yanai et al. (7). This index has a range of 0–1 with the score of 0 corresponding to ubiquitously expressed miRNAs (i.e. 'housekeepers') and a score of 1 for miRNAs that are expressed in a single tissue (i.e. 'tissue-specific' miRNAs). We calculated TSI for the 1364 miRNAs that have been detected in at least one tissue sample. For each miRNA, we compared TSI for the two bodies, for raw, quantile- and VSN-normalized data (Supplementary Table S2). Using the quantile-normalized data for the first body, 83.7% of all miRNAs showed an average abundance throughout the tissues with intermediate TSI values ranging from 0.15 to 0.85 (Figure 2A, Supplementary Figure S2A for VSN-normalized data). Only one miRNA (miR-3960) was ubiquitously expressed with a TSI < 0.15 and 222 miRNAs showed a highly tissue-specific expression with TSI > 0.85. For the second body, 88.8% of all miRNAs showed intermediate TSI values; one miRNA (miR-6089) showed a TSI < 0.15 and 152 miRNAs a TSI > 0.85 (Figure 2B, Supplementary Figure S2B for VSN-normalized data). The correlation of the VSN-normalized TSI values with the quantile-normalized TSI values was 0.88 ($P < 10^{-10}$).

The overall most tissue-specific miR-1–3p is presented in Figure 3. For all 61 samples raw-, quantile- and VSN-normalized expression intensities are presented as bar plot. Respective bar plots for all miRNAs can be generated using the online repository.

Clustering of tissue patterns and analysis of miRNA families

Beyond the analysis of single miRNAs, we determined the overall similarity/dissimilarity of the miRNA pattern between the different tissues. We performed hierarchical clustering of miRNAs and tissues using normalized expression intensities. We found two major clusters, the first of which containing mainly nervous system tissues and muscle tissues from both bodies. In the second cluster, the organs of the two individuals frequently did not cluster together (Figure 4A). Since the large number of miRNAs used for this clustering likely caused substantial noise, we restricted the clustering analysis to the 100 miRNAs with the highest data variance (Figure 4B). Here, we found three main clusters with the first one containing kidney, liver, stomach and small intestine of both bodies. The second cluster exclusively contained all brain tissue samples of both bodies and nervous system related tissue, i.e. spinal cord and dura mater. The third cluster contained thyroid, nerve, muscle, myocardium and colon each of both bodies. Other organs were found in different clusters, e.g. the lung samples and the brain coverings dura mater and arachnoid mater. For VSN-normalized data we observed a similar pattern, however, we found a stronger tendency of clustering of individuals in the different sub-clusters (Supplementary Figure S3).

To gain further insights into expression of tissue-specific miRNAs, we performed clustering with the 25 miRNAs displaying a TSI > 0.85 for both bodies in raw-, quantile- and VSN-normalized data (Figure 5). We found several groups of miRNAs with tissue-specific expression. In detail, we detected high expression of miR-133b, miR-133a-3p, miR-1–3p and miR-206 in both muscle samples and, with the exception of miR-206 also in both myocardial samples. Additionally, we found a cluster of four miRNAs specifically expressed in various brain tissues, i.e. miR-338–3p, miR-219a-5p, miR-124–3p and miR-9–5p. Another group of miRNAs, miR-507, miR-514a-3p and miR-509–5p was almost exclusively expressed in the testis samples. Besides these miRNA clusters, we also found single miRNAs that were expressed in a highly tissue-specific manner, i.e. miR-122–5p, miR-7–5p and miR-205–5p were each exclusively expressed in liver, pituitary gland and skin, respectively.

Tissue specificity of miRNA families

To further determine to what extend miRNA families show similar abundances in specific organs, we calculated the TSI not only for single miRNAs but also for mature miRNAs inside each miRNA family. Out of 187 miRNA families from the miRBase with at least two family members, we analysed 25 miRNA families with at least five mature forms (Figure 6A; Supplementary Table S3). We found several miRNA families with high TSI values including the above-mentioned mir-378 family with most of the family members showing a high abundance in muscle tissues and the myocardium. Similarly, the mir-506 family with 18 family members showed generally a high abundance in testis tissue while they were less expressed in other tissues. Other families, such as the mir-449 family with five members, did not show a common pattern in the different tissues: MiR-449c-3p was expressed specifically in spleen tissue, miR-449c-5p and -449b-5p in kidney and small intestine, miR-449a in lung, kidney and brain and miR-449b-3p in spleen. To extend this analysis we searched for miRNAs co-expression patterns in specific tissues. We used a high correlation cut-off and considered only miRNA-pairs with Pearson correlation exceeding 0.95. Altogether, we identified 73 miRNA pairs with tissue co-expression. In addition to pair-wise interactions, we also found sub-networks with at least four participants. The networks have been visualized using Cyto-Scape (Figure 6B). While we frequently observed co-expression among mature members of specific families (e.g. the mir-548 family), we also found correlations of miRNAs from different miRNA families. For example, miR-4312 was co-expressed with miRNAs from the let-7 family. Performing the same analysis with raw data, we detected an increased number of co-expressions, but generally confirmed the observation that has been based on the normalized data.

Tissue specificity of -3p and -5p mature forms

We asked whether -3p and -5p mature forms of miRNAs have different tissue specificity. To limit the bias of miR-

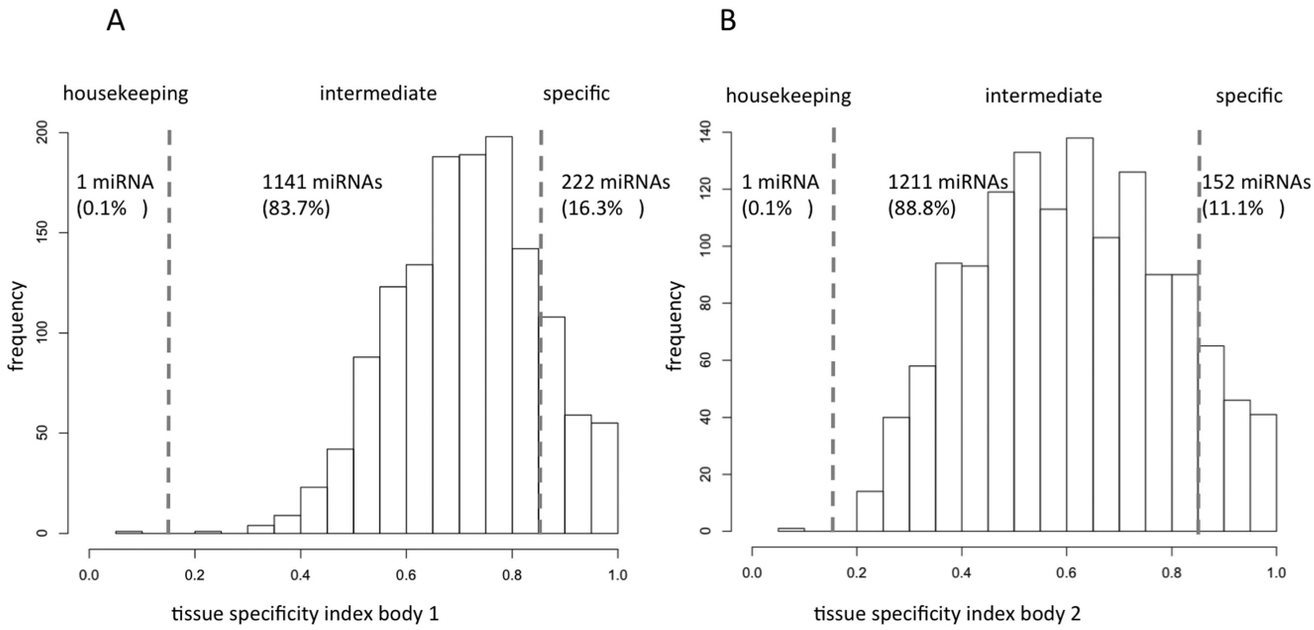


Figure 2. Histogram plot for the frequency of TSI of miRNAs in different tissues. Panel **A** represents TSI of the first, panel **B** of the second body. The vertical dotted lines correspond to the threshold originally proposed for defining housekeeping and specifically expressed miRNAs of <0.15 and >0.85. The same representation for VSN-normalized data is presented in Supplementary Figure S2.

NAs that are annotated with only one mature form, we only included those miRNAs that have two mature forms annotated and carried out the analyses in a paired manner (41% of the 1364 mature miRNAs were included). First, we investigated whether -3p or -5p mature forms are overall higher expressed. For both quantile- and VSN-normalized data, we calculated significantly higher expression of the -5p mature forms. The effects in VSN exceeded the quantile-normalized effects. Mature -5p forms were on average 21% higher expressed as compared to -3p forms (paired t-test *P*-value of 3.6×10^{-10}). To estimate whether the two mature forms are more or less specific for tissues, we calculated and compared the TSI values for the -3p and -5p forms. For both, TSI values based on VSN- and quantile-normalized data, we did not found significant differences between -3p and -5p forms ($P > 0.5$ in both cases). Having a detailed look at single miRNAs, we discovered that in all cases where -3p and -5p mature forms were tissue specific independent on the normalization technique the tissue patterns matched. The best matching profiles were found for hsa-miR-140, hsa-miR-378a, hsa-miR-509, hsa-miR-122, hsa-miR-124, hsa-miR-192 and hsa-miR-455. Only for one miRNA, miR-150, no significant correlation for -5p and -3p mature form was calculated (Supplementary Figure S4). The -3p form was specific for pancreas and the -5p form for stomach. All TSI values for -3p and -5p mature forms of quantile- and VSN-normalized data are available in Supplementary Table S4.

Inter-individual variations

In the previous analyses, we suggested that miRNAs are tissue specific. From two bodies it is impossible to extrapolate inter-individual variations within specific organs. In a first approach we searched for miRNAs that are overall higher or lower in all tissues of one of the two bodies, independent of the normalization technique. Two miRNAs, hsa-miR-548n and hsa-miR-548ap-5p, fulfilled these stringent criteria. Although these (and similarly differentially abundant miRNAs between both individuals) miRNAs had low TSI values and are not considered tissue specific the differences emphasize the importance of incorporating inter-individual variations.

We exemplarily analysed 16 lung tissue biopsies of 16 different individuals. Here, we expect miRNAs to be more homogenously expressed, leading to overall lower TSI values. For the quantile- and VSN-normalized data, we calculated significantly decreased TSI values in the individuals ($P < 10^{-16}$). The respective TSI values for biological replicates of lung tissue and the two bodies are presented in Supplementary Figure S5A (quantile normalized) and 5B (VSN normalized). These figures also indicate that few miRNAs have higher TSI in lung as compared to the overall TSI, i.e. variations between organs are smaller than variations between individuals. Inspecting the respective miRNAs, we found that they usually were specific for other organs than the lung and expressed to a very moderate limit in the lung. Here, already small variations lead to artificially high TSI values.

