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

A novel ultra-deep sequencing and computational analysis framework for non-coding small RNAs

Dissertation zur Erlangung des Grades eines
Doktors der Naturwissenschaften
der Medizinischen Fakultät der

UNIVERSITÄT DES SAARLANDES
2021

vorgelegt von Yongping Li
geb.am 27.02.1986 in Hunan, China

Tag der Promotion: 07.01.2022
Dekan: Prof. Michael D. Menger
Berichterstatter: Prof. Andreas Keller
Prof. Eckart Meese

“LIVING MATTER EVADES THE DECAY TO EQUILIBRIUM”

WHAT IS LIFE? by Erwin Shrödinger
First published in 1944.

A novel ultra-deep sequencing and computational analysis framework for non-coding small RNAs

Abstract

Small non-coding RNAs (sncRNAs) are essential players in all pathological and pathophysiological processes. The high evolutionary conservation, especially for microRNAs (miRNAs) from nematodes to humans make them very interesting research objects. To advance the translation and application of sncRNAs in human healthcare, it is mandatory to have a profound understanding of their expression in health and diseases, especially in the context of aging. With such knowledge we can then model the contribution of non-coding RNAs to challenging diseases such as Alzheimer's disease and other neurodegenerative disorders.

Two factors adding to an improved understanding of non-coding RNAs are high-resolution experimental approaches that measure the molecules in an as least as possible biased manner and advanced computational analysis. The latter topic covers two aspects, primary data analysis such as genome mapping but more importantly also statistical analysis with respect to the biological function and altered molecular pathways.

In my phD thesis, I contributed to the unbiased measurement of small non-coding RNAs by using combinatorial probe-anchor synthesis (cPAS) sequencing [1]. Using the new sequencing approach called DNBSEQ now, we were able to demonstrate that cPAS is not only more accurate as compared to microarrays but also that it generates a physiological distribution of non-coding RNAs that is partially lost in other sequencing approaches. Further, cPAS demonstrated a great technical reproducibility, making it of potential use for medical application. Available as high-throughput approach now, the work in my thesis is fundamental to characterize thousands of samples with millions of reads each in a reproducible and affordable manner, even limiting the hands-on time of technicians. Based on the success of the initial cPAS sequencing I worked on the advanced and even less biased analysis using the CoolMPS technology [2]. The key difference in this sequencing approach is that the detection signal is not generated by a chemically modified nucleotide incorporated in synthesized DNA but that a highly specific secondary antibody emits a light signal to sequence DNA or RNA [3]. This improved the sequencing quality significantly, at the same time lowering the sequencing cost.

As very first application for the sequencing of sncRNAs, we selected Alzheimer's disease, generating data of sufficient quality for application as clinical biomarker. As specimen types, we intentionally selected whole blood, containing the information from white blood cells, red blood cells, free circulating sncRNAs in plasma and extracellular vesicles. Notably, I also contributed to make the molecular measurements feasible as home-sampling [4], a topic that gains rapid traction not only because of the recent Sars-Cov2 pandemic.

Independent on the technology, it is essential to extract the relevant biological information from small non-coding RNA data. Using data from neurological disorders but also from other diseases such as lung cancer and from controls we first modeled how aging affects the molecular patterns [5]. Our results clearly suggest a dependency of small non-coding RNAs from the age of patients, calling for age specific diagnostic tests. One challenge is however to differentiate between causative and correlated effects. Finally, we thus collected our and others knowledge on one specific class of small non-coding RNAs, microRNAs, and model how these molecules regulate the gene expression. We included this information to miRTargetLink2 [6], a web server that can model specific gene regulatory effects in one disease such as Alzheimer's disease but that also can be applied to any other biomedical research question. Using miRTargetLink2, others and we now can answer highly complex questions – e.g., which genes are targeted by sets of miRNAs or which miRNAs target gene sets in a disease within minutes.

In sum, the advanced and least biased measurement of non-coding RNAs by deep sequencing and the advanced computational analysis developed in this work contributes to advance our understanding of the molecules in health and diseases. Further, the framework can be applied by other researchers in the context of any physiological or pathological processes in humans, mice and other animals.

Zusammenfassung

Kleine nicht-kodierende RNAs (small non-coding RNAs, sncRNAs) sind wesentliche Akteure in allen pathologischen und pathophysiologischen Prozessen. Ihr hoher Grad an evolutionärer Konservierung, von Fadenwürmern bis hin zum Menschen, machen sie zum interessanten Forschungsgegenstand. Um die Translation von sncRNAs zum Patientenwohl zu ermöglichen ist es zwingend notwendig, ein tiefes Verständnis ihrer Expression in Gesundheit und Krankheit zu haben, insbesondere im Kontext des Alterns. Mit diesem Wissen können wir dann den Beitrag von nicht-kodierenden RNAs zu herausfordernden Krankheiten wie der Alzheimer- oder Parkinson-Krankheit modellieren.

Zwei Faktoren, die zu einem verbesserten Verständnis der nicht-kodierenden RNAs beitragen, sind hochauflösende experimentelle Ansätze, die die Moleküle auf eine möglichst unvoreingenommene Weise messen, und fortgeschrittene rechnerische Analysen. Letzteres umfasst zwei Aspekte, zum einen die primäre Datenanalyse wie das Genom-Mapping, aber vor allem auch die statistische Analyse im Hinblick auf die biologische Funktion und veränderte molekulare Pfade.

In meiner Doktorarbeit habe ich einen Beitrag zur unvoreingenommenen Messung von kleinen nicht-kodierenden RNAs mit Hilfe der kombinatorischen Sonden-Anker-Synthese (cPAS) Sequenzierung geleistet [1]. Mit dem neuen Sequenzieransatz DNBSEQ konnten wir zeigen, dass cPAS im Vergleich zu Microarrays nicht nur genauer ist, sondern auch eine physiologische Verteilung der nicht-kodierenden RNAs erzeugt, die bei anderen Sequenzieransätzen teilweise verloren geht. Weiterhin zeigte cPAS eine große technische Reproduzierbarkeit, was es für den medizinischen Einsatz interessant macht. Die Methode ist inzwischen als Hochdurchsatz Methode verfügbar und erlaubt es Kohorten mit tausenden Patienten, jeweils mit Millionen an Datenpunkten, Zeit- und Kosten-effizient zu messen. Basierend auf dem Erfolg der ersten cPAS-Sequenzierung arbeitete ich an der weiterentwickelten und noch weniger verzerrten Analyse mit der CoolMPS-Technologie [2]. Der entscheidende Unterschied bei diesem Sequenzieransatz ist, dass das Detektionssignal nicht durch ein chemisch modifiziertes Nukleotid erzeugt wird, das in die synthetisierte DNA eingebaut ist, sondern dass ein hochspezifischer sekundärer Antikörper ein Lichtsignal zur Sequenzierung von DNA oder RNA aussendet [3]. Dadurch konnte die Sequenzierqualität deutlich verbessert und gleichzeitig die Sequenzierkosten gesenkt werden.

Als erste Anwendung für die Sequenzierung wählten wir die Alzheimer-Krankheit und generierten Daten von ausreichender Qualität für die Anwendung als klinischer Biomarker. Unsere Resultate beruhen dabei auf der Analyse von Vollblutproben, die sowohl das Muster von Weißen Blutkörperchen, Roten Blutkörperchen als auch frei zirkulierender und Vesikel-gebundener Moleküle widerspiegeln. Insbesondere habe ich

auch dazu beigetragen, die molekularen Messungen als Home-Sampling möglich zu machen [4], ein Thema, das nicht Zuletzt wegen der Sars-Cov-2 Pandemie schnell an Bedeutung gewinnt.

Unabhängig von der Technologie ist es wichtig, die relevanten biologischen Informationen aus den nicht-kodierenden RNA-Daten zu extrahieren. Anhand von Daten von neurologischen Erkrankungen, aber auch von anderen Krankheiten wie Lungenkrebs und von Kontrollen haben wir zunächst modelliert, wie das Altern die molekularen Muster beeinflusst [5]. Unsere Resultate deuten eindeutig auf eine Abhängigkeit der kleinen nicht-kodierenden RNAs vom Alter der Patienten hin, was nach altersspezifischen diagnostischen Tests verlangt. Eine Kern-Herausforderung ist es allerdings, zwischen Ursächlichen und Korrelierten Effekten zu unterscheiden. Daher bündeln wir unser Wissen und das von anderen Forschern über eine bestimmte Klasse von nicht-kodierenden RNAs, die microRNAs, und wie diese Moleküle die Genexpression regulieren. Diese Informationen fügten wir miRTargetLink [6] hinzu, einem Webserver, der spezifische genregulatorische Effekte bei einer Krankheit wie der Alzheimer-Krankheit modellieren kann, der aber auch auf jede andere biomedizinische Forschungsfrage angewendet werden kann. Mit Hilfe von miRTargetLink können andere und wir nun innerhalb von Minuten hochkomplexe Fragen beantworten - z. B. welche Gene von Sets von miRNAs angegriffen werden oder welche miRNAs bei einer Krankheit auf Gensets zielen.

Zusammenfassend lässt sich sagen, dass die fortschrittliche und am wenigsten verzerrte Messung von nicht-kodierenden RNAs durch Deep Sequencing und die fortschrittliche rechnerische Analyse, die in dieser Arbeit entwickelt wurde, dazu beiträgt, unser Verständnis dieser Moleküle in Gesundheit und Krankheit zu verbessern. Darüber hinaus kann das Framework von anderen Forschern im Zusammenhang mit beliebigen physiologischen oder pathologischen Prozessen bei Menschen, Mäusen und anderen Tieren angewendet werden.

Overview of Scientific Peer Reviewed Manuscripts

This is a cumulative thesis based on the following five published and peer-reviewed manuscripts. The publications included herein are identical to the published versions.

a) Fehlmann, T., ….**Li,Y.,**..., *et al.* cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin Epigenetics* **8**, 123, doi:10.1186/s13148-016-0287-1 (2016) [**YL** is contributing author]

b) **Li, Y.** *et al.* CoolMPS: evaluation of antibody labeling based massively parallel non-coding RNA sequencing. *Nucleic Acids Res* **49**, e10, doi:10.1093/nar/gkaa1122 (2021) [**YL** is first author]

c) Kern, F., Aparicio-Puerta, E., **Li, Y.**, *et al.*, miRTargetLink 2.0—interactive miRNA target gene and target pathway networks, *Nucleic Acids Research*, 2021; gkab297, [**YL** is shared first author]

d) Pirritano, M., …, **Li, Y., …,** *et al.* Next Generation Sequencing Analysis of Total Small Noncoding RNAs from Low Input RNA from Dried Blood Sampling. *Anal Chem* **90**, 11791-11796, doi:10.1021/acs.analchem.8b03557 (2018). [**YL** is contributing author]

e) Fehlmann, T., ..., **Li, Y.,** ..., et al. Common diseases alter the physiological age-related blood microRNA profile. Nat Commun. 2020;11(1):5958. Published 2020 Nov 24. doi:10.1038/s41467-020-19665-1 [**YL** is contributing author]

Table of Contents

1. INTRODUCTION 1

1.1 MIRNAS AND MIRNA BIOLOGY 1
1.2 OTHER NON-CODING RNAS 3
1.3 CHALLENGES AND PROSPECTS OF MIRNAS AS BIOMARKERS 3
1.4 ALZHEIMER'S DISEASE AND MIRNA IN ALZHEIMER'S DISEASE 7
1.5 SUMMARY AND CONTRIBUTION OF THE RESEARCH IN THIS THESIS IN CONTEXT 11

2.METHODS 13

2.1 HIGH THROUGHPUT TECHNOLOGIES FOR MIRNA SEQUENCING 13
2.1.1 INTRODUCTION OF CPAS (COMBINATORIAL PROBE ANCHOR SYNTHESIS) SEQUENCING 13
2.1.2 INTRODUCTION OF COOLMPS 17
2.1.3 SINGLE CELL RNA SEQUENCING USING DNB-C LAB 20
2.2 WEB SERVER FOR MIRNA ANALYSIS 20
2.2.1. MIRMASTER 21
2.2.2 MIRNATARGETLINK 21

3. GOAL OF THE THESIS 25

4. RESULTS 27

4.1 CPAS SEQUENCING (CLINICAL EPIGENETICS) 29
4.2 COOLMPS SEQUENCING (NUCLEIC ACIDS RESEARCH) SEQUENCING (CLINICAL EPIGENETICS) 41
4.3 MIRTRAGETLINK2 (NUCLEIC ACIDS RESEARCH) 55
4.4 DRIED BLOOD SPOT SNCRNA SEQUENCING (RNA BIOLOGY) 64
4.5 MIRNAS IN AGING AND AGE-RELATED DISEASES (NATURE COMMUNICATIONS) 71

5. CONCLUSION 86

6. BIBLIOGRAPHY 88

List of Figures

Figure 1 (miRNA biogenesis) 2
Figure 2 (sncRNA classes) 3
Figure 3 (heterogeneity in miRNA biomarkers) 4
Figure 4 (Alzheimer's disease projection in Germany 8
Figure 5 (research in context) 11
Figure 6 (Sequencing technologies for H. Sapiens) 13
Figure 7 (cPAS - DNA single strand circulation) 14
Figure 8 (DNA Nanoball generation) 14
Figure 9 (Patterned array for loading DNBs) 15
Figure 10 (core of the cPAS sequencing technology) 16
Figure 11 (Base calling in CPAS) 16
Figure 12 (CoolMPS sequencing overview) 17
Figure 13 (comparison standard MPS to CoolMPS) 18
Figure 14 (scRNA Seq workflow) 19
Figure 15 (miRNA gene binding) 21
Figure 16 (miRNA gene targeting complexity classes) 22

Abbreviations

A	**Adenine**
AD	**A**lzheimer's **D**isease
C	**C**ytosine
cPAS	**C**ombinatorial **P**robe **A**nchor **S**ynthesis
COPD	**C**hronic **O**bstructive **P**ulmonary **D**isease
CNGB	**C**hina **N**ational **G**ene**B**ank
DBS	**D**ried **B**lood **S**pots
DM	**D**iabetes **M**ellitus
DNA	**D**eoxyribonucleic **A**cid
DNB	**D**NA **N**anoballs
dsDNA	**D**ouble **s**tranded **d**eoxyribonucleic **a**cid
dsRNA	**D**ouble **s**tranded **r**ibonucleic **a**cid
EDTA	**E**thylene**d**iamine**t**etraacetic **a**cid
FACS	**F**luorescence-**a**ctivated **c**ell **s**canning
G	**G**uanine
miRNA	**M**icro **r**ibonucleic **a**cid
MPS	**M**assive **p**arallel **s**equencing
NGS	**N**ext **g**eneration **s**equencing
Nt	**n**ucleotides
RCA	**R**olling **c**ircle **a**mplifcation
RT	**R**everse **T**ranscription
RTPCR	**R**everse **T**ranscription **P**olymerase **C**hain **R**eaction
scRNA	**S**ingle **c**ell **RNA**
sncRNA	**s**mall **n**on-**c**oding **RNA**s
SRA	**S**equence **R**ead **A**rchive
ssRNA	**S**ingle **s**tranded **R**NA
U	**U**racil
UMI	**u**nique **m**olecular **i**dentifier
UTR	**U**n**t**ranslated **R**egion

1. Introduction

1.1 miRNAs and miRNA biology

MicroRNAs (miRNAs) are endogenous 22 nt RNAs that can play important regulatory roles in animals and plants by targeting mRNAs for cleavage or translational repression [7]. The small and very stable molecules that have big potential are transcribed in all the living organisms and excel by their high degree of evolutionary conservation. miRNA structures are typically predicted in stem loops structures, with one or two mature miRNAs produced from this stem loop by processing using the enzyme Dicer. The precise sequences of the mature miRNAs could be defined by cloning. The processing in a canonical- and non-canonical manner is a very complex process that is sketched in Figure 1 (miRNA biogenesis).

Single stranded RNAs can complementarily recognize, and then bind, via highly selective hydrogen bonding, to specific complementary ribonucleotide targets in the 3' prime untranslated region (3'-UTR) of specific messenger RNAs (mRNAs), and in doing so, down-regulate their expression [8 16]. As mRNAs are essential to bridge the genetic information delivery from DNA to proteins, it can be understood that miRNAs are very important for cellular processes and also will impact the phenotypes of the organisms. Since the discovery of miRNA in the early 1990s [17] , continuous and significant progress has been made on how miRNAs
[18]. Moreover, of the
features for miRNA biogenesis and genomics [19]. This is very helpful review to intrigue more scientific work on how the miRNAs could be involved in different life processes, especially, their mode of actions.

A key feature of mature miRNAs is their length, miRNAs typically consist of 25 to 30 nucleotides. Elementary ribonucleic acid sequence analysis and bioinformatics predict that a 'typical' 22 nucleotide single strand RNA that is comprised of 4 different ribonucleotides (adenine, guanine, cytosine and uridine; A,G,C,U) could have over 10^{13} possible sequence combinations or structures [20]. However, indeed, miRNAs are highly conserved between different eukaryotic tissues and for humans a little over 2,500 miRNAs are annotated in the reference database miRbase. Many experimental observations indeed indicated that there are typically only about $2{\times}10^{3}$ different miRNAs so far identified in all eukaryotic tissues [21]. That miRNAs are highly developmental stage-, tissue- and cell-specific, even in adjacent cell types suggests an extremely high evolutionary selection pressure to use only specific miRNA sequences [20-25]. In this aspect, miRNAs studies are even more promising in getting their mode of mechanism clarified in most of species.

So far, miRNAs have been identified as regulators for many essential biological processes, such as development, growth, differentiation, and neurodegenerative processes [26].

While bind at the 3′ untranslated region of target mRNA to induce translational repression [27], the 5′ untranslated region, coding sequence or promoter regions also exist [28]. Moreover, RNA-binding proteins have an important role in the regulation of miRNA activity [29] (see also Figure 1 (miRNA biogenesis)).

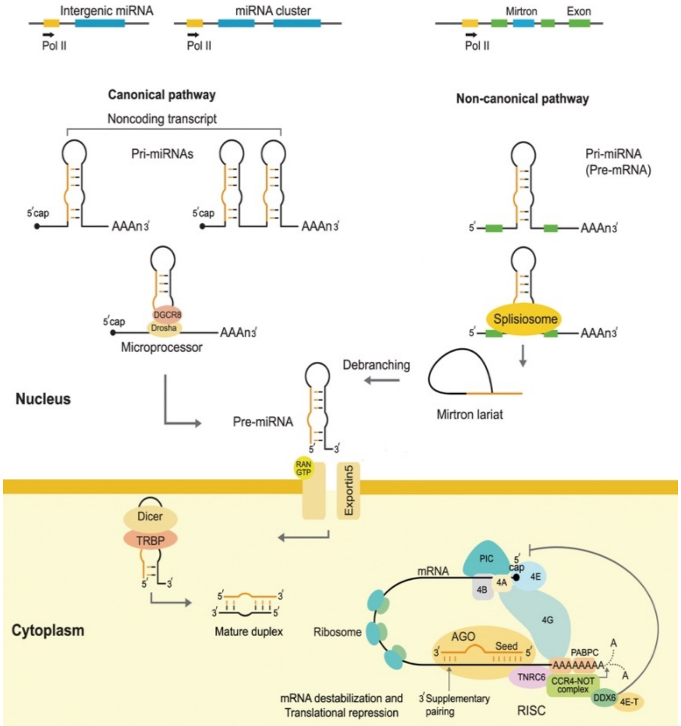


Figure 1 (miRNA biogenesis)

miRNA Biogenesis. In the usual (canonical) processing pathway pri-miRNAs are cleaved into pre-miRNAs. Pre-miRNAs are exported from the nucleus to the cytoplasm. Next, they are cleaved into small double stranded RNAs. The RIS complex mediates the recognition of the mRNA to be targeted. The non-canonical pathway, described as so-called Mirtrons, is characterized by additional splicing. The resulting RNA adopts a pre-miRNA like form and is likewise exported to the cytoplasm. This file is licensed under the Creative Commons Attribution-Share Alike 4.0 International license https://commons.wikimedia.org/wiki/File:MicroRNAs_biogenesis.jpg

Therefore, it is of great significance to study the miRNA -mRNA-proteins interactions as well as their pathways for more deeper understanding of their roles in the life processes. As miRNAs biomarkers are important for many pathological mechanisms and having great potential for diagnostic applications, miRNA-based therapies are also an extremely important research topic nowadays as well [30 37].

Some companies are also making miRNA therapies in clinical research phase such as RGENIX, Curamir Therapeutics Inc, Mirna Therapeutics Inc, Santaris Pharma etc. Those companies provided and updated their status on their homepages as well. However, there are no miRNA therapy clinically approved so far [38], which call for more solid research evidence and comprehensive understanding of miRNAs structures as well as their associated life processes.

1.2 Other non-coding RNAs

In addition to the fore mentioned miRNAs, many other non-coding RNA classes, large and small RNAs exist. While the most important class within my thesis are miRNAs, I want to briefly mention other non-coding RNA classes. These are listed in Figure 2. It is worth to explicitly highlight that not all the classes shown in this figure are small non-coding RNA. LincRNAs for example have a length exceeding several hundred bases. The data we describe in this thesis and the methodology are however based on small RNA sequencing. This means that all expression features considered herein are derived from short read sequencing. For the sake of simplicity, we thus term all the different classes described and used in this thesis as sncRNAs. While the emphasis is on miRNAs, the class with most annotated representatives are lincRNAs (49,803), followed by piRNAs (28,733). All representatives from the different classes have been extracted from the respective references, e.g. the piRNAs from piRBase and the miRNAs from miRbase. The details on the different classes are provided later in the results section in the respective manuscripts.

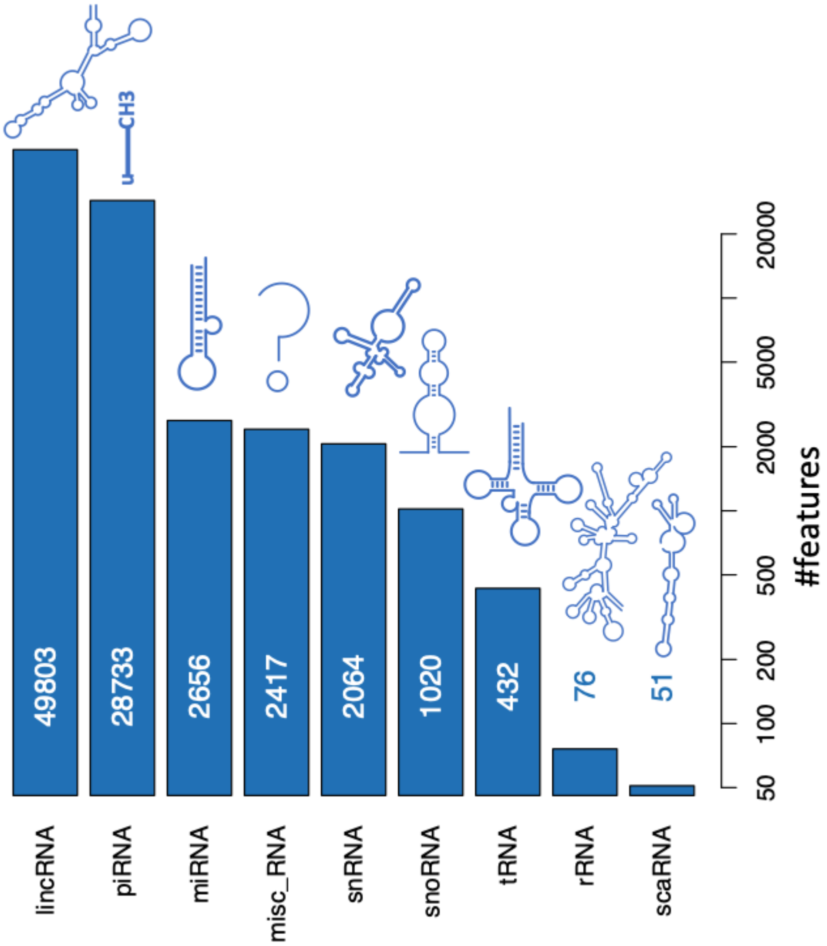


Figure 2 (sncRNA classes)

The chart displays the different small RNA classes and the number of representatives in H. Sapiens according to the respective reference databases. Please note that not all the listed RNA classes are small RNAs, since we derive the expression of the molecules however from short reads, we denote them as “small” in the context of this thesis. Source: own graphic.