As the second example we downloaded expression values from 32 prostate tissues from the GEO (not affected tissues as part of a case-control cancer study, GSE76260). The TSI values were calculated for quantile- and VSN-normalized intensity values. Only the 625 miRNAs that were included in both studies were considered. In this analysis the variations between individuals were even lower as compared to

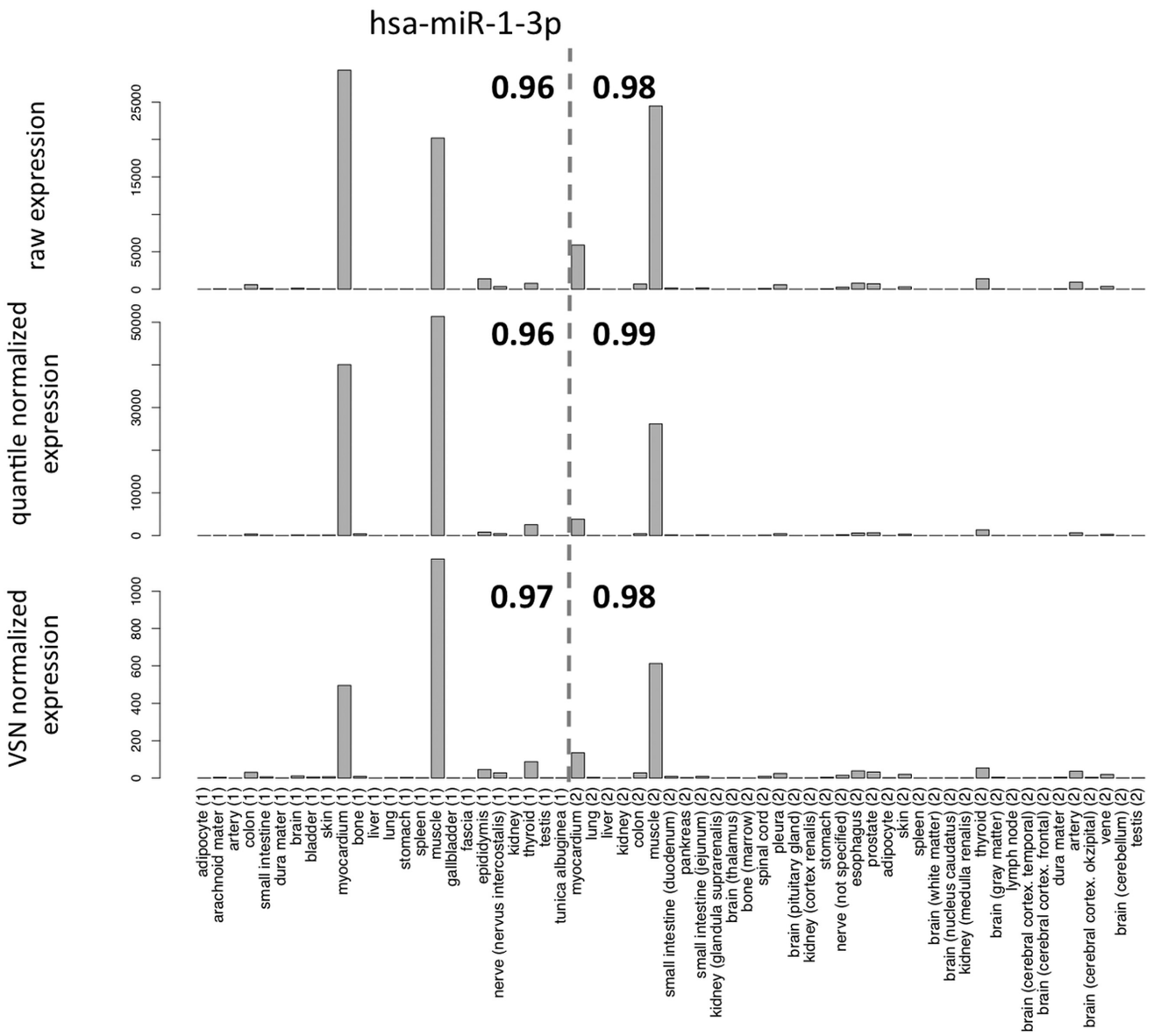


Figure 3. Bar plots for all 61 samples for miR-1–3p, the miRNA with highest overall TSI in the first and second body. The vertical dashed line separates the first from the second body. TSI values for both bodies are highlighted in the figure. The miRNA is high expressed in muscle and myocardium. Raw-, quantile- and VSN-normalized expression intensities for this miRNA match well across all different tissues.

the variations between organs. Again, TSI values were significantly lower for prostate tissue ($P < 10^{-16}$). The scatter plots are analogously to the lung tissues presented in Supplementary Figure S6. Also for the other tissues extracted from the GEO, which are also available on the tissue atlas web resource, lower TSI values were observed. In sum our results thus indicate that the inter-individual variations are smaller as compared to inter-organ variability.

Homology of tissue specificity in humans and rats

To addressed the question to what extend a tissue-specific abundance of the miRNA pattern is conserved between human and rodents, we matched the data of our study to data published in a recent study, which used the same miRNA platform (Agilent) (8). From all miRNAs expressed in our tissue collection, 230 matched in sequence identically between human and rat. Of the tissues included in the human and rat studies, 42 organs could be matched. For all these miRNAs and organs, we calculated the TSI values in human and rat, showing an overall correlation of 0.362 (P-value of 9×10^{-8}). To determine the significance of this finding, we additionally performed 1 million permutation tests, which showed an average correlation value of 0. While these results indicate an overall matching of miRNA abundances in humans and rats, the TSI does not acknowledge the origin of the miRNAs, i.e. a value of 1 for a rat miRNA may indicate

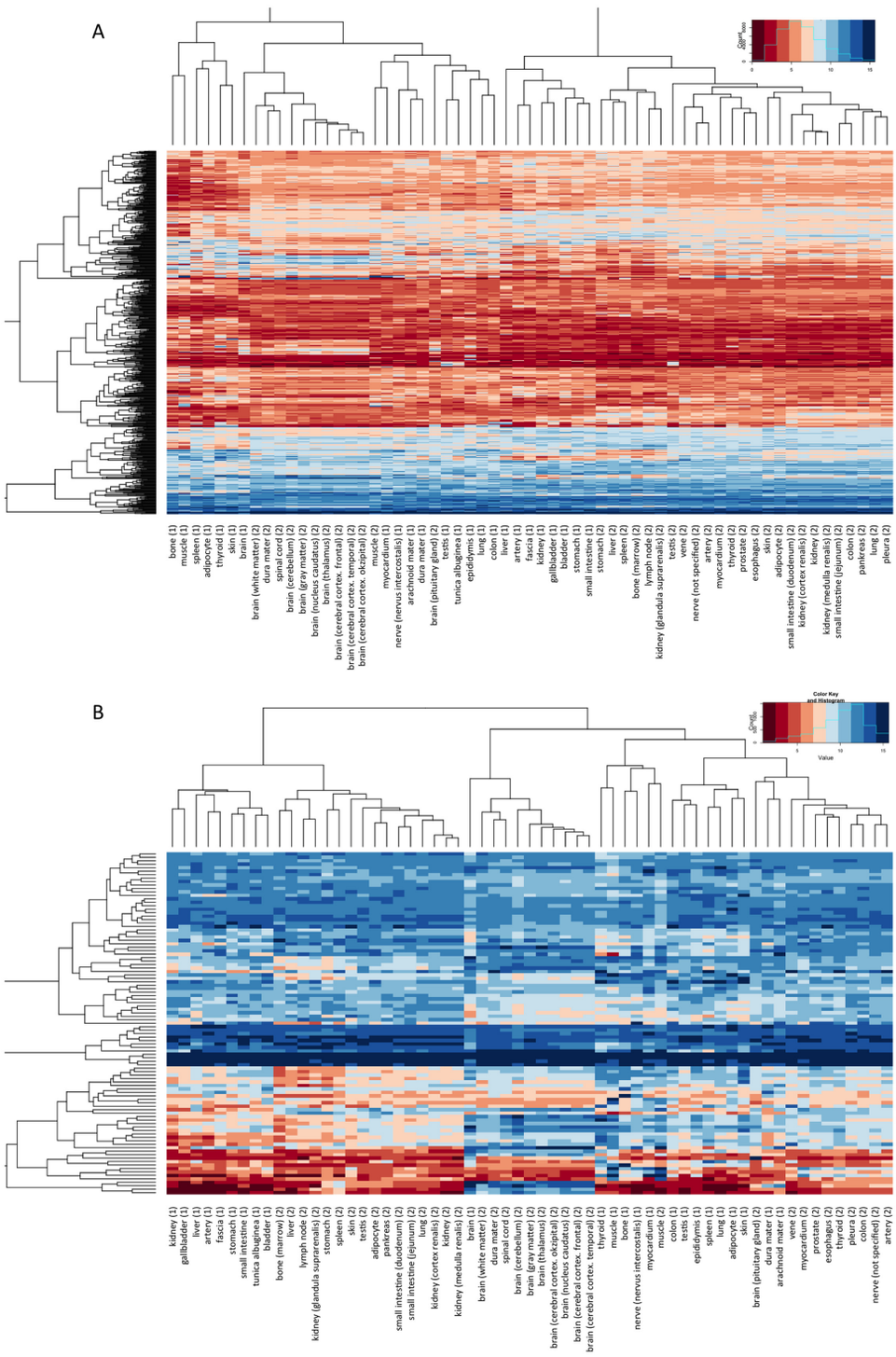


Figure 4. Hierarchical clustering of all tissues in both bodies. Log_2 transformed quantile normalized intensity values were used for clustering. The intensity value distribution is shown in the upper right corner of the figures. Panel **A** shows significantly expressed miRNAs, while panel **B** focuses on the 100 miRNAs with overall highest data variance. The respective representation for VSN-normalized data is presented in Supplementary Figure S3.

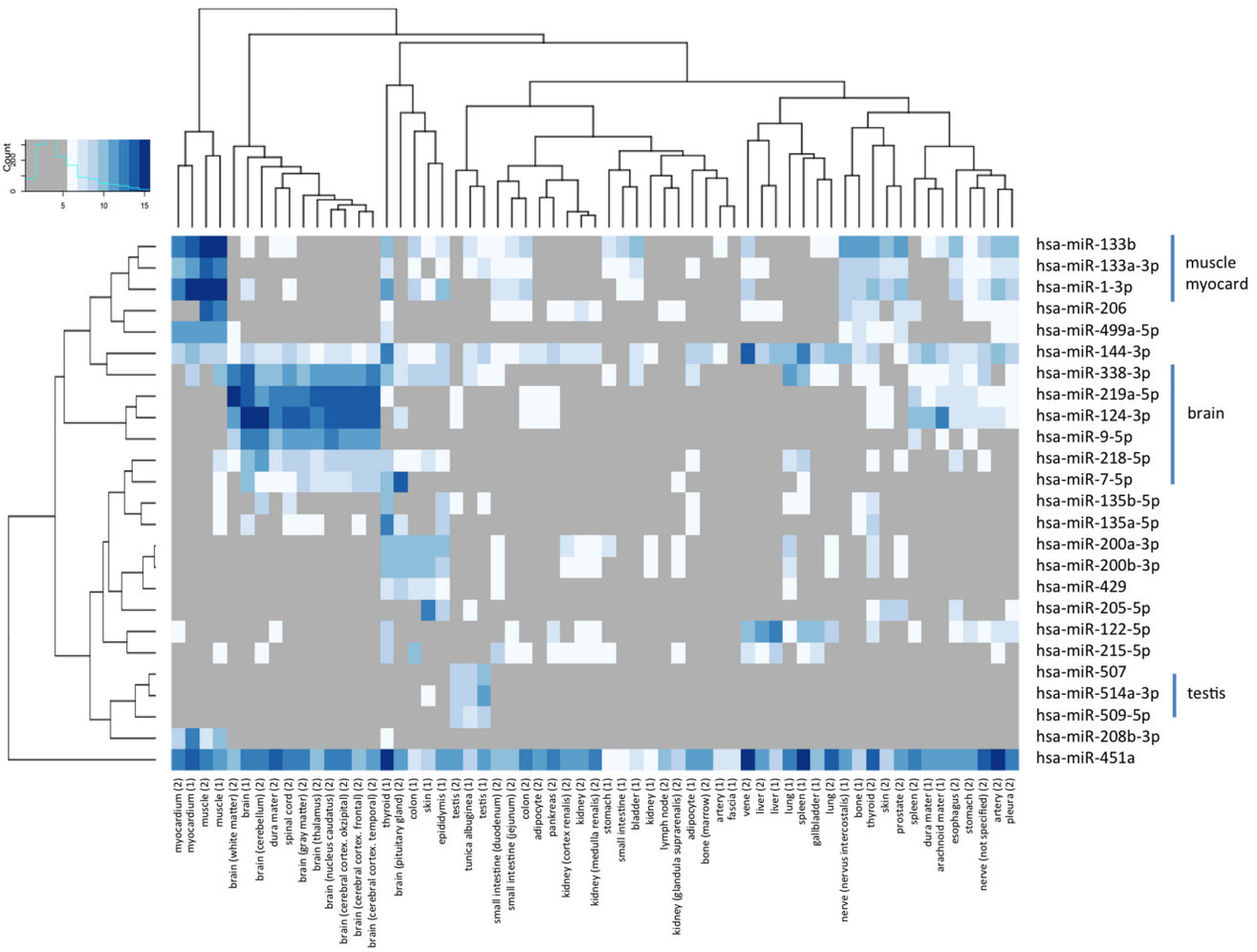


Figure 5. Heat map for the 25 miRNAs that have TSI values of >0.85 in both bodies. Log_2 transformed expression intensities of quantile normalized expression values are presented. To facilitate the interpretation of specific miRNAs in organs or organ groups low expressed miRNAs were greyed out (see also color distribution scheme in the upper left corner). The analysis highlights tissue-specific miRNAs that are exemplarily presented on the right-hand side of the plot, such as hsa-miR-1–3p that has already been described in Figure 3 as most specific miRNA overall.

specificity for spleen and for the same miRNA specificity for brain in humans. However, the overall correlation of the expression values of rat and human miRNAs was 0.361 ($P < 10^{-16}$), indicative of a significant matching of human and rat expression profiles. Similar to the results for humans in Figure 5, we clustered the miRNAs with high TSI values in human and rat. Altogether, we focused on very specific miRNAs: 54 miRNAs with TSI values exceeding 0.9 were considered. The resulting heat map where maximal rat and human miRNA expression was set to 100% to make both data sets comparable to each other is presented in Figure 7. In this analysis we did not observe a predominant clustering in humans and rats but a strong tendency of organs to cluster together. Examples of directly matching pairs include the spleen, myocardium, muscle, pancreas, kidney, liver, stomach, skin, brain or spinal cord. The miRNAs in this heat map matched the specific miRNAs in Figure 5 very well such as miR-133a-3p, and miR-133b for muscle and myocardium or miR-9–5p, miR-219a-5p, miR-7–5p and miR-124–3p for brain and spinal cord. Bar plots comparing each miRNA directly for specificity in tissues of rat and human are provided in the supplementary material.