1.3 Challenges and prospects of miRNAs as biomarkers

As described in the previous section, miRNAs exhibit a significant role in most physiological processes. Their targets or associated mRNAs appear to be also dysregulated in conditions of specific pathological processes.

As miRNAs are detectable in various biological fluids and are stable in dry blood spot transportation [69] and also clinical lab processing, they are considered as very valuable biomarker candidates. There are numerous studies also reporting miRNAs to be indicative for chronic diseases [39], nevertheless, it is still indeed a big challenge to get miRNAs applied in clinical patients care diagnostics. There are several processes crucially influencing this translational step, including the right collection of the samples, the measurement using an accurate profiling system and the evaluation of substantially large cohorts in a clinical setting. In the next sections I want to sketch the most relevant factors that impact the translation of miRNAs from bench to bedside. The main four reasons are sketched in Figure 3 below. The human miRNA and disease database (HMDD) currently available in the third version[1], lists serval thousand associations between miRNAs and pathologies in different specimen types, disease states, and evaluated by different technologies. In turn, the results stored in the HMDD can at the first view look contradicting each other (a miRNA going up in a disease in one study but going down in the same diseases in another study), but the results might be justified by different conditions. Examples are that the one study considers whole blood and the second plasma, or that one study considers low grades of that disease the other higher grades. Such biological factors are often hard to be separated from technical factors.

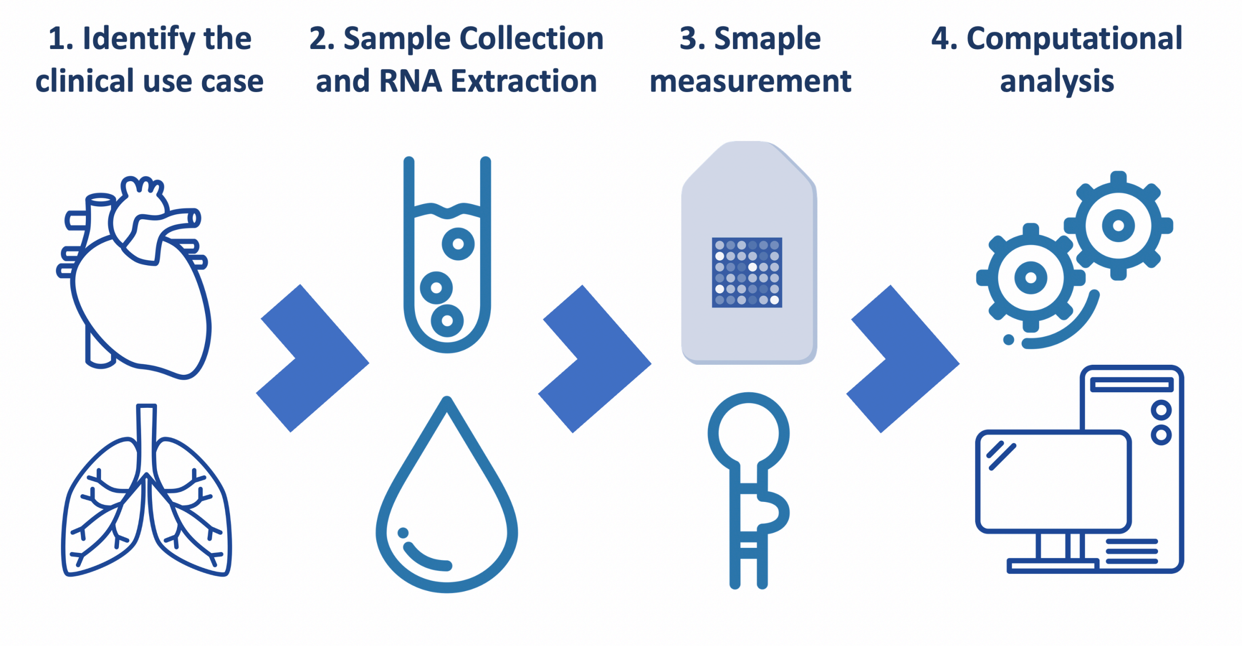

1. Identify the clinical use case
2. Sample Collection and RNA Extraction
3. Smaple measurement
4. Computational analysis

Figure 3 (heterogeneity in miRNA biomarkers)

The figure displays four important sources for heterogeneity. These include the exact definition of the clinical use case, the selection of the right sample type (referred to as matrix), the selection of the right profiling technology and the computational analysis. Image source: self-designed.

1) **Defining the right clinical use case:** As described in the first part of the Introduction and elaborated later in this section and the methods, miRNAs are relevant for

[1] http://www.cuilab.cn/hmdd

basically all human pathologies. In the context of heterogeneity, I am not referring to different disease conditions such as Alzheimer's disease (AD) and Lung Cancer (LCa) where obviously different sncRNA patterns are expected. But when comparing different research works proposing miRNA biomarkers, subtitle differences (e.g. the one study relies on low stage Lung Cancer while another relies on late stage Lung Cancers) can make substantial differences in the profile. In this context, also confounding factors have a severe impact, here, especially the age can affect miRNA profiles as we will demonstrate in the results section.

2) **Sampling:** Sampling of miRNAs could generate differences in miRNA quantity and type determination, such as storage time [40], blood collection tubes [41]. This means already the information whether anti-coagulants such as EDTA are added has an impact on profiles, of course also whole blood and plasma profiles differ. Lastly, sample matrices such as dried blood spots (DBS) again show different patterns though in principle they contain the same RNAs as whole blood. Especially the differences between DBS and whole blood will be elaborated in the results section. In a similar direction, I want to emphasize that also the purification of RNAs can impact the profiles, even between column based and gel-based size selection, differences in non-coding RNA profiles can exist.

3) **Measurement:** Further analytic techniques are also creating differences for same sample results 42 43 . For instance, next generation sequencing analysis of miRNAs was reported to introduce new isoforms of miRNAs during the RNA library preparation processes by RNA ligase [44]. While many of the earlier manuscripts on miRNAs and other RNAs in general were based on microarrays, now next generation sequencing is the de facto standard. In sum, on the one hand, technically, different methods such as RT PCR, genome arrays, next generation sequencing are applied to study the miRNAs in the samples, those data generate differences for the quantity and also different types of miRNAs information for same cohort of samples. In the context of this thesis, we will focus on stable and efficient measurement of small non-coding RNAs using such massive parallel sequencing (MPS).

4) **Data analysis:** For the data analysis, statistic methods and algorithms can also lead to different miRNAs type information. This leads from the substantial influence in using different normalization approaches (e.g. housekeeping genes or miRNAs versus distribution based normalization or others), to differences because of the application of other hypothesis tests (e.g. non-parametric Wilcoxon Mann-Whitney versus parametric t-Test) and many others. Also, this aspect is covered in my thesis as described in the Results chapter.

Besides these general challenges, often for the NGS studies of miRNA, only limited cohort sizes are available. This is due to the high cost associated with the measurement as well as the availability of clinical samples of highest quality and challenges related to the very large data set sizes. From such small – and often monocentric and retrospective studies – it is difficult to get an overall pattern of dysregulated miRNAs for the diseases being studied and also fully validate those miRNAs discovered to be true miRNAs.

Again, it is important to emphasize, that not all factors are driven by technical nature or limitations in study set up, but that the biological factors add to the complexity. Patients with same diseases or phenotypes also express several different miRNAs, since they have other confounding factors such as gender, age, or other accompanying chronic diseases or physiological conditions [5]. Or at cellular levels, miRNAs differ between plasm and serum preparations. And at molecular level, extracellular miRNAs are different between microvesicles or bounded to proteins [45]. The limitations on sample size, individual biological differences, technology leave a lot of space to further concretely depict the real expression status of miRNAs in vivo.

Given those challenges mentioned above, many further studies are extensively carried by applying new technologies and also streamline the disease cohort types in order to improve the specificity of miRNAs as well as the reproducibility of data analysis of miRNAs signatures of different disorders such as type 2 diabetes mellitus (DM), Chronic Obstructive Pulmonary Disease (COPD), breast cancer, lung cancer, Alzheimer's disease (AD) and many others [46].

Those studies, especially bundled with in-depth literature review made comprehensive comparison and careful analysis of miRNAs in different disease types, therefore, those studies indicate that a small set of miRNAs in the same sample types with similar study set-up, patient recruitment criteria seem to be largely concordant [46]. As a consequence, there are still a few percentages of each cohort, whose miRNAs are different to verify. Curated databases for storing circulating miRNAs and diseases relations could facilitate the detection of miRNAs that are applicable for diseases diagnosis [46]. As sketched above, respective databases are still hard to interpret even though they are partially manually curated to a significant amount.

Those unverified results call for more longitudinal follow up studies that could monitor changes of miRNAs along with disease progression or treatment in regular timelines. Using the new measurement devices such as next generation sequencing machines could increase the speed of data generation, novel biomarker discovery and enable large scale sample analysis significantly. However, being high throughput, their turn round time is however often more than 2-5 weeks if the samples need to be shipped to centralized labs. So, for the clinical routine diagnosis, the high throughput technology is on the one hand good for database accumulation and reference establishment, but they should be also as fast as cheap as possible. But most importantly, the quality of data generation and data analysis should be optimized [47].

Given the challenges that could potentially affect the diagnosis specificity and sensitivity of miRNAs in clinical routine, technology is further improved, and also more resources are used for studying miRNAs. This gives rise to more extensive miRNAs databases corresponding to different disease types. In this thesis, one of the topics are applying new next generation sequencing technology for miRNAs studies as well. And it is hoped that with the suitable technology, miRNAs databases on the population level, tissue level and even single cell level can be extended.

It is probably then possible to have a reasonable application - a multiplexed test of typical list of miRNAs for a specific type of diseases such as Alzheimer at specific age range for facilitating healthcare monitoring or diagnosis. In this light, AD is a promising use case for my research. While the main goal is conceptually, i.e. defining a stable, efficient and high-throughput measurement pipeline with sophisticated computational algorithms, it is important to select a clinical use case that allows to showcase the power of a potential diagnostic test.

1.4 Alzheimer's disease and miRNA in Alzheimer's disease

(source- "Dementia Fact sheet", World Health Organization, September 2020). As global aging progresses, many countries such as China, Japan, Germany, USA and others are now facing the fact that the population structures are imbalanced. For instance, people over 60s are consisting around 18-20% of the whole population number, which will definitely lead to more Alzheimer's prevalence.

It is estimated that in 2010 there were 35.6 million worldwide; these numbers are expected to 100 every 20 years by 2050 [48]. In Germany, a similar development is expected, as data from the Statistisches Bundesamt demonstrate (see also Figure 4).

As a neurodegenerative disease, Alzheimer's diseases become worse with time and is thought to begin 20 years or more before symptoms arise [48 49]. While brains change in an unnoticeable manner for imaging technologies, individuals experience also unnoticeable symptoms such as memory loss and language problems [50 53]. Often when individuals experience those symptoms, they are already too late to get disease prevention or cured.

Alzheimer symptoms occur mainly because nerve cells in parts of the brain involved in thinking, learning and cognition have been damaged or destroyed in a way that cannot be compensated or rescued.

Eventually with time, patients develop huge problems in carrying out basic body functions, need to depend other caregivers around the clock and ultimately fatal [54]. Unfortunately, so far, there is no cure or vaccine for Alzheimer's diseases. As Alzheimer's patients are dependent on others for their daily care, this gives huge burden for public health.

By the year of 2030, it is estimated that the global cost of dementia could grow to around US$2 trillion, which could overwhelm health and social care systems [55]. For families that have Alzheimer's patients, they are under huge emotional stress and sadness since communications with patients are almost blocked.