DISCUSSION

As miRNAs emerge as important regulators of protein expression during tissue development and homeostasis, there is an increasing need for a standardized atlas of miRNA expression in multiple human tissues. Although there is ample evidence for differential miRNA expression in different human tissues, the majority of studies investigate differential expression in only one organ/tissue. Due to the different identification methods and normalization strategies, the results of these studies are not easily comparable limiting their value for comparison of miRNA expression in different tissues. The optimal human miRNA tissue atlas would be based on different fresh tissues each obtained from the same donor; different donors should be of different age and gender both of which are known to influence the miRNA

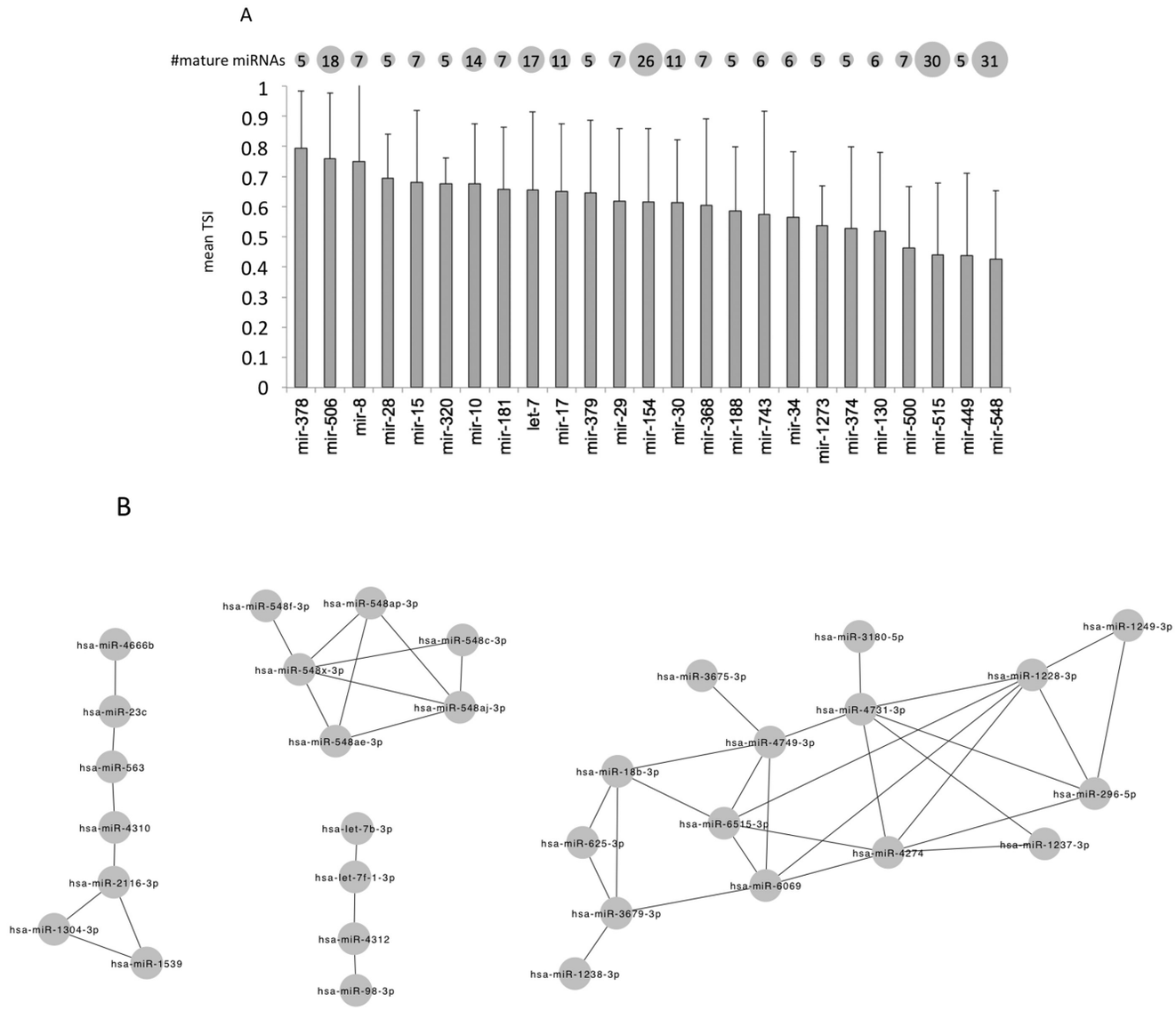


Figure 6. **A**: Average and standard deviation of TSI value in different miRNA families. For each miRNA family with at least five members the mean and standard deviation of all family members TSI is presented as bar plot. Families are sorted with decreasing average tissue specificity from left to right. Highest tissue specificity was observed for the miR-378 family, predominantly being specific for myocardium and muscle. The number of mature family members is shown above the columns with balloons, representing the family size. **B**: Co-expression network of miRNAs. Each miRNA pair connected by an edge has co-expression across all samples with Spearman correlation coefficient above 0.95.

pattern (12). As this ideal scenario is not possible in human studies, fresh biopsy material could be used for miRNA isolation with the advantage of yielding high-quality RNA. There are, however, several disadvantages: (i) biopsies will be mostly taken from patients with affected organs, (ii) high inter-individual differences can mask tissue-specific differences of miRNA abundances, (iii) a bias is likely introduced by multiple centres that are involved in tissue collections and (iv) samples of vital organs, e.g. thalamus, spinal cord or cerebellum, are not available. Alternatively, miRNAs can be isolated from tissues collected from the same individuals upon autopsy. The advantage of the latter approach is the availability of multiple tissues from the same individuals, even from vital organs, with the disadvantage of RNA degradation in the samples due to the storage duration of the body and the advanced age or the disease status of the body donors. In context of our tissue atlas, the main question is whether the differences in the abundance of miRNAs induced by post-mortem RNA degradation, which is different from *in-vitro* RNA degradation by UV light or heat, are higher than the differences between the tissues profiled. There is scant evidence for extended post-mortem stability of individual miRNAs (13,14). In case of whole miRNA tissue profiles, Ibberson et al. found that RNA degradation due to prolonged inadequate tissue storage has a random effect on miRNAs and compromises the reliability of miRNA

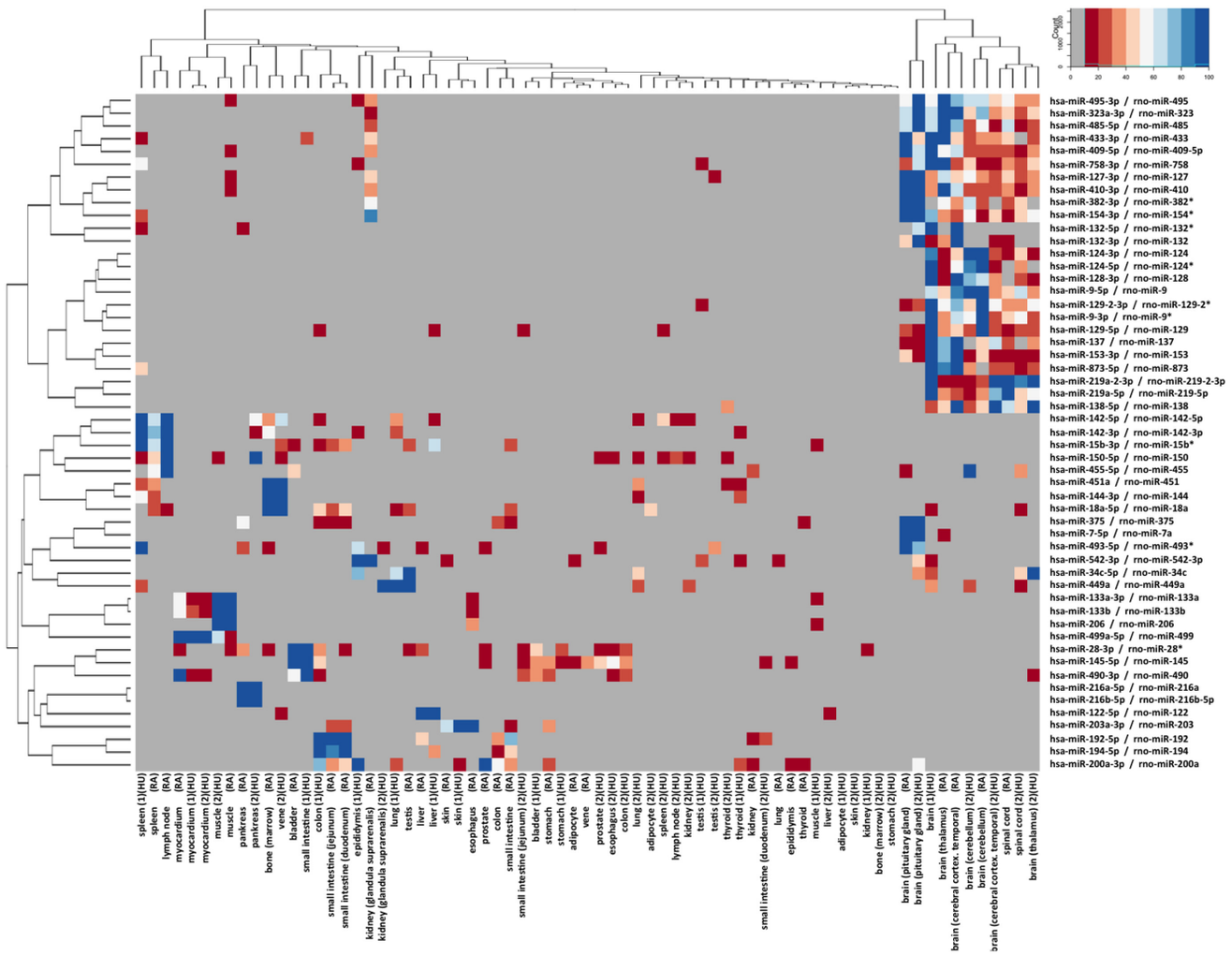


Figure 7. Conservation of tissue-specific expression of miRNAs in human and rat. Matching miRNAs (100% matching of mature miRNA sequence) from organ expression in rats and humans were calculated. For each miRNA in rats and humans the TSI was calculated and highly specific miRNAs were clustered. Since overall expression in humans and rats varied, the maximal intensity of each miRNA in the two organs was set to 100% and all other miRNAs were linearly scaled. All miRNAs with below 10% expression of maximal intensity are shown in grey to facilitate data interpretation (see also colour gradient presented in the upper right corner). On the right-hand side the human/rat miRNA identifiers are shown, below the heat map the matched tissues are presented (HU for human; RA for rat). For rat tissues the average intensity of replicated measurements is presented.

profiles, generating false positive deregulated miRNAs (15). But they also clearly state that 'even samples with the most degraded RNAs still preserve a tissue-specific miRNA signature'. This finding is in line with our observations in the present study. For lung and heart tissue we investigated short- and long-term degradation, highlighting an overall limited impact on the tissue specificity of miRNA profiles. Only very few miRNAs were affected at all. Given the data from two organs, we however cannot exclude the possibility that some tissue-specific miRNAs might be affected by degradation of the sample. We are also aware that the autopsy samples of the two male individuals provide only a snapshot of the full variability of miRNA expression. While we aim at adding more full body profiles we supported the data in the present study by tissue collections extracted from the literature (e.g. gastric and prostate tissues) and by own measurements (lung tissue).