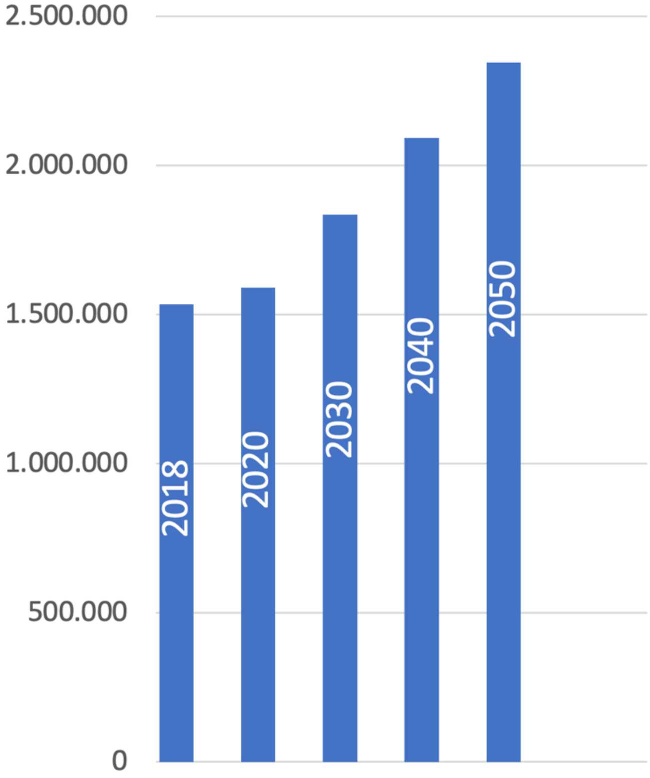


Figure 4 (Alzheimer's disease projection in Germany

The figure presents the projection of AD cases in Germany till 2050. Source: own visualization of data extracted from the Statistisches Bundesamt.

Nowadays, people are becoming aware that the diagnosis of Alzheimer should be as early as possible. However, currently, magnetic resonance tomography technology or other imaging technologies can only detect disease status in a macro-tissue level, which cannot meet the challenges to detect the molecular pathologies of Alzheimer to enable early intervention. In order to solve this biological, social challenges in the aging earth, 'omics'-based, hypothesis-free, exploratory big data pathway, will enable collection of genomics, transcriptomics, epigenomics and proteomics data from progressing asymptomatic, preclinical and clinical neurodegenerative diseases populations. Those omics data is key to the ultimate understanding of the Alzheimer's diseases and development of early diagnosis and effective individualized treatment of Alzheimer's disease [56].

Indeed, many research projects have been consistently investigating the molecular process of Alzheimer's disease using genomics and proteomics technologies [57]. The documented candidate biomarkers are described in different literatures [58 65]. However, many famous Alzheimer's biomarkers like Aß, tau, ApoE4 and others reported still are lack of enough accuracy and their prediction value remains further to be combined with other biomarkers [66 67].

Until now for Alzheimer's diagnosis, it is still challenging to have molecular diagnostics markers that are easily accessible, for instance, from peripheral blood or urine, saliva, highly specific and sensitive, less expensive and technically applicable by laboratories with standard equipment and interpretation tools. While Alzheimer's diseases are progressing with time, it is also very important to have regular and highly standardized sampling methods to enable longitudinal home sampling.

their potential as non-invasive biomarkers for human pathologies [68]. Besides, they are stable from Dried Blood Spot (DBS) samples, which could enable remote sampling and potentially good for longitudinal follow ups [69].

In a multicenter study published in 2011, Keller and co-workers profiled the expression 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 [68]. While this study did not contain AD patients or any other form of dementia, it was the basis for later AD miRNA research. The important result of the first version of the human disease miRNOme study was that it is highly important to consider a broad range of diseases as controls. Many miRNAs were dys-regulated in the same direction in various diseases, most importantly a striking down-regulation of miR-144. If such a miRNA is reported in a case-control study on AD it has likely no consequence for medical applications since it is down-regulated in the majority of all human pathologies. In continuation of the original study and beyond, miRNA patterns have been extensively studied in Alzheimer's patients' tissue samples and cell cultures [70 71]. In addition, serum profiling of circulating miRNAs by Geekiyanage etc al provided first evidence that miRNAs expression differences might be valuable biomarkers for Alzheimer's diseases diagnosis [72].

In 2013, Leidinger et al applied next generation sequencing (Illumina platform Hiseq2000) to investigate the miRNA expression profiles from blood samples, where 140 unique differentially expressed miRNAs between Alzheimer patients and controls. Further, they compared the NGS vs RT PCR methods to validate 12 miRNA biomarkers with accuracy pecificity sensitivity 90 in Alzheimers' blood samples vs health controls [66]. This study gave a very good starting point for studying deregulated miRNAs in blood for molecular diagnosis of Alzheimer's diseases. Again, the study was affected by a limitation mentioned in the previous section, which is the technology, most importantly related to the fact that whole blood miRNAs were considered. In fact, 90% of all reads measured in the study by Leidinger et al come from one single mature miRNA, namely miR-486-5p. This greatly affects the profiles of the less abundant miRNAs, one key point in my thesis.

As age is highly related to the prevalence of Alzheimer’s diseases, many further studies on miRNA have been carried to provide insights into changes in microRNAs abundance associations on age, neurodegenerative diseases etc [73 77]. The aging process is also involving different other diseases in parallel, which will in general influence the normal aging progression of miRNAs profiles in blood. For instance, people at age 40-50 have largest effect size in lung and heart diseases while the neurological diseases are more effected in age 60-70 years old. This means for a healthy individual, his or her miRNA patterns might be influenced for a given period of time by some other physiological or pathological conditions such as acute infection etc.

Therefore, it is complex to resolve the miRNA biomarkers in age dependency or related pattern, since aging is accompanied with different physiological or pathological status for individuals. And individuals with same age might have different environmental conditions as well. Besides, miRNAs have complex gene regulatory networks, it is demanding to see how the entire networks change with age. Since it would be also challenging to define at which age range, the miRNA tests should be suggested for Alzheimer’s risk analysis. To further develop understanding of age-related miRNAs changes, Fehlmann et al have used computational deconvolution methods to characterize all 2549 annotated miRNAs in 4393 whole blood samples from both genders across the lifespan (30-90 years) [78].

This study showed aging is a confounder in biomarkers discovery, which needs to be incorporated into different scenarios of other diseases. So far, many different technologies are available for high-throughput studies for different diseases but not for specific diseases at specific ages. As sketched above, to monitor the miRNome, usually microarrays and high-throughput sequencing are applied. The suitable technology will not only give good technical performances but also give good biological findings. Microarray technology could of advantage to give high dynamic range of blood miRNAs, while NGS reads could be only matched to few miRNAs [1]. Moreover, for Alzheimer’s disease, iso-forms are characteristic. This means that in patients, a modification at the 3’ end of the mature miRNA can be dominant that is less abundant in control patients. However, because microarrays can only identify known miRNAs since it is designed in a probe capture manner [79], such iso forms are frequently missed in microarray studies. It is mandatory to consider that miRNAs are often building isomiRs and basically all human miRNAs have different isoforms [80] and to apply a methodology that takes this fact into account.

To better characterize those isomiRs, next generation sequencing technology as SNP level are required. However, for the next generation sequencing data quality, also need good sample preparation and library preparation together with sequencing chemistry methods. Moreover, more biological samples need to be analyzed, which call for low-costs NGS strategies. Besides the high throughput data generation should be of high quality, low costs, the interpretation also requires integrative and intelligent strategy to mapping the miRNAs patterns in aging associated diseases such as Alzheimer [81 82]. Therefore, in this PhD thesis, many efforts are paid for investigating sequencing performance technology as described in the methods.

1.5 Summary and contribution of the research in this thesis in context

In Figure 3 on page 4 I sketched challenges in the translational research. In my thesis I tried to address topics covering the different aspects. Within a single PhD thesis not all issues related to such a complex research topic can be addressed. Most of the contributions of my work are related to the improved measurement of small non-coding RNAs (this includes miRNAs but also other sncRNAs) as well as computational analysis to add biological interpretation to the vast amount of data. The manuscripts included in my cumulative dissertation are presented in the context of the four main challenges and my contributions below in Figure 5.

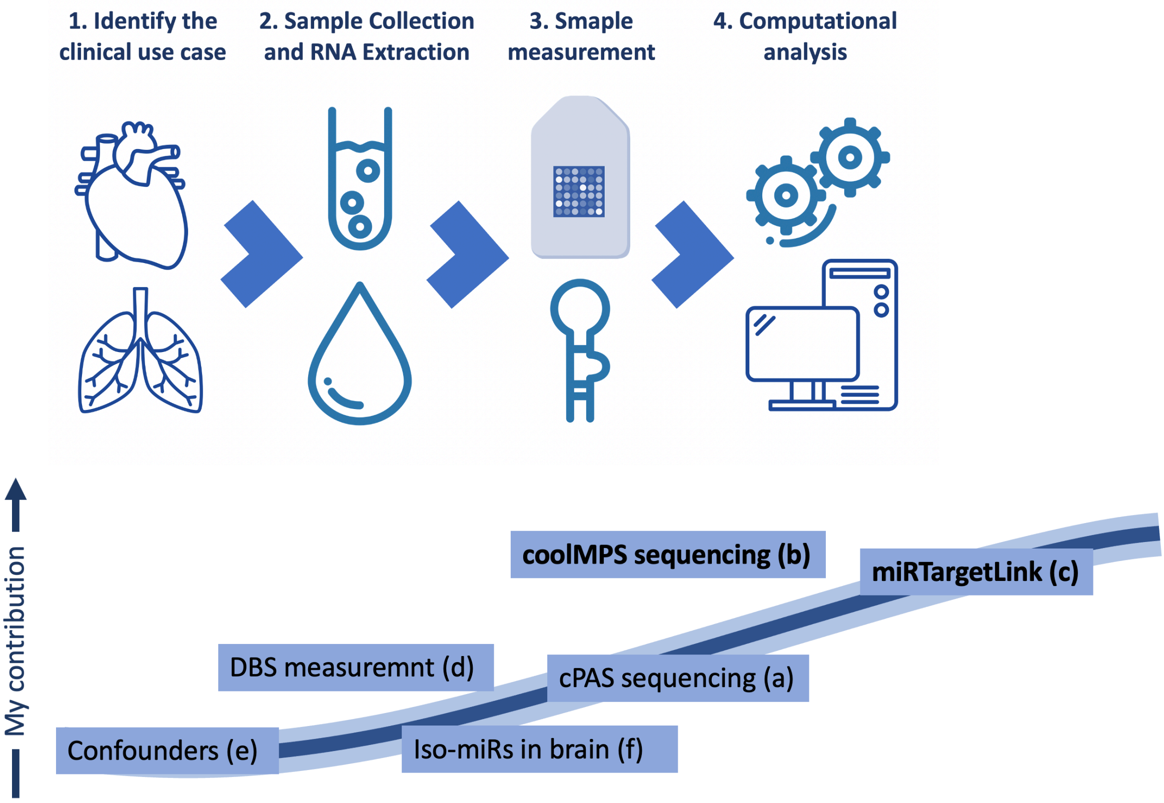


Figure 5 (research in context)

This figure shows the four main challenges described above and the contributions of my thesis to these challenges in six manuscripts. As an outlook, we additionally present a single cell study at the end. The main contributions where I am first author are highlighted in bold.

2.Methods

2.1 High throughput Technologies for miRNA Sequencing

2.1.1 Introduction of cPAS (combinatorial Probe Anchor Synthesis) Sequencing

As outlined in previous chapter, it is important to have an unbiased next generation sequencing approach to further study miRNAs in a high throughput way [83 84]. Most studies have been using Illumina sequencing platforms. Other technologies that have partially been discontinued include the 454 system by Roche and the ABI SoLiD system. In addition to these short-read sequencing systems, also long read sequencing (e.g. from PacificBio) exist. Another emerging technology includes nanopore sequencing as for example developed by the company Oxford Nanopores. Figure 6 presents a current overview of data sets available for H. Sapiens, irrespectively on whether DNA or RNA has been sequenced (data extracted from the Sequence Read Archive on May, 25th 2021). In this figure, the complete genomics technology as well as the BGISEQ technology are listed. This sequencing relies on so-called cPAS sequencing, invented by Drmanac and co-workers [85].