We used a microarray platform for miRNA expression detection since this platform shows a high reproducibility as evidenced by the miRQC study (5). In our study, analysis of technical replicates of nine samples processed in different batches reached high correlation values above 0.986 for all samples. In previous studies, we observed a substantial bias introduced in Next Generation Sequencing (NGS) data by sample preparation of blood samples (10). However, NGS analysis would enable to detect presently unknown miRNAs as well miRNAs iso-forms that have demonstrated to target biological pathways in a cooperative manner (16). A key challenge with microarray data is normalization. Many techniques that are frequently applied such as variance stabilizing normalization or quantile normalization can have a substantial influence on the results. Quantile normalization e.g. assumes an overall similar distribution of all miRNAs. We thus performed the relevant analyses on raw data,

quantile- and VSN normalization. Irrespective of the normalization technique we found higher TSI values for miRNAs as, e.g. known from mRNAs (7). This result suggests that miRNA expression is more tissue specific as compared to mRNA expression.

The, as of now, most comprehensive study on tissue-specific miRNAs in humans was published by Landgraf et al. in 2007 (3). They sequenced 256 small RNA libraries from 26 different organ systems and cell types of humans and rodents, with ∼1000 clone each. The human samples included normal samples from 16 tissues most of them brain and reproductive tissues. They identified 340 mature human miRNAs including 33 novel miRNAs not listed in the miRBase version 9.1, which was the current version at the time of the study (17). For canonical miRNAs they found a high concordance of tissue-specific expression in humans and rodents. When we compared our data to a data set on 55 different rat tissues available at GEO database (8), we could confirm conserved tissue-specific expression of several miRNAs, including miR-133b, miR-124 and miR-9. Amongst others, Landgraf et al. detected tissue-specific expression of miR-122 in liver, of miR-9, miR-124 and miR 128a/b in brain, of miR-7, miR-375, miR-141 and miR-200a in pituitary gland and of miR-142, miR-144, miR-150, miR-155 and miR-223 in hematopoietic cells. Overall, our results correlated well with this data, confirming specific expression of miR-122, miR-9, miR-124 and miR-7 in the respective organs. Consistent with Landgraf's results, we found miR-122–5p as highest expressed miRNA in the liver of both bodies. Our study, however, also identified low expression of miR-122–5p in spleen, gall bladder and veins. MiR-124 (miR-124–3p) was identified as the third most specific miRNA in the nervous system by Landgraf et al. We observed expression of this miRNA in different areas of the brain but not in other tissues. For miR-144, we found highest expression in vein and spleen, consistent with the assumption of residual hematopoietic cells in these samples; additionally, we found high expression of this miRNA in thyroid. Of note, miR-144 has been found highly expressed in normal thyroid and downregulated in papillary thyroid carcinoma (18). We also found high expression of miR-1–3p, miR-133a-3p, miR-133b and miR-206 in myocard and muscle. These miRNAs are known as myomiRs that regulate key genes in muscle development (19,20). Additionally, we detected a highly specific expression of miR-205–5p, miR-514a-3p and miR-192–5p in skin, testis and colon samples of one of the bodies, respectively. MiR-205–5p that is highly expressed in melanocytes and downregulated in melanoma is inverse correlated with melanoma progression (21). MiR-514a-3p belongs to the miR-506 family; the mouse orthologue of miR-506, mmu-201, has been shown to be specifically expressed in reproductive tissues (3). A significant decrease in expression of miR-192–5p in colorectal cancer compared to normal mucosa has been reported (22).

The knowledge of the expression pattern of miRNAs in different tissues is essential for understanding normal development and disease development of the respective tissue. In addition, knowing the tissues that express specific miRNAs helps to develop a miRNA found in whole blood or serum into a biomarker for a specific disease. Elevated serum levels of liver-specific miR-122 have been detected in patients with drug induced liver injury, steatosis, hepatitis-B and -C infections and in patients with hepatocellular carcinoma (23–26). Elevated levels of circulating myomiRs, i.e. miR-1, miR-206 and miR-133a/b, have been proposed as biomarker for heart failure and different forms of muscle dystrophy, but are also elevated after half-marathon run (27–29).

In summary, we provide an atlas of miRNA expression in multiple human tissues. This atlas can be used as starting point for elucidation of the role of miRNAs in tissue development and tissue-specific diseases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We acknowledge the support of Siemens Healthcare.

FUNDING

Saarland University and Siemens Healthcare. Funding for open access charge: Saarland University and Siemens Healthcare; funded in part by FP7 project BestAgeing.
Conflict of interest statement. None declared.

REFERENCES

1. Petryszak,R., Burdett,T., Fiorelli,B., Fonseca,N.A., Gonzalez-Porta,M., Hastings,E., Huber,W., Jupp,S., Keays,M., Kryvych,N. *et al.* (2014) Expression atlas update–a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **42**, D926–D932.
2. Ponten,F., Jirstrom,K. and Uhlen,M. (2008) The human protein atlas–a tool for pathology. *J. Pathol.*, **216**, 387–393.
3. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
4. Leidinger,P., Backes,C., Meder,B., Meese,E. and Keller,A. (2014) The human miRNA repertoire of different blood compounds. *BMC Genomics*, **15**, 474.
5. Mestdagh,P., Hartmann,N., Baeriswyl,L., Andreasen,D., Bernard,N., Chen,C., Cheo,D., D'Andrade,P., DeMayo,M., Dennis,L. *et al.* (2014) Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nat. Methods*, **11**, 809–815.
6. Huber,W., von Heydebreck,A., Sultmann,H., Poustka,A. and Vingron,M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**(Suppl. 1), S96–S104.
7. Yanai,I., Benjamin,H., Shmoish,M., Chalifa-Caspi,V., Shklar,M., Ophir,R., Bar-Even,A., Horn-Saban,S., Safran,M., Domany,E. *et al.* (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21**, 650–659.
8. Minami,K., Uehara,T., Morikawa,Y., Omura,K., Kanki,M., Horinouchi,A., Ono,A., Yamada,H., Ohno,Y. and Urushidani,T. (2014) miRNA expression atlas in male rat. *Sci. Data*, **1**, 140005.
9. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
10. Backes,C., Leidinger,P., Altmann,G., Wuerstle,M., Meder,B., Galata,V., Mueller,S.C., Sickert,D., Stahler,C., Meese,E. *et al.* (2015) Influence of next-generation sequencing and storage conditions on miRNA patterns generated from PAXgene blood. *Anal. Chem.*, **87**, 8910–8916.

11. Backes,C., Sedaghat-Hamedani,F., Frese,K., Hart,M., Ludwig,N., Meder,B., Meese,E. and Keller,A. (2016) Bias in high-throughput analysis of miRNAs and implications for biomarker studies. *Anal. Chem.*, **88**, 2088–2095.
12. Meder,B., Backes,C., Haas,J., Leidinger,P., Stahler,C., Grossmann,T., Vogel,B., Frese,K., Giannitsis,E., Katus,H.A. *et al.* (2014) Influence of the confounding factors age and sex on microRNA profiles from peripheral blood. *Clin. Chem.*, **60**, 1200–1208.
13. Nagy,C., Maheu,M., Lopez,J.P., Vaillancourt,K., Cruceanu,C., Gross,J.A., Arnovitz,M., Mechawar,N. and Turecki,G. (2015) Effects of postmortem interval on biomolecule integrity in the brain. *J. Neuropathol. Exp. Neurol.*, **74**, 459–469.
14. Lv,Y.H., Ma,K.J., Zhang,H., He,M., Zhang,P., Shen,Y.W., Jiang,N., Ma,D. and Chen,L. (2014) A time course study demonstrating mRNA, microRNA, 18S rRNA, and U6 snRNA changes to estimate PMI in deceased rat's spleen. *J. Forensic Sci.*, **59**, 1286–1294.
15. Ibberson,D., Benes,V., Muckenthaler,M.U. and Castoldi,M. (2009) RNA degradation compromises the reliability of microRNA expression profiling. *BMC Biotechnol.*, **9**, 102.
16. Cloonan,N., Wani,S., Xu,Q., Gu,J., Lea,K., Heater,S., Barbacioru,C., Steptoe,A.L., Martin,H.C., Nourbakhsh,E. *et al.* (2011) MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol.*, **12**, R126.
17. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
18. Swierniak,M., Wojcicka,A., Czetwertynska,M., Stachlewska,E., Maciag,M., Wiechno,W., Gornicka,B., Bogdanska,M., Koperski,L., de la Chapelle,A. *et al.* (2013) In-depth characterization of the microRNA transcriptome in normal thyroid and papillary thyroid carcinoma. *J. Clin. Endocrinol. Metab.*, **98**, E1401–E1409.
19. Callis,T.E., Chen,J.F. and Wang,D.Z. (2007) MicroRNAs in skeletal and cardiac muscle development. *DNA Cell Biol.*, **26**, 219–225.
20. Thum,T., Catalucci,D. and Bauersachs,J. (2008) MicroRNAs: novel regulators in cardiac development and disease. *Cardiovasc. Res.*, **79**, 562–570.
21. Liu,S., Tetzlaff,M.T., Liu,A., Liegl-Atzwanger,B., Guo,J. and Xu,X. (2012) Loss of microRNA-205 expression is associated with melanoma progression. *Lab. Invest.*, **92**, 1084–1096.
22. Karaayvaz,M., Pal,T., Song,B., Zhang,C., Georgakopoulos,P., Mehmood,S., Burke,S., Shroyer,K. and Ju,J. (2011) Prognostic significance of miR-215 in colon cancer. *Clin. Colorectal Cancer*, **10**, 340–347.
23. Akamatsu,S., Hayes,C.N., Tsuge,M., Miki,D., Akiyama,R., Abe,H., Ochi,H., Hiraga,N., Imamura,M., Takahashi,S. *et al.* (2015) Differences in serum microRNA profiles in hepatitis B and C virus infection. *J. Infect.*, **70**, 273–287.
24. Krauskopf,J., Caiment,F., Claessen,S.M., Johnson,K.J., Warner,R.L., Schomaker,S.J., Burt,D.A., Aubrecht,J. and Kleinjans,J.C. (2015) Application of high-throughput sequencing to circulating microRNAs reveals novel biomarkers for drug-induced liver injury. *Toxicol. Sci.*, **143**, 268–276.
25. Pirola,C.J., Fernandez Gianotti,T., Castano,G.O., Mallardi,P., San Martino,J., Mora Gonzalez Lopez Ledesma,M., Flichman,D., Mirshahi,F., Sanyal,A.J. and Sookoian,S. (2015) Circulating microRNA signature in non-alcoholic fatty liver disease: from serum non-coding RNAs to liver histology and disease pathogenesis. *Gut*, **64**, 800–812.
26. Xu,J., Wu,C., Che,X., Wang,L., Yu,D., Zhang,T., Huang,L., Li,H., Tan,W., Wang,C. *et al.* (2011) Circulating microRNAs, miR-21, miR-122, and miR-223, in patients with hepatocellular carcinoma or chronic hepatitis. *Mol. Carcinog.*, **50**, 136–142.
27. Akat,K.M., Moore-McGriff,D., Morozov,P., Brown,M., Gogakos,T., Correa Da Rosa,J., Mihailovic,A., Sauer,M., Ji,R., Ramarathnam,A. *et al.* (2014) Comparative RNA-sequencing analysis of myocardial and circulating small RNAs in human heart failure and their utility as biomarkers. *Proc. Natl Acad. Sci. USA*, **111**, 11151–11156.
28. Gomes,C.P., Oliveira-Jr,G.P., Madrid,B., Almeida,J.A., Franco,O.L. and Pereira,R.W. (2014) Circulating miR-1, miR-133a, and miR-206 levels are increased after a half-marathon run. *Biomarkers*, **19**, 585–589.
29. Cacchiarelli,D., Legnini,I., Martone,J., Cazzella,V., D'Amico,A., Bertini,E. and Bozzoni,I. (2011) miRNAs as serum biomarkers for Duchenne muscular dystrophy. *EMBO Mol. Med.*, **3**, 258–265.

Fehlmann et al. Clinical Epigenetics (2016) 8:123
DOI 10.1186/s13148-016-0287-1

Clinical Epigenetics

RESEARCH Open Access

cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs

Tobias Fehlmann[1], Stefanie Reinheimer[3], Chunyu Geng[2*], Xiaoshan Su[2], Snezana Drmanac[2,4], Andrei Alexeev[2,4], Chunyan Zhang[2], Christina Backes[1], Nicole Ludwig[3], Martin Hart[3], Dan An[2], Zhenzhen Zhu[2], Chongjun Xu[2,4], Ao Chen[2], Ming Ni[2], Jian Liu[2], Yuxiang Li[2], Matthew Poulter[2], Yongping Li[2], Cord Stähler[1], Radoje Drmanac[2,4], Xun Xu[2*], Eckart Meese[3] and Andreas Keller[1*]

Abstract

Background: We present the first sequencing data using the combinatorial probe-anchor synthesis (cPAS)-based *BGISEQ-500* sequencer. Applying cPAS, we investigated the repertoire of human small non-coding RNAs and compared it to other techniques.

Results: Starting with repeated measurements of different specimens including solid tissues (brain and heart) and blood, we generated a median of 30.1 million reads per sample. 24.1 million mapped to the human genome and 23.3 million to the *miRBase*. Among six technical replicates of brain samples, we observed a median correlation of 0.98. Comparing BGISEQ-500 to HiSeq, we calculated a correlation of 0.75. The comparability to microarrays was similar for both BGISEQ-500 and HiSeq with the first one showing a correlation of 0.58 and the latter one correlation of 0.6. As for a potential bias in the detected expression distribution in blood cells, 98.6% of HiSeq reads versus 93.1% of BGISEQ-500 reads match to the 10 miRNAs with highest read count. After using miRDeep2 and employing stringent selection criteria for predicting new miRNAs, we detected 74 high-likely candidates in the cPAS sequencing reads prevalent in solid tissues and 36 candidates prevalent in blood.

Conclusions: While there is apparently no ideal platform for all challenges of miRNome analyses, cPAS shows high technical reproducibility and supplements the hitherto available platforms.

Keywords: Next-generation sequencing, miRNA, Biomarker discovery, BGISEQ

Background

Currently, high-throughput analytical techniques are massively applied to further the understanding of the non-coding transcriptome [1]. Still, the full complexity of non-coding RNAs is only partially understood. One class of well-studied non-coding RNAs comprises small oligonucleotides, so-called miRNAs [2, 3].

Among the techniques most commonly used for miRNA profiling are microarrays, RT-qPCR, and next-generation sequencing (NGS), also referred to as high-throughput sequencing (HTS). An excellent review on the different platforms and a cross-platform comparison has been recently published [4]. A detailed examination of technologies, however, frequently reveals a bias. One reason for the respective bias is the ligation step, as, e.g., reported by Hafner and co-workers [5]. For example, the quantification of miRNAs differs between NGS and microarrays as it is dependent on base composition [6]. Especially, the guanine and uracil content of a miRNA seems to influence the abundance depending on the platform used. A substantial strength of NGS is the ability to support the completion of the non-coding transcriptome. Unlike microarrays and RT-qPCR, NGS allows the discovery of novel miRNA candidates. To this end, different algorithms have been implemented, with *miRDeep* being one of the most popular ones [7]. A substantial part of small RNA sequencing data has been obtained using HiSeq and MiSeq platforms (Illumina) based on stepwise sequencing by polymerase on DNA microarrays prepared by bridge PCR [8], as well as the

* Correspondence: gengchunyu@genomics.cn; xuxun@genomics.cn; andreas.keller@ccb.uni-saarland.de
[2]BGI-Shenzhen, Shenzhen, China
[1]Clinical Bioinformatics, Saarland University, 66125 Saarbrücken, Germany
Full list of author information is available at the end of the article

IonTorrent systems from Thermo Fisher Scientific using a different type of polymerase-based stepwise sequencing on micro-bead arrays generated by emulsion PCR, the first method proposed for making microarrays for massively parallel sequencing [9]. Another approach is the ligase-based stepwise sequencing also using micro-bead arrays, applied for example by ThermoFisher Scientific's SOLiD sequencing platform, and which has also been used to analyze and present novel miRNAs [10].

In the current study, we applied the new combinatorial probe-anchor synthesis (cPAS)-based BGISEQ-500 sequencing platform that combines DNA nanoball (DNB) nanoarrays [11] with stepwise sequencing using polymerase. An important advantage of this technique compared to the previously mentioned sequencing systems is in that no PCR is applied in preparing sequencing arrays. Applying cPAS, we investigated the human non-coding transcriptome. We first evaluated the reproducibility of sequencing on standardized brain and heart samples, then compared the performance to Agilent's microarray technique and finally evaluated blood samples. Using the web-based miRNA analysis pipeline *miRmaster* and the tool *novoMiRank* [12], we finally predicted 135 new high-likely miRNA candidates specific for tissue and 35 new miRNA candidates specific for blood samples.

Methods

Samples

In this study, we examined the performance of three sample types using three techniques for high-throughput miRNA measurements (Illumina's HiSeq sequencer, Agilent's miRBase microarrays, and BGI's BGISEQ-500 sequencing system, see details below). The three specimens were standardized HBRR sample ordered from Ambion (catalog number AM6051) and UHRR sample ordered from Agilent (catalog number 740000). UHRR and HBRR samples were measured in two and six replicates, respectively. As third sample type, we used *PAXGene* blood tubes. Here, two healthy volunteers' blood samples were collected and miRNAs were extracted using PAXgene Blood RNA Kit (Qiagen) according to manufacturer's protocol. The study has been approved by the local ethics committee.

Next-generation sequencing using BGISEQ-500

We prepared the libraries starting with 1 μg total RNA for each sample. Firstly, we isolated the microRNAs (miRNA) by 15% urea-PAGE gel electrophoresis and cut the gel from 18 to 30 nt, which corresponds to mature miRNAs and other regulatory small RNA molecules. After gel purification, we ligated the adenylated 3′ adapter to the miRNA fragment. Secondly, we used the RT primer with barcode to anneal the 3′ adenylated adapter in order to combine the redundant unligated 3′ adenylated adapter. Then, we ligated the 5′ adapter and did reverse transcript (RT) reaction. After cDNA first strand synthesis, we amplified the product by 15 cycles. We then carried out the second size selection operation and selected 103–115 bp fragments from the gel. This step was conducted in order to purify the PCR product and remove any nonspecific products. After gel purification, we quantified the PCR yield by Qubit (Invitrogen, Cat No. Q33216) and pooled samples together to make a single strand DNA circle (ssDNA circle), which gave the final miRNA library.

DNA nanoballs (DNBs) were generated with the ssDNA circle by rolling circle replication (RCR) to enlarge the fluorescent signals at the sequencing process as previously described [11]. The DNBs were loaded into the patterned nanoarrays and single-end read of 50 bp were read through on the BGISEQ-500 platform for the following data analysis study. For this step, the BGISEQ-500 platform combines the DNA nanoball-based nanoarrays [11] and stepwise sequencing using polymerase, as previously published [13–15]. The new modified sequencing approach provides several advantages, including among others high throughput and quality of patterned DNB nanoarrays prepared by linear DNA amplification (RCR) instead of random arrays by exponential amplification (PCR) as, e.g., used by Illumina's HiSeq and longer reads of polymerase-based cycle sequencing compared to the previously described combinatorial probe-anchor ligation (cPAL) chemistry on DNB nanorrays [11]. The usage of linear DNA amplification instead of exponential DNA amplification to make sequencing arrays results in lower error accumulation and sequencing bias.

Next-generation sequencing using HiSeq

Samples have been sequenced using Illumina HiSeq sequencing according to manufacturer's instructions and as previously described [16, 17].

Agilent microarray measurements

For detection of known miRNAs, we used the SurePrint G3 8×60k miRNA microarray (miRBase version 21, Agilent Technologies) containing probes for all miRNAs from miRBase version 21 in conjunction with the miRNA Complete Labeling and Hyb Kit (Cat. No. 5190-0456) according to the manufacturer's recommendations. In brief, 100 ng total RNA including miRNAs was dephosphorylated with calf intestine phosphatase. After denaturation, Cy3-pCp was ligated to all RNA fragments. Labeled RNA was then hybridized to an individual 8×60k miRNA microarray. After washing, array slides were scanned using the Agilent Microarray Scanner G2565BA with 3-μm resolution in double-pass mode. Signals were retrieved using Agilent AGW Feature Extraction software (version 10.10.11).

Data availability

The new sequencing data using BGISEQ-500 data are available in the Additional file of this manuscript (Additional file 1: Table S3).

Bioinformatics analysis

The raw reads were collapsed and used as input for the web-based tool miRMaster, allowing for integrated analysis of NGS miRNA data. On the server side, mapping to the human genome was carried out using *Bowtie* [18] (one mismatch allowed). miRNAs were quantified similar to the popular *miRDeep2* [19] algorithm. The prediction of novel miRNAs was performed using an extended feature set built up on novoMiRank [12]. For classification, an *AdaBoost* model using decision trees was applied. Novel miRNAs were cross-checked against other RNA resources, including the *miRBase* [20], *NONCODE2016* [21], and *Ensembl* non-coding RNAs. The assessment of the quality of new miRNAs was carried out using the novoMiRank algorithm. A downstream analysis of results including cluster analysis was performed using R. For target prediction, we applied TargetScan 7.1 (http://www.targetscan.org/vert_71/) and predicted for all new miRNAs the targets. With the predictions, we extracted the context ++ scores and used them for prioritizing the targets, miRNA-target interactions with context++ scores below 1 were considered as high-likelihood targets. Target networks were constructed using an offline version of MiRTargetLink [22] and visualized in Cytoscape. miRNA target pathway analysis has been carried out using GeneTrai2 [23]. For the GeneTrail2 analysis, all available categories were analyzed, the minimal category size was set to 4 and all p values were adjusted using Benjamini-Hochberg adjustment.

Results

Raw data analysis

We sequenced six brain, two heart, and two blood samples using the BGISEQ-500 system. The resulting reads were mapped to the human genome allowing one mismatch per read. The 10 samples had a median of 30.1 million reads. Of these, 24.1 million reads mapped to the human genome and 23.3 million reads to miRNAs annotated in the human miRBase version 21. The remaining 0.7 million reads per sample contain potentially new miRNAs.

Technical reproducibility of the BGISEQ-500 and comparison to microarrays

To assess the technical reproducibility of the sequencing platform, we evaluated the six technical replicates of the human brain sample (see correlation matrix in Fig. 1). The median correlation between the six replicates was 0.98, and the 25 and 75% quantile were 0.98 and 0.99, respectively. These data suggest an overall high correlation for technical replicates on the BGISEQ-500 platform.

Comparing the BGISEQ-500 data to the measurements of the brain sample with microarrays (miRBase version 21) that have also been carried out as six technical replicates (median correlation of the microarrays was 0.999), we observed a log correlation of 0.48. A direct comparison is presented in the scatter plot in Fig. 2a. This plot highlights many miRNAs that can be measured at a comparable level on both platforms. However, a subset of the small non-coding RNAs is shifted towards higher expression on the array platform. The same behavior can be observed in the cluster heat map in Fig. 2b. This heat map graphically represents the 50 miRNAs with most different detection between both techniques. To compare rather the ranks of miRNAs instead of the absolute read counts, the replicated brain samples on both platforms were jointly quantile normalized. Three miRNAs, in particular, showed highly significant deviations (multiple testing adjusted p values below 10^{-20}). Hsa-miR-8069 was almost not detected in the BGISEQ-500 but had 0.9 million normalized intensity counts on the array platform, hsa-miR-4454 had 51.6 normalized reads on the BGISEQ-500 versus 1.9 million normalized counts on the microarrays, and hsa-miR-7977 had 343.2 normalized reads on the BGISEQ-500 versus 1.3 million normalized counts on the microarrays. This means that the three miRNAs were orders of magnitudes more abundant on microarrays as compared to the sequencing system. The secondary structures of the three precursors are presented in Additional file 2: Figure S1. These results match well to previously published platform comparisons between NGS and microarrays [6]. Here, several miRNAs such as hsa-miR-941 (not detected in any array experiment, not detected in RT-qPCR, average read count of ~1000 reads using Illumina HiSeq sequencing) had expression levels differing several orders of magnitude between the miRBase microarrays and using HiSeq sequencing.

The full list of miRNAs with raw and adjusted p values in t test and Wilcoxon-Mann-Whitney test comparing BGISEQ-500 and microarrays is presented in Additional file 3: Table S1. Overall, the results are well in-line with those obtained between HiSeq NGS and the same microarray platform [6]. Reasons that explain differences between arrays and NGS include different sensitivity levels of the platforms, cross-hybridization of miRNAs with similar sequences on the microarrays or bias in library preparation. Further, effects of the normalization can lead to variations in miRNA quantification.