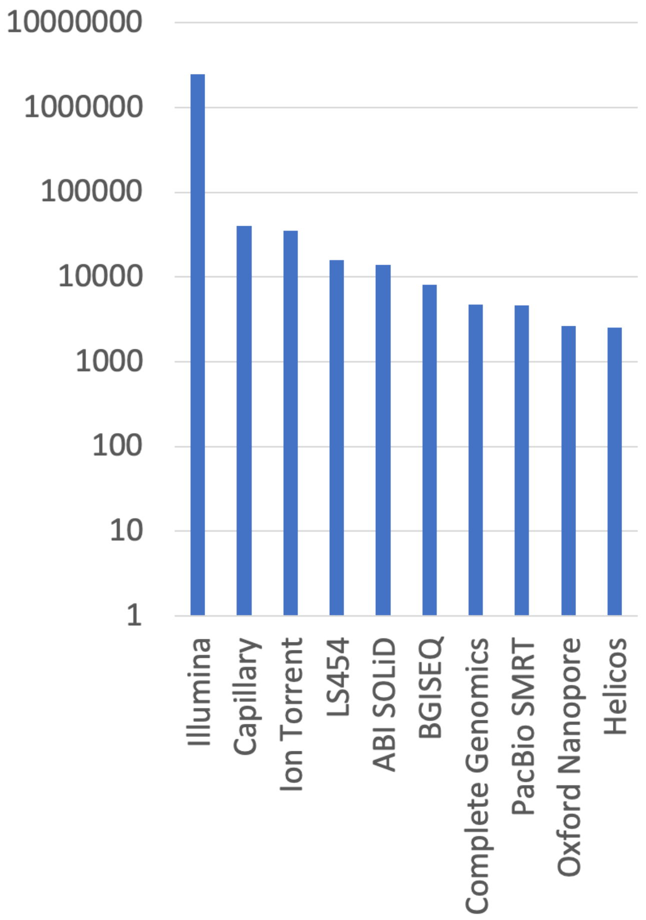


Figure 6 (Sequencing technologies for H. Sapiens)

The graphic presents the number of sequencing experiments for H. Sapiens depending on the used sequencing platform. Most of all samples has been sequenced using Illumina technology. Please note that the number of experiments on the y-axis is presented on a log10 scale. Source: own graphic based on data from the sequence read archive.

The small RNA sequencing protocol presented in this thesis relies on the same technology. I thus introduce the main principle in sufficient details. The DNA Nanoball-based sequencing technology starts from genomic DNA fragmentation which can be done from enzymatic digestions or sonic processes. The fragmented double stranded DNA (dsDNA) pieces are exposed in to 95-degree heating to generated single stranded DNAs, which are used as template for DNA single strand circularization.

The Figure 7 on page 14 schematically shows how the single strand circular DNA is produced.

After DNA single strand circulation step, DNA nanoball are generated using a high-fidelity enzymatic process. In the next step, the DNB nanoballs are generated (Figure 8).

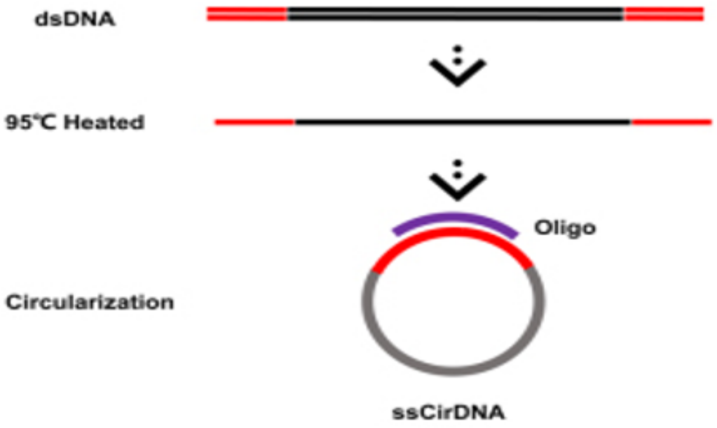


Figure 7 (cPAS - DNA single strand circulation)

DNA Single strand circulation. D DNA with is heated generate ssDNA (single stranded DNA). A splint oligonucleotide with a complementary sequence to both the 5' and 3' terminal ends of one strand of the target dsDNA will hybridize to both the 5' and 3' terminal ends of the same target ssDNA to form a nicked circle. sing DNA ligase to form a single stranded circle. (source- MGI Tech Co.,Ltd Sequencing DNB Platform Product Brochure).

After DNBs are created, their concentration are measured or quantified with Qubit devices before loading onto the sequencing chip. The quality control step of DNBs is important for the sequencing quality and can be done with common devices inexpensively. For the DNB sequencing technology, there are some benefits. Firstly, it used the high-fidelity DNA polymerase during the rolling circle amplification (RCA) step, this could help reduce the errors by amplification. In addition, as depicted in Figure 7, the amplification always uses the original single strand DNA circle as template, which will avoid the propagation or accumulation of amplification errors. So, these two factors

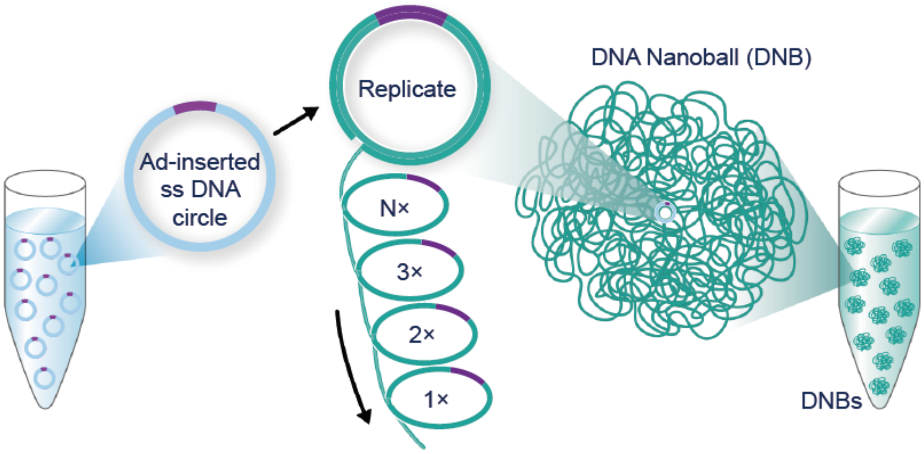


Figure 8 (DNA Nanoball generation)

DNA nanoballs are by rolling RCA template. 100-1000 copies of DNA fragments with various sizes. (source-MGI Tech Co.,Ltd DNBSEQ technology introduction)

In addition, RCA technology avoids errors , GC biases and dropouts observed with other amplification , such as PCR according to the MGI Tech Co., Ltd DNBSEQ introductions. In summary, the DNB preparation is thus a very promising method during sequencing sample preparation and could help improve the data quality significantly.

After sample preparation, DNBs are then loaded and sequenced on a patterned array. DNBs since they have phosphate backbone, and the array surface is positively charged. In this way, DNBs will stay on the array surface. The DNB binding sites are created on the surface of a silicon chip from state-of-art semiconductor manufacturing process. The size of binding site spots is designed in a way that only one sing DNB can bind to each binding spot. In addition, the distance between active spots is uniform

This patterned array method can help get high sequencing accuracy, high chip usage and optimal reagent usage. The positive and negative charge interaction helps maintain the DNBs once they are loaded on the slide surface of flow cells.

The sizes of DNBs and the active binding sites on the slide surface are balanced, so that they are same or similar size, which could ensure the high spots yield. The loading process of DNBs is depicted in Figure 9.

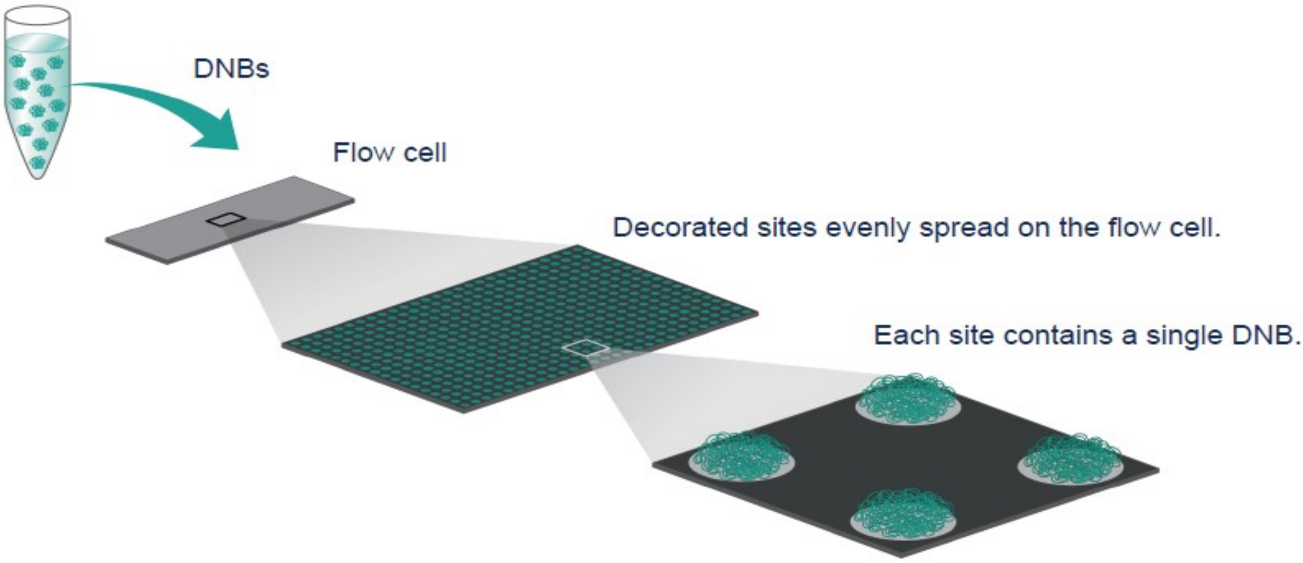

DNBs
Flow cell
Decorated sites evenly spread on the flow cell.
Each site contains a single DNB.

Figure 9 (Patterned array for loading DNBs)

Active spots on the chip surfaces are distributed uniformly to ensure only one single DNB is bound to each binding site. Source: MGI Technology

After loading DNBs to the flow cells or slide surfaces, sequencing is done using CPAS technology. The primers are hybridized to the adapter region of the DNBs, and during the sequencing by synthesis step, the fluorescently labeled dNTPs probes will be detected and imaged (Figure 10).

The sequencing reaction time has been reduced to less than one minute. This is reached by significant improvements in sequencing biochemistry, as well as the identification of a superior sequencing polymerase screened from tens of thousands of mutants.

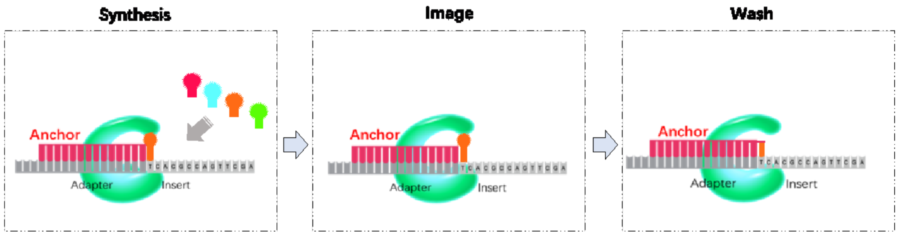
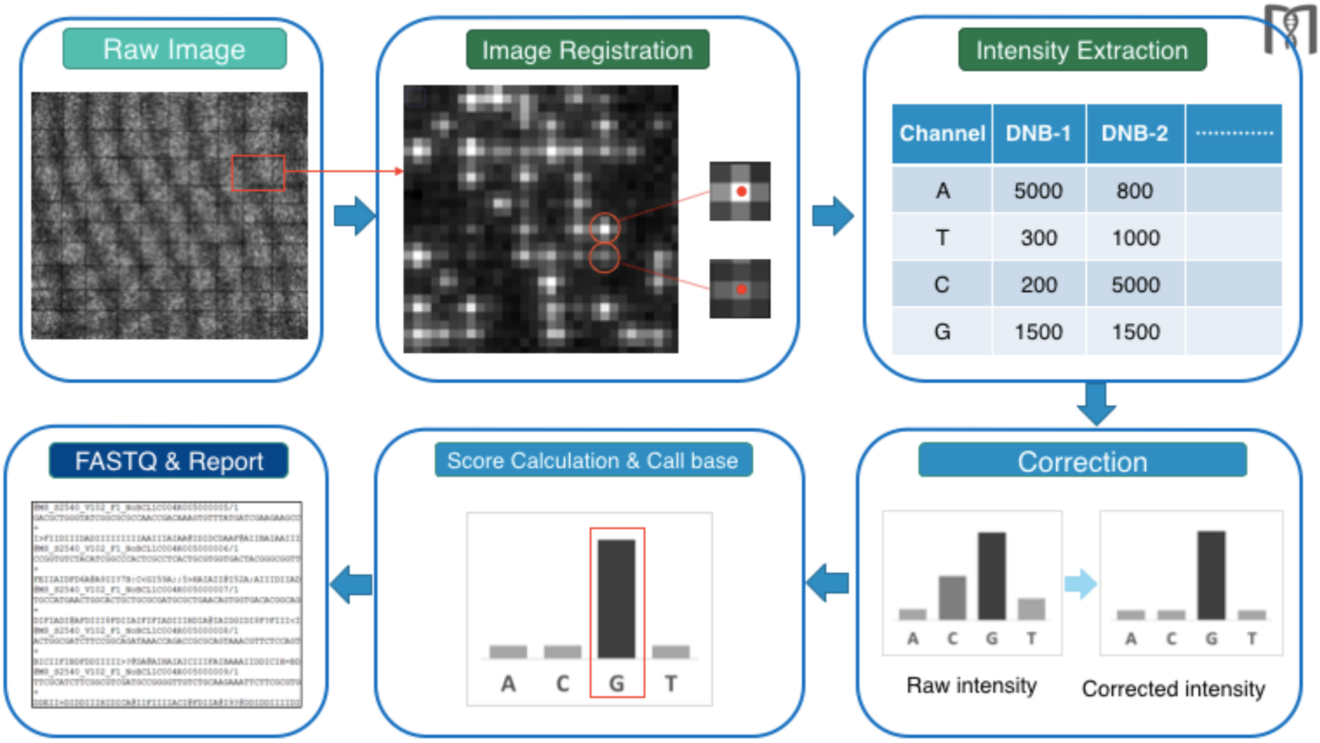