Biological replicates of blood samples and comparison to other platforms

One of the most promising applications in small RNA analysis is biomarker profiling in body fluids. We

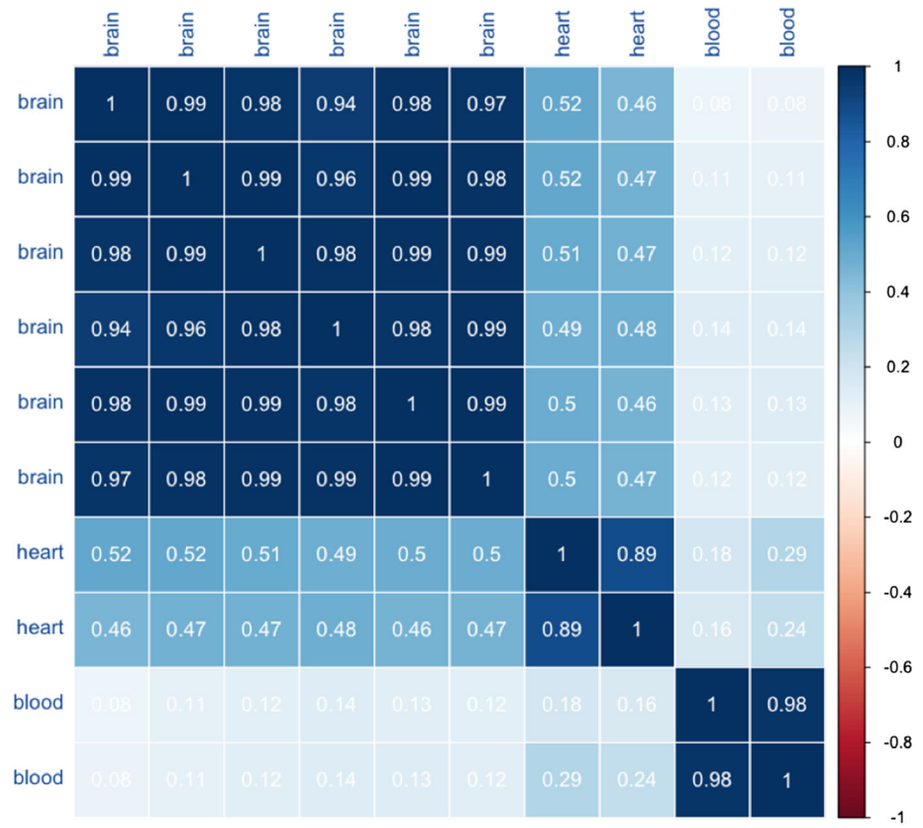


Fig. 1 Correlation matrix of the brain (six technical replicates), heart (two technical replicates), and blood (two biological replicates) sequenced by the BGISEQ-500 system

previously analyzed over 2000 blood samples on Agilent microarrays [17, 24, 25] and about 1000 samples using HiSeq sequencing [26, 27] and compared both platforms [6]. We correlated two newly sequenced blood samples using the BGISEQ-500 system to the data generated by HiSeq and Agilent microarrays. When interpreting the results, it is important to keep in mind that the microarrays and HiSeq data are from the same samples [6] while the newly sequenced blood drawings are from other individuals and thus biological but no technical replicates. To minimize a potential bias between the platforms with respect to different miRNA sets, we first reduced the marker set to the 2525 human miRNAs that were profiled on all platforms and next to the subset of 658 miRNAs that were discovered in all three platforms. For each, platform data were normalized using quantile normalization. Due to the wide dynamic range of miRNAs in blood samples, which is approximately 10^7, we present the three pairwise comparisons (BGISEQ-500 to microarrays, BGISEQ-500 to HiSeq, and HiSeq to microarrays) on a log scale. The scatter plots are presented in Fig. 3. The highest correlation was observed for BGISEQ-500 to Illumina (0.75, Fig. 3a). Even the correlation between microarrays and HiSeq was below this value (0.6, Fig. 3c). Especially since technical replicates have been measured for these platforms, the increased correlation of sequencing platforms is remarkable. The comparison of BGISEQ-500 and microarrays revealed correlation values in the same range as for the brain samples (0.58, Fig. 3b). The 3D scatter plot in Fig. 3d compares the expression of the three platforms directly to each other. The coloring of the miRNAs has been carried out with respect to the GC content.

Expression distribution of miRNAs

As mentioned, miRNA expression is highly variable and can scatter across many orders of magnitude. We thus compared the distribution of the sequencing reads in blood samples on the HiSeq to the BGISEQ-500. Blood samples, including blood cells (especially red blood cells) are known to be enriched for few miRNAs that are highly expressed. The diagram in Fig. 4 (panel A) highlights that 90.8% of all blood sequencing reads from the HiSeq match to one single miRNA: hsa-miR-486-5p. The second most abundant miRNA miR-92a-3p takes further 5.5%, and already the third most abundant marker miR-451a has below 1% of all reads. In sum, 98.6% of all reads match to the top 10 miRNAs. For the

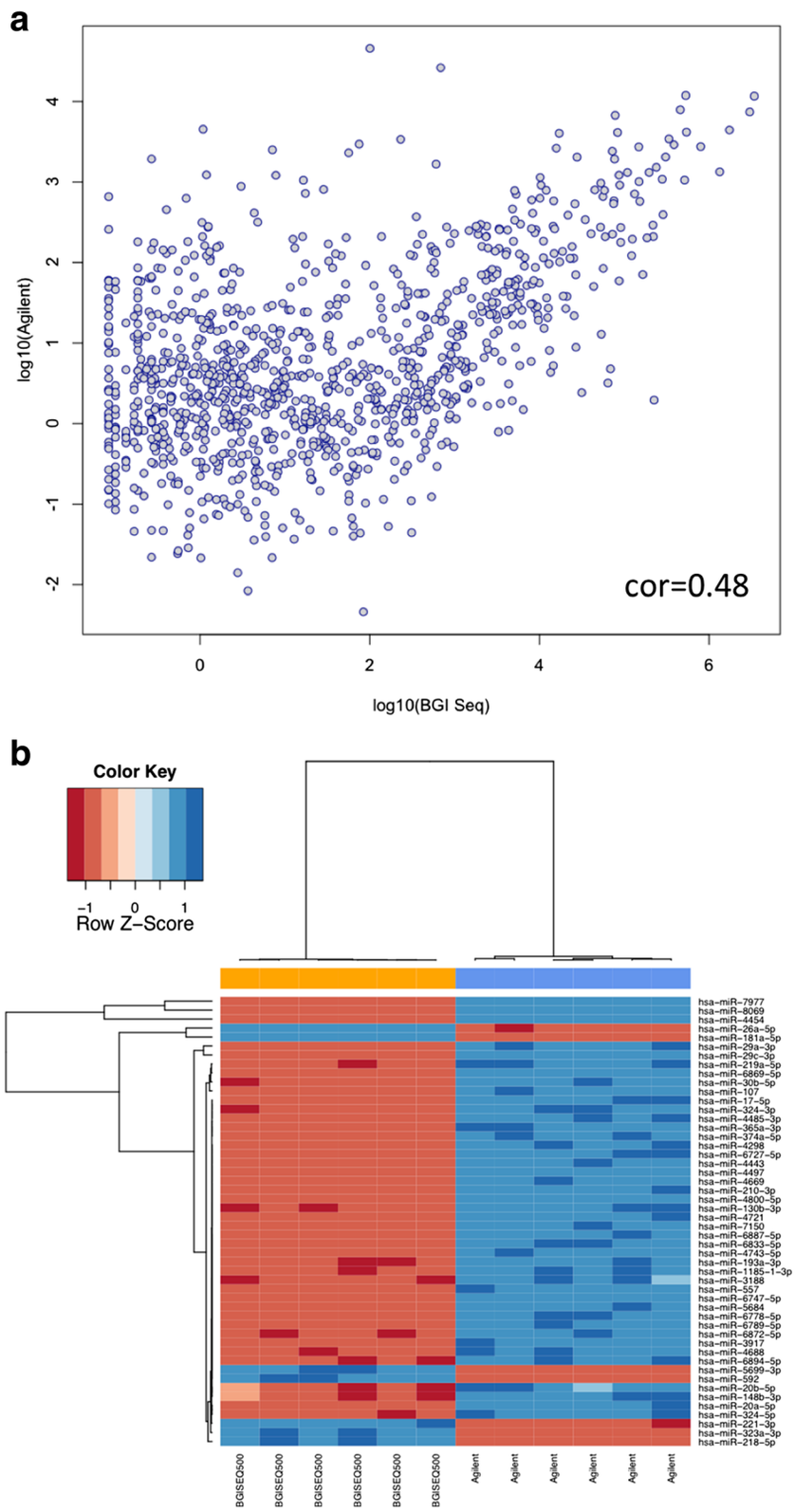


Fig. 2 a Log average expression of common miRNAs for the brain RNA on BGISEQ-500 and on Agilent microarrays (six technical replicates each). **b** Heat map with dendrogram for the 50 most differently detected miRNAs in the brain RNA between Agilent and BGISEQ-500 (six technical replicates each)

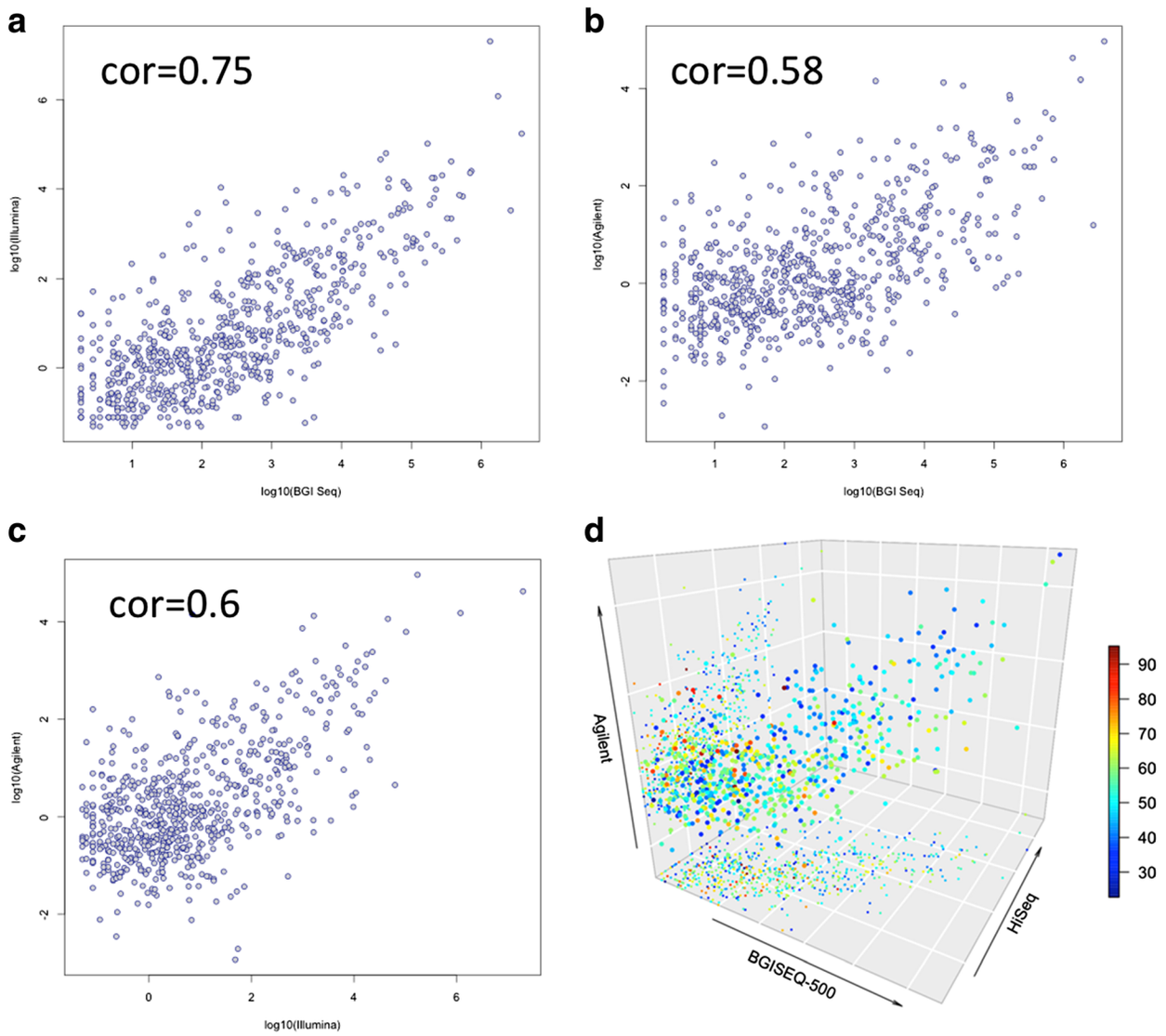


Fig. 3 **a**-**c** Pairwise scatter plots for comparing expression of miRNAs in blood cells on microarrays, HiSeq, and BGISEQ-500. Please note that for HighSeq and Agilent technical replicates were measured, for BGISEQ-500 biological replicates. **d** 3D scatter-plot colored by the GC content of miRNAs

BGISEQ-500 (panel B), 45.9% of reads match to miR-451a, further 20% map to miR-191-5p and 13.3% map to miR-92a-3p. The most abundant miRNA in HiSeq, miR-486-5p, is detected in 7.7% of all reads. 93.1% of all sequenced reads match to the top 10 miRNAs.