Figure 10 (core of the cPAS sequencing technology)

After sequencing primers are hybridized to the DNB , a dNTP probe is incorporated with a DNA polymerase (Figure 4). MGI's proprietary base-calling software . , (Source—MGI Tech Co.,Ltd DNBSEQ Platform Product Brochure).

Once the fluorescence dNTPs signals are converted to electrical signals, base calling and registration is done and data could be further filtered for high quality Q30, Q20 scores. Figure 11 describes the base calling step.



Channel	DNB-1	DNB-2	…………
A	5000	800	
T	300	1000	
C	200	5000	
G	1500	1500	

Figure 11 (Base calling in CPAS)

ignal intensities from all channels .

Quality scores are based on phred-33 standard.

(source: DNBSEQ Technology Information provided by MGI Tech Co.,Ltd)

In summary, the DNBSEQ Technology is using linear amplification of single circular DNA templates for generating billions of DNBs, which are then sequenced on a patterned array and gives high quality sequencing data. The linear amplification has unique advantages for miRNAs sequencing to avoid amplification errors propagation, as reported in literature [1]. Moreover, the error rate is independent of the sequencing length. This means, especially towards the end of the reads where the miRNA/gene targeting is determined the required low error rate is reached. In the first part of the results section, we describe in detail how we adapted the sample preparation and library generation step to achieve highly accurate sequencing data for small RNAs.

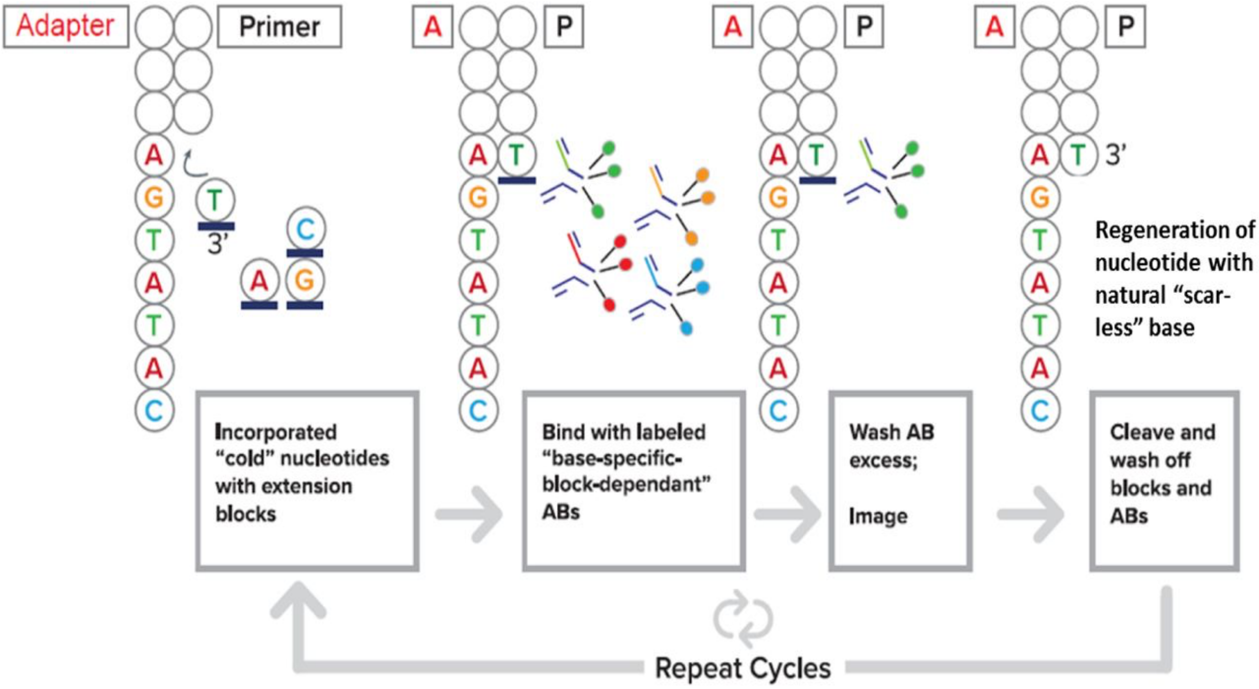

Adapter
Primer
A
G
T
A
T
A
C
T
3'
A
G
C
A
P
3'
Regeneration of nucleotide with natural "scar-less" base
Incorporated "cold" nucleotides with extension blocks
Bind with labeled "base-specific-block-dependant" ABs
Wash AB excess;
Image
Cleave and wash off blocks and ABs
Repeat Cycles

Figure 12 (CoolMPS sequencing overview)

The figure schematically presents the four steps in the CoolMPS sequencing procedure that are repeated in each sequencing cycle. Bars (-) on the nucleotides depict 3' chemical . RTs with natural nucleobase are with three dye molecules Source: Figure is taken from reference [3]

2.1.2 Introduction of CoolMPS

on DNA nanoarrays provides . It readily excels by a low amplification bias and highly accurate sequencing results across the whole read length and independent on the position in the read. Nonetheless, the signal detection can be further improved. While the other sequencing technology is applying modified nucleotides for blocking the synthesis step and using fluorescently labeled nucleotides for imaging, the disadvantages are that the signals will diminish with the sequencing progressing. Further the common MPS is limited by read length, usually only paired or single end 100, 150, 200 base pairs are available, which will cause difficulties of structural variants analysis during later bioinformatics. Therefore, further improvement in read length, sequence quality on the sequencing technology is necessary. Besides, cost reduction

of a novel MPS chemistry (CoolMPS™) utilizes nucleotides and four natural nucleobase (A, T, C, G)-specific fluorescently labeled antibodies with fast (30 seconds) binding. The CoolMPS sequencing method is antibody-based, but it is compatible with the DNB sample preparation. This in fact means that can sequence the same libraries with the same sequencer, only by exchanging the sequencing chemistry. Therefore, CoolMPS™ data generation can be implemented on MGI's DNBSEQ MPS platform using arrays of 3 .

The sequencing process is based on DNA polymerase catalytic activity. Natural unlabeled nucleotides or bases are binding with fluorescently labeled antibodies specifically. Each antibody can be labeled with one or more fluorescent molecules to get higher signal intensity. Therefore, for same copy numbers of DNB, CoolMPS sequencing can have stronger signals and higher resolution signals, which will lead to more accuracy of data generation (source-MGI Tech Co., Ltd CoolMPS Sequencing set

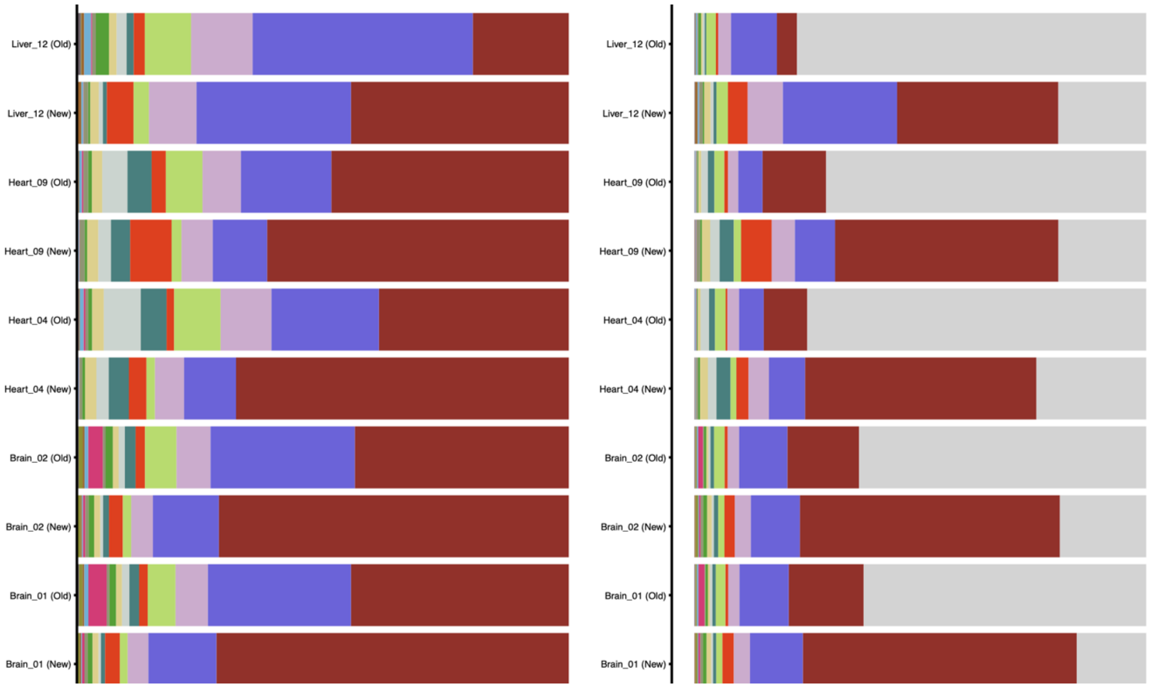


Figure 13 (comparison standard MPS to CoolMPS)

The figure presents the matching of reads to the different sncRNA classes across a wide variety of mouse tissues (left panel). For each tissue the old (standard MPS) and new (CoolMPS) performance is presented. The CoolMPS approach highlights the intended enrichment for miRNAs. The panel on the right-hand side presents the mapping also considers unmapped reads (grey fraction). Here, CoolMPS has a significantly higher mapping rate as compared to standard MPS. Source: own graphic based on data from the Tabula Muris Senis collection.

Brochure). It is reported that DNBs with less than 50 template copies were successfully sequenced by strong -signal CoolMPS with 3- times higher accuracy than in stand MPS [3].

Because CoolMPS is using the unlabeled nucleotides, this will avoid 'scars' on the bases during incorporation. Therefore, as sequencing is moving on, CoolMPS will not have the problems of 'scar' accumulation, which process is affecting the binding activity of the DNA polymerase and increase the error rate of the sequencing. While having no 'scar' accumulation issue, it is in principle that the sequencing read length can be prolonged with low out-of-phase incorporation for CoolMPS approach. CoolMPS™ chemistry mode of mechanism is sketched in Figure 12 on page 17.