Comparison of the distribution and abundance of miRNAs on the microarray platform is difficult since microarrays show a saturation effect. This means that for two miRNAs expressed in a range above the saturation, no difference can be observed. We nonetheless performed the same analysis as presented above, assuming that the sum of all expression counts equals to 100%. In this analysis, miR-451a which is found in 0.8% of HiSeq reads and 45.9% of BGISEQ-500 reads is the highest expressed in microarrays (37.2% of all expression counts), followed by 17% of miR-486-5p.

Prediction of novel miRNAs

Predicting new miRNAs from NGS data is a challenging task since many false positive miRNA candidates are observed. We implemented our own prediction tool for miRNAs from NGS data and filtered the candidates stringently to reduce the false discovery rate. Without any filtering steps, our initial predictor trimmed for maximizing the ROC AUC returned 25,086 candidates across all samples. The exclusion of the candidates with low abundance (less than 10 total reads) reduced the number of candidates to around 10% (2354 candidates).

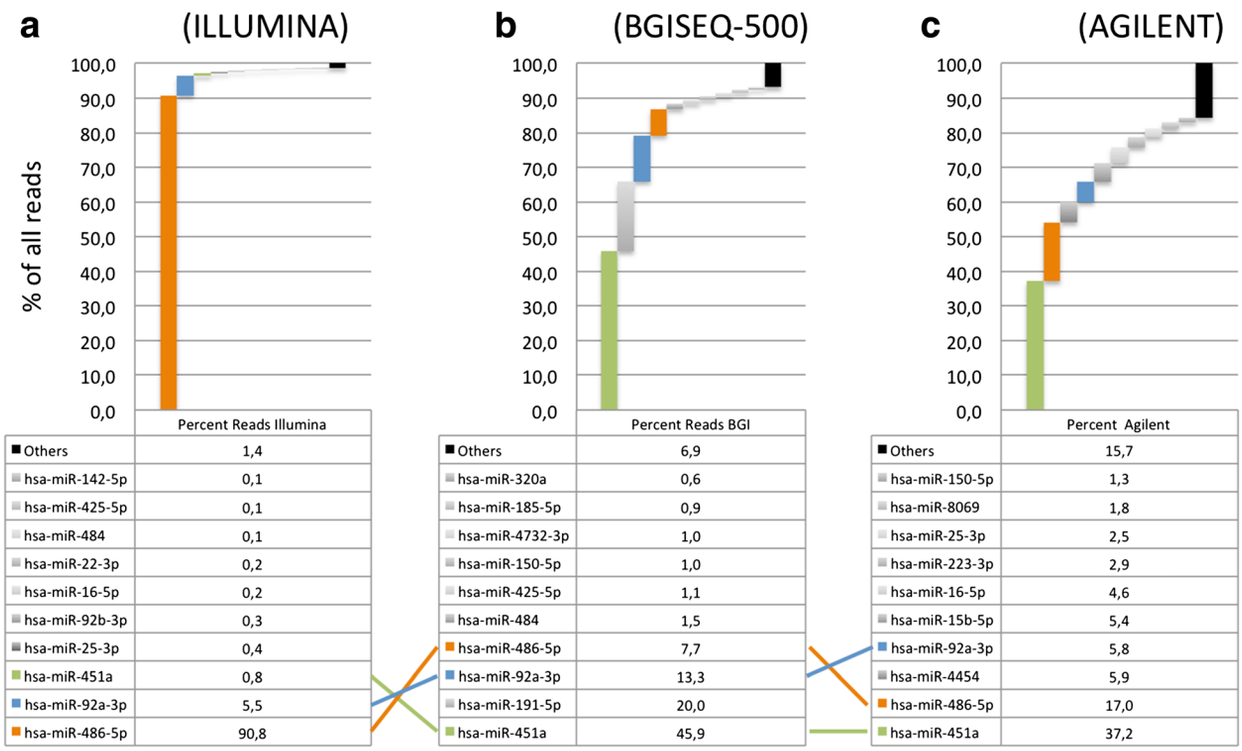


Fig. 4 Expression distribution of the 10 miRNAs with the highest detection in the blood RNA on the HiSeq system (**a**), BGISEQ-500 (**b**), and microarray system (**c**). Note that for the Agilent microarray system, the sum of all expression intensities was assumed to be 100%

Further analysis with *novoMiRank* (cutoff 1.5) filtered out more miRNAs, leaving 1553. The miRNAs were flagged by *novoMiRank* because of a high deviation from miRNAs in the first *miRBase* versions, including deviating length, free energy, or nucleic acid composition of miRNAs. Matching the remaining candidates to other RNA resource in a blacklisting step finally presented 926 miRNA candidates (Additional file 4: Table S2). Still, it is likely that this set contains many false positives. Additionally, low-throughput experimental validation of almost 1000 miRNA candidates, e.g., by Northern Blot is a very labor-extensive approach. We thus additionally compared the frequency of reads mapping to the blood versus tissue samples. As detailed in Fig. 5a, we observe a substantial variability between blood and tissue for the 926 miRNA candidates (correlation 0.18). Defining a miRNA as tissue/blood specific if it occurs with a factor of 100-fold higher in one of both sample types (normalized for the total number of samples) highlighted 74 new miRNA candidates specific for tissue and 36 new miRNA candidates specific for blood samples. Figure 5b shows bar plots for two miRNA precursors, the most tissue specific novel-mir-36616 (blue), only present in the brain samples, and the blood specific novel-mir-31007. The first miRNA, which is observed exclusively in the brain samples and not in the heart, reveals a significantly less expressed 3′ mature form as compared to the 5′ mature form. The second miRNA is exclusively observed in blood samples. Here, the 5′ mature form is lower expressed compared to the 3′ form. The boxes above the bar plots show the secondary structures of both miRNA candidates.

miRNA target analysis

For all 926 miRNAs, we predicted targets using TargetScan. To rank miRNA-target interactions, we used the context++ score (distribution of the context++ score across all predictions is provided in Additional file 5: Figure S2). Thereby, we observed an accumulation of high-likelihood targets for tissue-specific miRNAs. Of the 926 miRNAs, the tissue specific had an average 42.8 targets, the neither for blood nor for tissue-specific miRNAs 40.7 targets while for blood-specific miRNAs, only 34.5 targets were predicted. The complex miRNA-target network is presented in Additional file 6: Figure S3. It contains 6014 nodes (5088 genes and 926 miRNAs). Network characteristics such as degree distribution and shortest path length are presented in Additional file 7: Figure S4. The genes with largest numbers of predicted miRNAs targeting the gene were CYB561D1 (229 miRNAs), FBXL12 (174 miRNAs), PML (162 miRNAs), and VNN3 (154 miRNAs). The distribution of miRNAs in

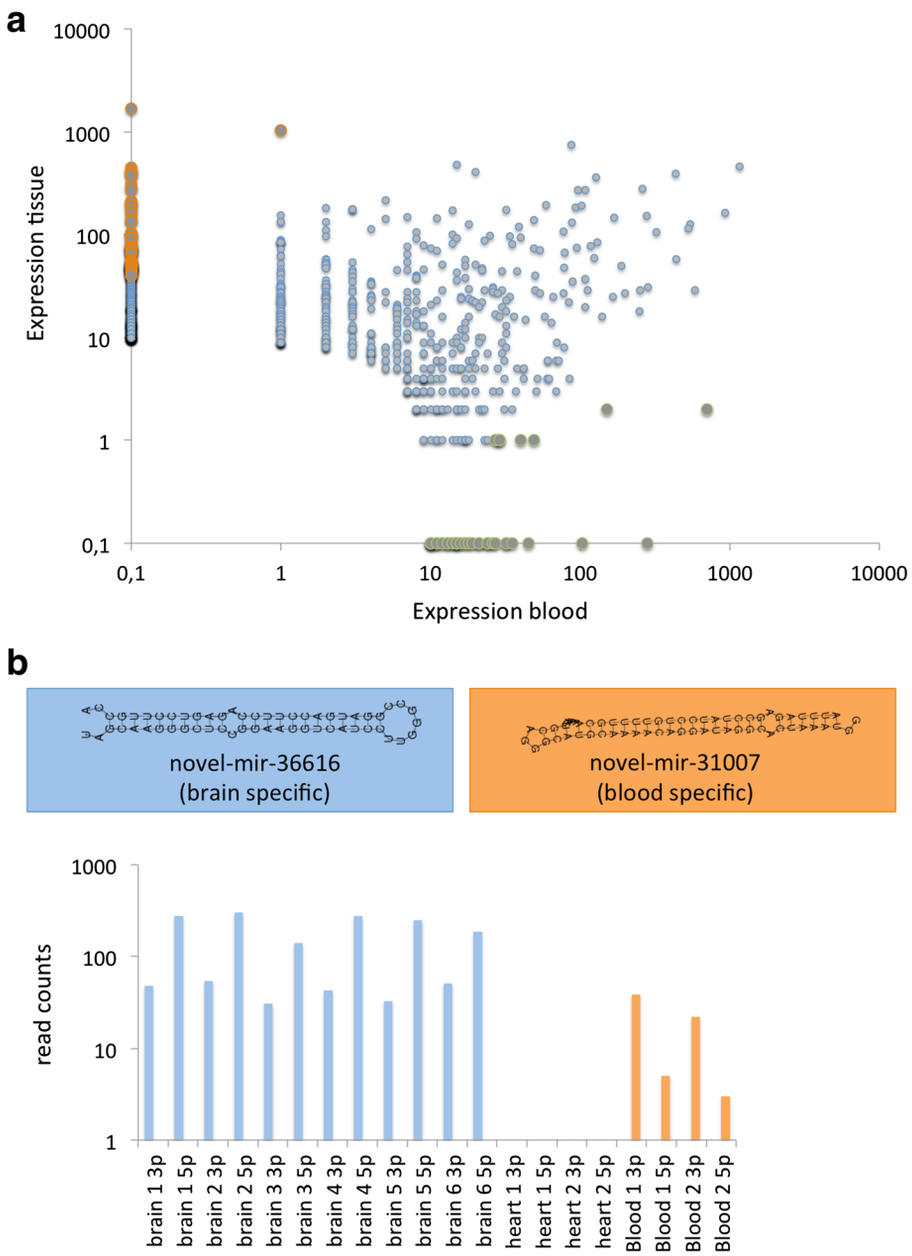


Fig. 5 a Expression of novel miRNAs in blood versus tissue. The *green* miRNAs are specific for blood, the *orange* miRNAs for tissue, and the *blue* miRNAs were detected in both specimens. **b** Bar plot for two different miRNAs. The first miRNA novel-mir-36616 in *blue* is detected only in the brain tissue and not in the heart tissue or blood, the -3p form is one order of magnitude less expressed compared to the -5p mature form. The second miRNA novel-mir-31007 is expressed in blood and not in tissue, here, the -3p form is more than an order of magnitude more abundant compared to the -5p form

the different group is presented as Venn diagram in Additional file 8: Figure S5). Among the predicted target genes that were found only for candidate miRNAs being blood specific was, e.g., HMOX1, heme oxygenase 1, mediating the first step of the heme catabolism by cleaving heme to build biliverdin or HPX, coding for hemopexin. The complex nature of the in silico calculated miRNA-target network requires further analyses to understand whether target genes accumulate in specific biochemical categories such as KEGG pathways or gene ontologies. We thus applied GeneTrail2 separately to the set of genes targeted by blood specific miRNAs, targeted by tissue specific miRNAs and by all other miRNAs. As the background sets, all genes predicted to be targeted by at least a single miRNA were selected and the functionality to compare different enrichment analyses by

GeneTrail2 has been used. Enriched pathways seem to be largely relevant for either blood or tissue miRNAs, as Additional file 9: Figure S6 highlights. Tissue specific miRNAs had target genes enriched for DNA damage response, the apoptosis, or RNA polymerase II regulatory region DNA binding while blood miRNAs target genes were, e.g., enriched for TP35 network. Interestingly, tissue miRNA target genes also clustered on specific genomic locations (e.g., 19p12 and 19.q13) while blood miRNA targets did not show such an enrichment. In contrast, blood miRNA targets were enriched for disease phenotypes such as carotid artery diseases. In sum, the enrichment analysis highlights very distinct patterns for blood and tissue miRNA targets. Of course, not only the new miRNAs themselves but also the predicted targets deserve detailed experimental validation.

Discussion

The advent of next-generation sequencing reduced the costs of sequencing while simultaneously increasing the speed of throughput [28]. Today, the costs for small RNA seq are almost equal to and even lower than miRNA microarrays, although small RNA-seq provides the additional possibility for detecting novel small RNA entities.

In the present study, we investigated two current sequencing approaches supporting massively parallel sequencing, which is of high relevance in small RNA research because of the high dynamic range of these molecules: DNA nanoball [11]-based sequencing by BGISEQ-500 and PCR cluster [8]-based sequencing by HiSeq. An important difference between these techniques is in that the first approach uses linear DNA amplification, and the second uses exponential DNA amplification to make sequencing arrays. The latter approach may in turn lead to amplification errors and some specific biases. Besides this fundamental difference, both approaches have their additional advantages and disadvantages. Specifically for the BGISEQ-500, the library preparation currently takes around three working days, the sequencing itself needs one or at maximum two working days. Each flowcell of the BGISEQ-500 has two lanes. On each of these lanes, 32 Gb data can be generated using single-end reads of length 50 bases. The cost of the reagent and material is around 200 USD for 20 million reads ensuring high-quality data at a reasonable cost.

Recently, we published a manuscript about bias in NGS and microarray analysis for miRNAs [6], highlighting that the expression of miRNAs on different platforms varies by, for example, the nucleic acid composition. In the validation by RT-qPCR, we focused on miRNAs discordant between the high-throughput platforms. Thereby, we observed cases where the RT-qPCR results were concordant with Illumina HiSeq, with microarrays or with none of the techniques. Therefore, we were especially interested how the BGISEQ-500 platform compares to the HiSeq platform and microarrays with the content from the *miRBase* for small RNA analysis.

Three miRNAs had high divergence between arrays and BGISEQ-500, among them hsa-miR-4454, which was high abundant in arrays but almost not detectable in BGISEQ-500. According to the miRBase, only 28% of users believe that this miRNA is real. Although such votes have only limited value, they at least indicate that this miRNA may be influenced by technological bias.

For high-throughput sequencing, the library preparation and the kits used play a crucial role for the quality of the sequencing results. Others and we noticed an overly abundance of the miRNA miR-486-5p when using the TruSeq kit (Illumina, San Diego), which seems to be independent of the source of the analyzed material [6, 29, 30]. Using the BGISEQ-500 platform, we observed lower read counts for this miRNA. However, in some cases, the miRNA abundance of BGISEQ-500 matches to the HiSeq sequencing results while microarrays show a different expression level, and in other cases, the BGISEQ-500 deviates from the other platforms and in several cases, all three techniques provide substantially divergent results. The more even distribution of reads of the BGISEQ-500 compared to the HiSeq results facilitates the discovery of new miRNAs, which are expected to be significantly less expressed as compared to the already known miRNAs, especially from early miRBase versions.

With respect to many miRNA currently annotated in miRBase and the rapidly growing number of new miRNAs, it is essential not only to have tools for filtering likely false-positives such as the NovoMiRank tool but also to carry out validation of miRNAs using other molecular biology approaches such as cloning and Northern blotting.

Focusing on the performance of the BGISEQ-500, we found a high technical reproducibility of sequencing results, which was however slightly below the technical reproducibility of microarrays. This fact can have different reasons, e.g., the different limit of detection of microarrays. In contrast to sequencing, microarrays have a saturation effect. With respect to the total number of discovered known miRNAs, performance of the BGISEQ-500 was comparable both to the Illumina and the microarray platform.

Conclusions

In sum, none of the mentioned platforms seems to provide the “ultimate solution” in miRNA analysis. All have their advantages and disadvantages and show some bias for the detection of certain sequence types.

Additional files

Additional file 1: Table S3. miRNA read count of the BGISEQ-500. (XLSX 250 kb)

Additional file 2: Figure S1. Predicted secondary structures for selected miRNAs. (PNG 241 kb)

Additional file 3: Table S1. Comparison of BGISEQ-500 to Agilent. (XLSX 135 kb)

Additional file 4: Table S2. List of novel miRNA candidates. (XLSX 6531 kb)

Additional file 5: Figure S2. Histogram of the decade logarithm of the context++ scores (multiplied by −1) of predicted targets for the candidate miRNAs. Since negative context++ scores are favorable, the miRNA targets on the right of the diagram are more likely true interactions. (PNG 78 kb)

Additional file 6: Figure S3. Full interaction network. Predicted miRNAs are represented in large nodes, colored by type (red: blood specific, blue: tissue specific, green: all others) and genes are represented by smaller gray nodes. (PNG 1033 kb)

Additional file 7: Figure S4. Core network characteristics as node degree distribution (*top*) and shortest path length (*bottom*). (PNG 129 kb)

Additional file 8: Figure S5. Venn diagram showing the distribution of predicted target genes for tissue-specific miRNA candidates, blood-specific miRNA candidates, and all other miRNA candidates. (PNG 156 kb)

Additional file 9: Figure S6. Comparison of the pathway enrichment analysis for the GeneTrail2 analysis with respect to the three target sets. *Red arrows* represent significant enrichments. (PNG 289 kb)

Acknowledgements
We acknowledge the support of BGI-Shenzhen and Complete Genomics.

Funding
The study has been funded by internal funds of Saarland University.

Availability of data and materials
Following publication expression data are available in the gene expression omnibus (GEO).

Authors' contributions
Setting up the assay were done by CG, XS, AA, SD, CZ, DA, JL, and RD. Generating miRNA data were done by SR, CZ, NL, MH, ZZ, CX, AC, and MN. Evaluation of data was done by TF, CB, NL, YL, and AK. Drafting and revision of the manuscript were done by EM, AK. Study design and set-up were done by YL, CS, XX, EM, and AK. All authors read and approved the final manuscript.

Competing interests
Authors with affiliations 1 and 2 are employed by BGI-Shenzhen, Shenzhen, China, and Complete Genomics (a BGI company), Mountain View, CA, USA.

Consent for publication
Not applicable.

Ethics approval and consent to participate
The study has been approved by the local ethics committee (Ärztekammer des Saarlandes).

Author details
[1]Clinical Bioinformatics, Saarland University, 66125 Saarbrücken, Germany. [2]BGI-Shenzhen, Shenzhen, China. [3]Department of Human Genetics, Saarland University, Saarbrücken, Germany. [4]Complete Genomics (a BGI company), Mountain View, CA, USA.

Received: 6 October 2016 Accepted: 4 November 2016
Published online: 21 November 2016

References

1. Veneziano D, Nigita G, Ferro A. Computational approaches for the analysis of ncRNA through deep sequencing techniques. Front Bioeng Biotechnol. 2015;3:77.
2. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75(5):843–54.
3. Ruvkun G. Molecular biology. Glimpses of a tiny RNA world. Science. 2001;294(5543):797–9.
4. Mestdagh P, Hartmann N, Baeriswyl L, Andreasen D, Bernard N, Chen C, Cheo D, D'Andrade P, DeMayo M, Dennis L, et al. Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. Nat Methods. 2014;11(8):809–15.
5. Hafner M, Renwick N, Brown M, Mihailovic A, Holoch D, Lin C, Pena JT, Nusbaum JD, Morozov P, Ludwig J, et al. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA. 2011;17(9):1697–712.
6. Backes C, Sedaghat-Hamedani F, Frese K, Hart M, Ludwig N, Meder B, Meese E, Keller A. Bias in high-throughput analysis of miRNAs and implications for biomarker studies. Anal Chem. 2016;88(4):2088–95.
7. Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol. 2008;26(4):407–15.
8. Mayer P, Farinelli L, Kawashima EHUhwgcpUS. Method of nucleic acid amplification. In.: Google Patents; 2011
9. Drmanc R, Crkvenjakov R. Prospects for a miniaturized, simplified and frugal human genome project. Sci Yugosl. 1990;16(1–2):97–107.
10. Keller A, Backes C, Leidinger P, Kefer N, Boisguerin V, Barbacioru C, Vogel B, Matzas M, Huwer H, Katus HA, et al. Next-generation sequencing identifies novel microRNAs in peripheral blood of lung cancer patients. Mol BioSyst. 2011;7(12):3187–99.
11. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010;327(5961):78–81.
12. Backes C, Meder B, Hart M, Ludwig N, Leidinger P, Vogel B, Galata V, Roth P, Menegatti J, Grasser F, et al. Prioritizing and selecting likely novel miRNAs from NGS data. Nucleic Acids Res. 2016;44(6):e53.
13. Canard B, Sarfati RS. DNA polymerase fluorescent substrates with reversible 3′-tags. Gene. 1994;148(1):1–6.
14. Tsien RY, Ross P, Fahnestock M, Johnston AJUhwgcpCAAce. Dna sequencing. In.: Google Patents; 1991
15. Church GM, Mitra RDUhwgcpEPAce. Nucleotide compounds having a cleavable linker. In.: Google Patents; 2003
16. Meder B, Backes C, Haas J, Leidinger P, Stahler C, Grossmann T, Vogel B, Frese K, Giannitsis E, Katus HA, et al. Influence of the confounding factors age and sex on microRNA profiles from peripheral blood. Clin Chem. 2014;60(9):1200–8.
17. Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, Haas J, Ruprecht K, Paul F, Stahler C, et al. A blood based 12-miRNA signature of Alzheimer disease patients. Genome Biol. 2013;14(7):R78.
18. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
19. Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2012;40(1):37–52.
20. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 2006;34(Database issue):D140–4.
21. Zhao Y, Li H, Fang S, Kang Y, Wu W, Hao Y, Li Z, Bu D, Sun N, Zhang MQ, et al. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. Nucleic Acids Res. 2016;44(D1):D203–8.
22. Hamberg M, Backes C, Fehlmann T, Hart M, Meder B, Meese E, Keller A. MiRTargetLink—miRNAs, genes and interaction networks. Int J Mol Sci. 2016;17(4):564.
23. Stockel D, Kehl T, Trampert P, Schneider L, Backes C, Ludwig N, Gerasch A, Kaufmann M, Gessler M, Graf N, et al. Multi-omics enrichment analysis using the GeneTrail2 web service. Bioinformatics. 2016;32(10):1502–8.
24. Keller A, Leidinger P, Vogel B, Backes C, ElSharawy A, Galata V, Mueller SC, Marquart S, Schrauder MG, Strick R, et al. miRNAs can be generally associated with human pathologies as exemplified for miR-144. BMC Med. 2014;12:224.

25. Keller A, Leidinger P, Bauer A, Elsharawy A, Haas J, Backes C, Wendschlag A, Giese N, Tjaden C, Ott K, et al. Toward the blood-borne miRNome of human diseases. Nat Methods. 2011;8(10):841–3.
26. Keller A, Backes C, Haas J, Leidinger P, Maetzler W, Deuschle C, Berg D, Ruschil C, Galata V, Ruprecht K, et al. Validating Alzheimer's disease micro RNAs using next-generation sequencing. Alzheimers Dement. 2016:12(5): 565-76.
27. Backes C, Leidinger P, Altmann G, Wuerstle M, Meder B, Galata V, Mueller SC, Sickert D, Stahler C, Meese E, et al. Influence of next-generation sequencing and storage conditions on miRNA patterns generated from PAXgene blood. Anal Chem. 2015;87(17):8910–6.
28. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30(9):418–26.
29. Huang X, Yuan T, Tschannen M, Sun Z, Jacob H, Du M, Liang M, Dittmar RL, Liu Y, Liang M, et al. Characterization of human plasma-derived exosomal RNAs by deep sequencing. BMC Genomics. 2013;14:319.
30. Burgos KL, Javaherian A, Bomprezzi R, Ghaffari L, Rhodes S, Courtright A, Tembe W, Kim S, Metpally R, Van Keuren-Jensen K. Identification of extracellular miRNA in human cerebrospinal fluid by next-generation sequencing. RNA. 2013;19(5):712–22.