As described in the previous section, CoolMPS has many potential advantages over standard MPS. This has been verified by investigating various tissues from the Tabula Muris Senis study and underlines that CoolMPS is the perfect sequencing technology for deciphering complex sncRNOmes (Figure 13). The data presented in this Figure contain an additional novel library preparation procedure. Instead of manual size selection from gels, an automated magnetic bead selection was used. By applying this method on the SP960 sample preparation robot we now can scale up the sequencing of sncRNA data sets to a 96-well plate format and produce 288 data sets each with 30 Million reads within a week.

In my PhD thesis work, I contributed to the development of the advanced CoolMPS approach for miRNA sequencing and the results are described in the paper published [2] and highlighted in Figure 5 on page 11.

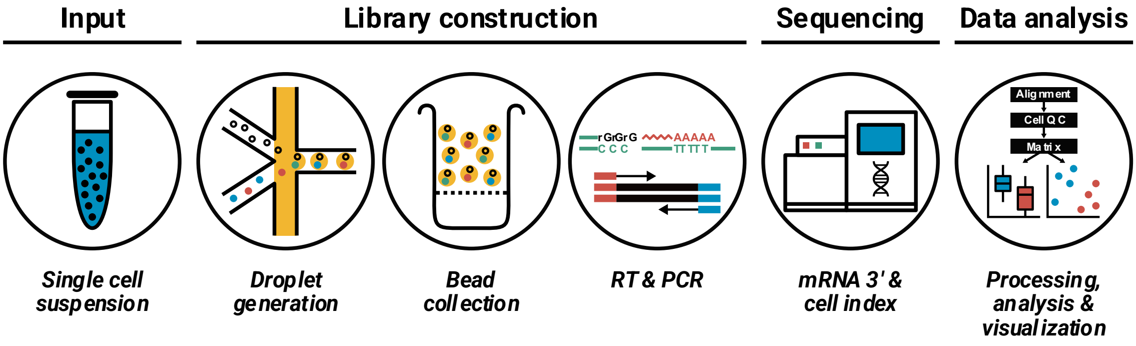


Figure 14 (scRNA Seq workflow)

The figure presents the single cell RNA seq workflow. Blood cells are brought into a single cell suspension, oil droplets with each having a blood cell and a barcode in the optimal case are generated. The beads are collected, the RNA in amplified and then all RNAs with the attached barcodes identifying the cell of origin are sequenced using cPAS sequencing. Finally, the data are analyzed (this step is done by Tobias Fehlmann). Source: Own graphic.

2.1.3 Single Cell RNA Sequencing using DNB-c Lab

As described in the previous sections we applied the cPAS and CoolMPS sequencing to measure sncRNAs in a so-called bulk manner. Here, the information from different blood cell types and the different representative cells from this blood cell type are mixed, adding significant noise to the data. For sncRNAs that have no poly-A tail, single cell sequencing does not exist at scale. Since my aim was nonetheless o get RNA single cell data, I worked towards the end of my thesis to get single cell RNA data. Basically, we made use of a droplet-based sequencing system as sketched in Figure 14. In some detail, we generated a single-cell suspensions for oil droplet generation. Thereafter, we did an emulsion breakage and beads collection step. To increase the number of molecules, we did reverse transcription (RT), and cDNA amplification. From the amplified products we in turn produced the barcoded libraries. The sequencing libraries contain the 3' UTR of a gene of interest, a unique molecular identifier (UMI) to determine the cell of origin and the sequencing adapter. The detailed read structure of the paired-end reads consist of Read 1 (covering 30 bases inclusive of 10-bp cell barcode 1, 10-bp cell barcode 2 as well as 10-bp UMI), and Read 2 (containing 100 bases of transcript sequence as well as 10-bp sample index). The sequencing of the libraries using DNA nanoballs was not done on the sequencing system installed at Saarland University, because they lack the required throughput. Instead, the sequencing was done as a service using the ultra-high-throughput DIPSEQ T1 sequencer at China National GeneBank (CNGB). Altogether, we sequenced 1.1 Million cells from 400 individuals, including AD patients, patients with mild cognitive impairment, Parkinson's disease and unaffected controls of matched gender and age. The data analysis has been mainly performed by Tobias Fehlmann and he will describe the results in his PhD thesis.

2.2 Web Server for miRNA Analysis

Having the right tools for stable measurement of sncRNAs with an unprecedented single base resolution of highest quality calls for improved computational analyses. Here, two factors are to be differentiated, the primary analysis and the secondary analysis. While the primary analysis in the context of this thesis includes the step from fastq files to biomarker profiles, the secondary analysis takes care on adding biological sense to the data. Remarkably, this is different form a frequent other definition where the primary analysis ends already with the sncRNA expression matrix. Since the purpose of this work is however to make complex AD biomarker panels explainable, this letter important step is considered separately. In this direction, I did not contribute to the development of the miRMaster tool that I sketch only for the sake of completeness in Section 2.2.1. miRMaster, but focused on the development of miRTargetLink described in Section 2.2.2 MiRNATargetLink.

2.2.1. miRMaster

The CoolMPS technology generates millions and millions of small RNA reads that must be processed. With standard software implemented in scripting languages such as Perl, we would not reach the required computing speed. Thus, the Chair for Clinical Bioinformatics developed the miRMaster tool, which is now available in the second version [85]. The program is available as web service and in the backend implemented in efficient C++. In brief, small RNA read data sets are uploaded in fastq file format and mapped to the reference genomes and small RNA resources mentioned in the Introduction (see Figure 2 on page 3). If not mentioned explicitly, I applied miRMaster using the standard parameters. Remarkably, miRMaster has own modules for processing also CoolMPS data generated using MGISEQ platforms. As output, miRMaster generates tables and images that describe the expression of the different representatives of RNAs across the samples and computes biomarker profiles in case-control studies. From the data returned by miRMaster it is however hard to conclude whether biomarkers are just correlated to a disease phenotype or whether they have a potential to be causative. For making respective conclusions, I contributed to the development of miRTargetLink, a web-based software that computes complex regulatory networks between miRNAs and target genes. miRMaster has been mainly developed by Tobias Fehlmann and I appreciate his guidance in evaluating my sncRNA data using miRMaster.

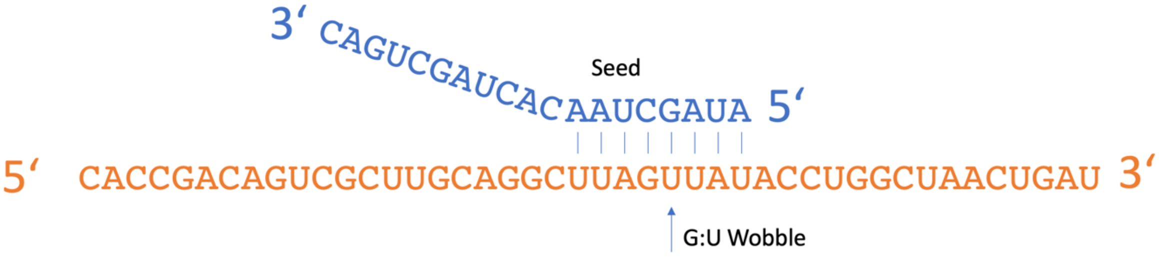

3‘ CAGUCGAUCACAAUCGAUA 5‘
Seed
5‘ CACCGACAGUCGCUUGCAGGCUUAGUUAUACCUGGCUAACUGAU 3‘
G:U Wobble

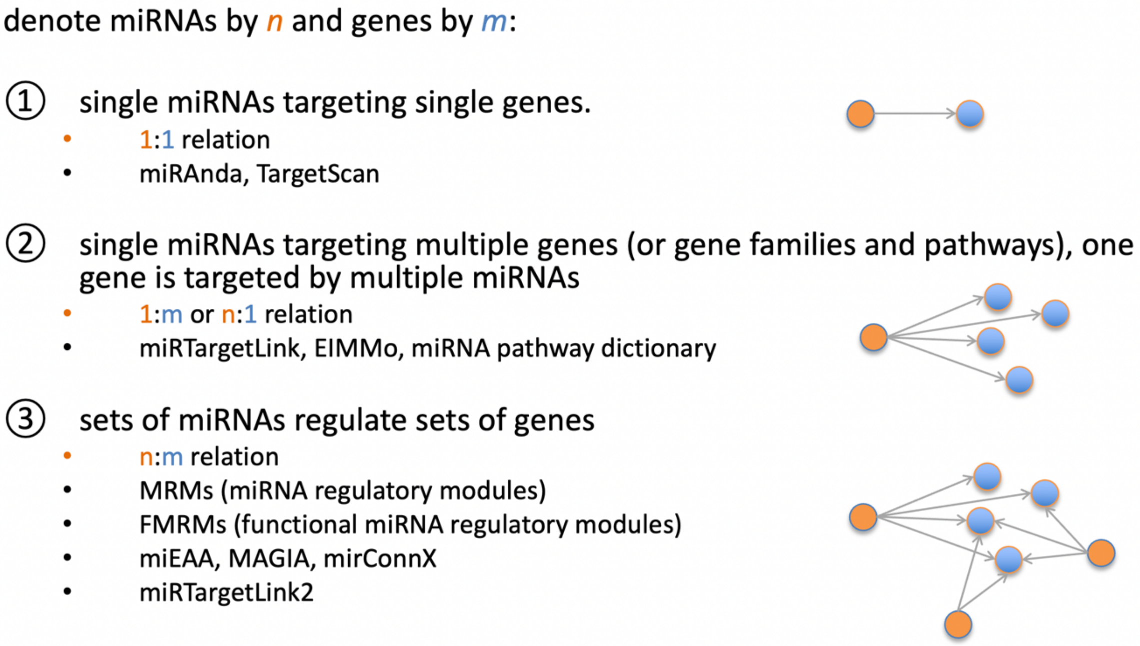
Figure 15 (miRNA gene binding)

The figure describes the binding of a miRNA to a target gene. The 3‘ UTR of the target gene is shown in orange, the miRNA in blue. The seed region of around 7-8 nucleotides is deterministic for the binding although shorter pattern can also initiate binding and gene repression. Often, a G:U wobble can be observed in the seed region, which is not impairing the binding. Source: own graphic.

2.2.2 MiRNATargetLink

With the application of high throughput sequencing technologies as well as other proteomics technology, more and more research projects have been carried out for miRNA involved life processes. For miRNAs, the canonical pathway includes the binding of target genes, containing the seed region of a miRNA and the 3' Untranslated Region (UTR) of a gene (see Figure 15).

Theoretically, in *H. Sapiens*, 2,500 miRNAs can regulate the expression of 25,000 genes, over 60 million combinations. Of course, only a fraction of those is biologically functional. Computational approaches to predict which of the miRNA – gene interactions are functional have limited accuracy, specificity, and sensitivity. An excellent overview on miRNA gene targeting in general and with a focus on computational target prediction has been published by Kern and co-workers [86]. Nonetheless, databases store such predicted interactions, partially as consensus predictions. Other data bases rely on validated interactions, including weak evidence targets (such as negative correlation of miRNA and gene expression) and stronger evidence targets (such as reporter assays). There are eight typical databases available for miRNAs, partially tailored for target information and partially including target information as part of a broader functionality. Examples include miRBase for miRNA annotation; miRTarBase, mirDIP, miRDB, miRATBase for miRNA targets; miRPathDB for miRNA target pathway database, miEAA for miRNA set enrichment analysis and GeneTrail for Gene set enrichment analysis [6]



denote miRNAs by *n* and genes by *m*:

① single miRNAs targeting single genes.

- 1:1 relation
- miRAnda, TargetScan

② single miRNAs targeting multiple genes (or gene families and pathways), one gene is targeted by multiple miRNAs

- 1:m or n:1 relation
- miRTargetLink, EIMMo, miRNA pathway dictionary

③ sets of miRNAs regulate sets of genes

- n:m relation
- MRMs (miRNA regulatory modules)
- FMRMs (functional miRNA regulatory modules)
- miEAA, MAGIA, mirConnX
- miRTargetLink2

Figure 16 (miRNA gene targeting complexity classes)

The figure presents different complexity classes of how interactions between miRNAs and genes can be modeled. The most realistic assumption is the n:m relation presented in the lower panel of the figure.

Those databases contain high amount of miRNAs information, for instance, miRbase in its most recent release version 22 (October 2018) contains 38,589 entries from 271 species and MirGeneDB contains 10,899 curated miRNAs from 45 different organisms [2]. Besides, there are around 11,000 annual publications on miRNAs [2], which add to an important resource of miRNA knowledge, however, it is a demanding search work about how and what miRNAs regulate which specific genes, pathways. In fact, miRNAs and genes are not following simple 1:1 relation, one miRNA is regulating one gene. Indeed, one gene can be repressed by multiple miRNAs and one miRNA can repress multiple genes. Typically, cooperative effects exist, including several binding sites of the same miRNA in one UTR of a gene or co-located binding sites of different miRNAs in the UTR. Computational tools must consider these complexity stages (Figure 16).

Therefore, an integrative solution from different data sources to present quick and comprehensive answers to help visualize the miRNAs targets are needed urgently, miRNA TargetLink 2 is developed to enable search for miRNA targets in bi-directional mode. In the miRNA TargetLink 2 project, data is selected from existing databases such as miRBase (v22.1) and validated targets from miRTarBase (v.8) and miRATBase. In addition, mirDIP (v4.1) is used for predicted miRNA targets for human, miRDB (v.6) for mouse and rat.

Besides, miRTargetLink supports target pathways from miRPathDB 2.0. miEAA 2.0 and GeneTrail 3 are the tools used for gene set annotation analysis. An overview of databases and tools used for miRTargetLink 2 is provided in Table 1.

Table 1 Overview of in-house and third-party resources included in miRTargetLink 2.0

Database	function	type	entries	version	source
miRBase	miRNA annotation	database		22.1	third party
miRTarBase	miRNA annotation	database	553 000	8.1	third party
mirDIP	miRNA target database	database	1519000	4.1	third party
miRDB	miRNA target database	database	1173000	6.0	own
miRATBase	miRNA target database	database	300	1.0	own
miRPathDB	miRNA target pathway database	database	13000	2.0	own
miEAA	miRNA set enrichment analysis	tools	n.a	2.0	own
GeneTrail	Gene set enrichment analysis	tools	n.a	3.0	own

Those applications of tools and databases enable a comprehensive background resources for the miRNA search analysis, so researchers can freely choose for one specific miRNA to understand their targets and associated pathways. Or they could apply the miRTargetLink 2 to get the miRNA results under the pathways of interests.

3. Goal of the thesis

As outlined in the previous chapter, the importance of miRNAs in biology needs high throughput sequencing data generation and analysis in a cost-effective way. From the cPAS-based sequencing on BGISEQ-500 to explore small non-coding RNAs, I was motivated significantly to explore further new technologies that can enable further data improvements and cost reduction on miRNA research. Being high throughput, advanced computational analysis is also of great importance.

Therefore, the primary goal of this thesis focuses on the new technology CoolMPS, to evaluate their performances in a comprehensive way. Subsequently, we are applying integration of data analysis solve the computing challenge by providing an interface of input miRNAs and get their targetome network information.

Therefore, the analysis of a large collection of miRNA samples and their profiles are done. The studies of cPAS and CoolMPS technologies on miRNA were performed in collaboration with industry partners such as BGI Research Institute and MGI Tech Co., Ltd and Complete Genomics.

To offer a good computational analysis solution for miRNAs, a web server, miRNA Target Link, was implemented to provide access to interactively analyze the miRNA in the uni-directional and multi-directional way. While setting up the framework of data generation as well as data interaction analysis, we also investigated the miRNA sequencing in aging associated diseases such as Alzheimer diseases.

One core observation in my research was that data from mixtures of cells have certainly a diagnostic and prognostic value but that single cell resolution data will most likely add significantly to reduce the noise. Especially for the bulk sequencing of whole blood, two factors contribute to data noise. First of all, the sncRNAs are measured from all blood cell types, irrespectively of their origin. This could be cured by positive selection of one or several cell types, by negative selection or by fluorescence-activated cell scanning (FACS). Still, we would only measure signals from many cells of that cell type. Here, single cell RNA sequencing is a promising option to improve the situation. While for sncRNAs, single cell sequencing approaches are available only in a low-throughput manner, they are available for gene expression profiling at scale. Towards the end of my work, I thus supported the development of a complete map of neurodegenerative gene expression across a blood cells.

4. Results

This cumulative thesis is mainly based on following peer-reviewed publications whose published versioned are included in this chapter.

a) Fehlmann, T., ….**Li,Y.,**..., *et al.* cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin Epigenetics* **8**, 123, doi:10.1186/s13148-016-0287-1 (2016) [**YL** is contributing author]

b) **Li, Y.** *et al.* CoolMPS: evaluation of antibody labeling based massively parallel non-coding RNA sequencing. *Nucleic Acids Res* **49**, e10, doi:10.1093/nar/gkaa1122 (2021) [**YL** is first-author]

c) Kern, F., Aparicio-Puerta, E., **Li, Y.**, *et al.*, miRTargetLink 2.0—interactive miRNA target gene and target pathway networks, *Nucleic Acids Research*, 2021;, gkab297, https://doi.org/10.1093/nar/gkab297 [**Li,Y** is shared first author]

d) Pirritano, M., …, **Li, Y., …,** *et al.* Next Generation Sequencing Analysis of Total Small Noncoding RNAs from Low Input RNA from Dried Blood Sampling. *Anal Chem* **90**, 11791-11796, doi:10.1021/acs.analchem.8b03557 (2018). [**Li,Y** is contributing author]

e) Fehlmann, T., ..., **Li, Y.,** ..., et al. Common diseases alter the physiological age-related blood microRNA profile. Nat Commun. 2020;11(1):5958. Published 2020 Nov 24. doi:10.1038/s41467-020-19665-1 [**Li,Y** is contributing author]

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

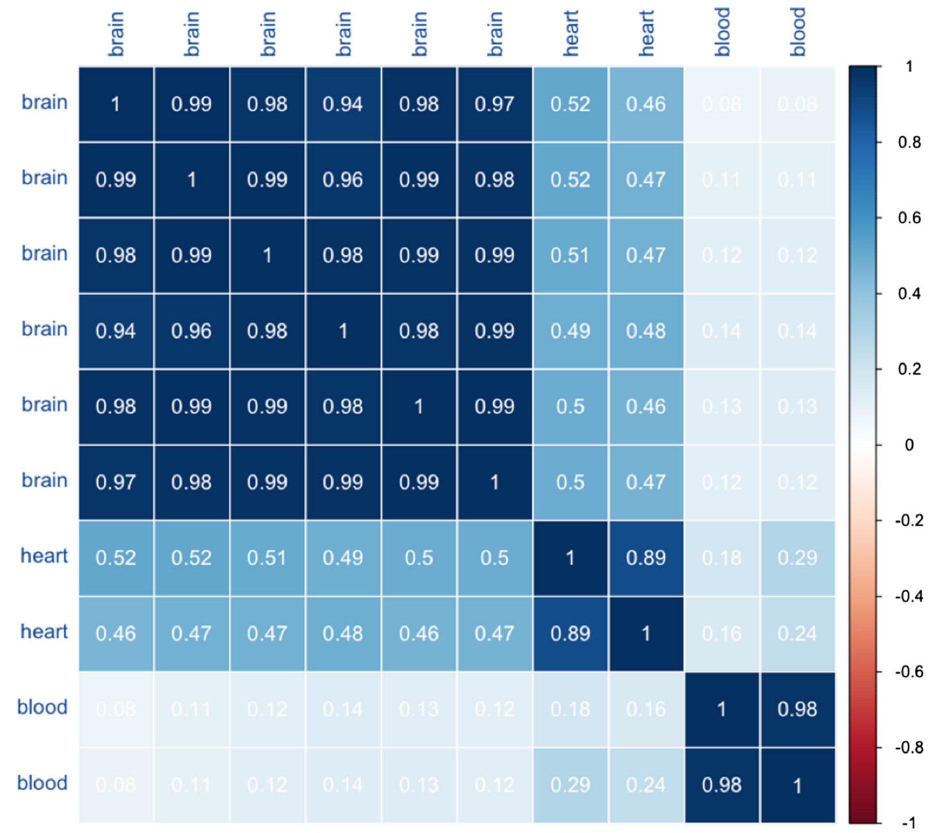


Fig. 1 Correlation matrix of the brain (six technical replicates), heart (two technical replicates), and blood (two biological replicates) sequenced by the BGISEQ-500 system

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

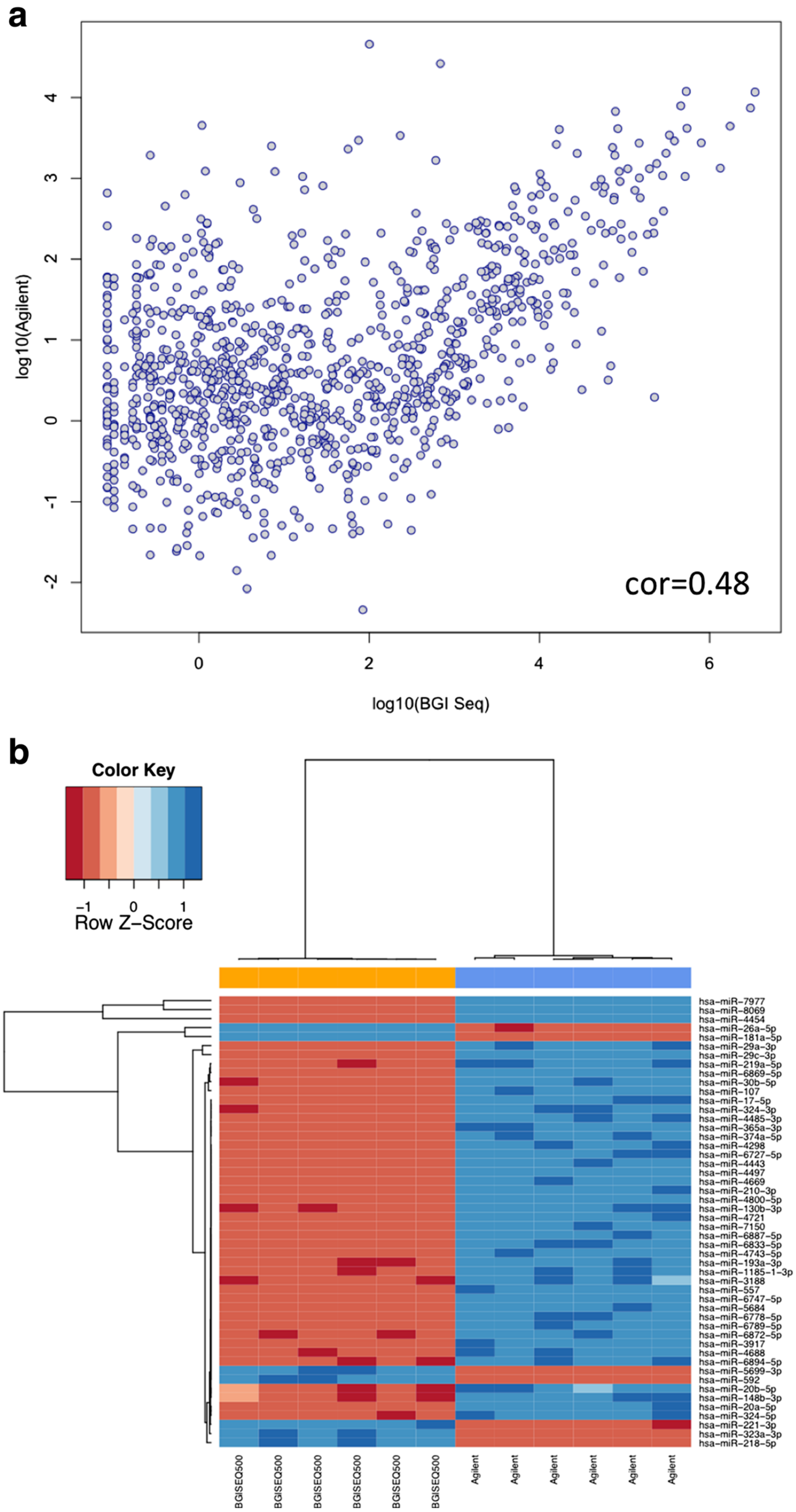


Fig. 2 a Log average expression of common miRNAs for the brain RNA on BGISEQ-500 and on Agilent microarrays (six technical replicates each). **b** Heat map with dendrogram for the 50 most differently detected miRNAs in the brain RNA between Agilent and BGISEQ-500 (six technical replicates each)

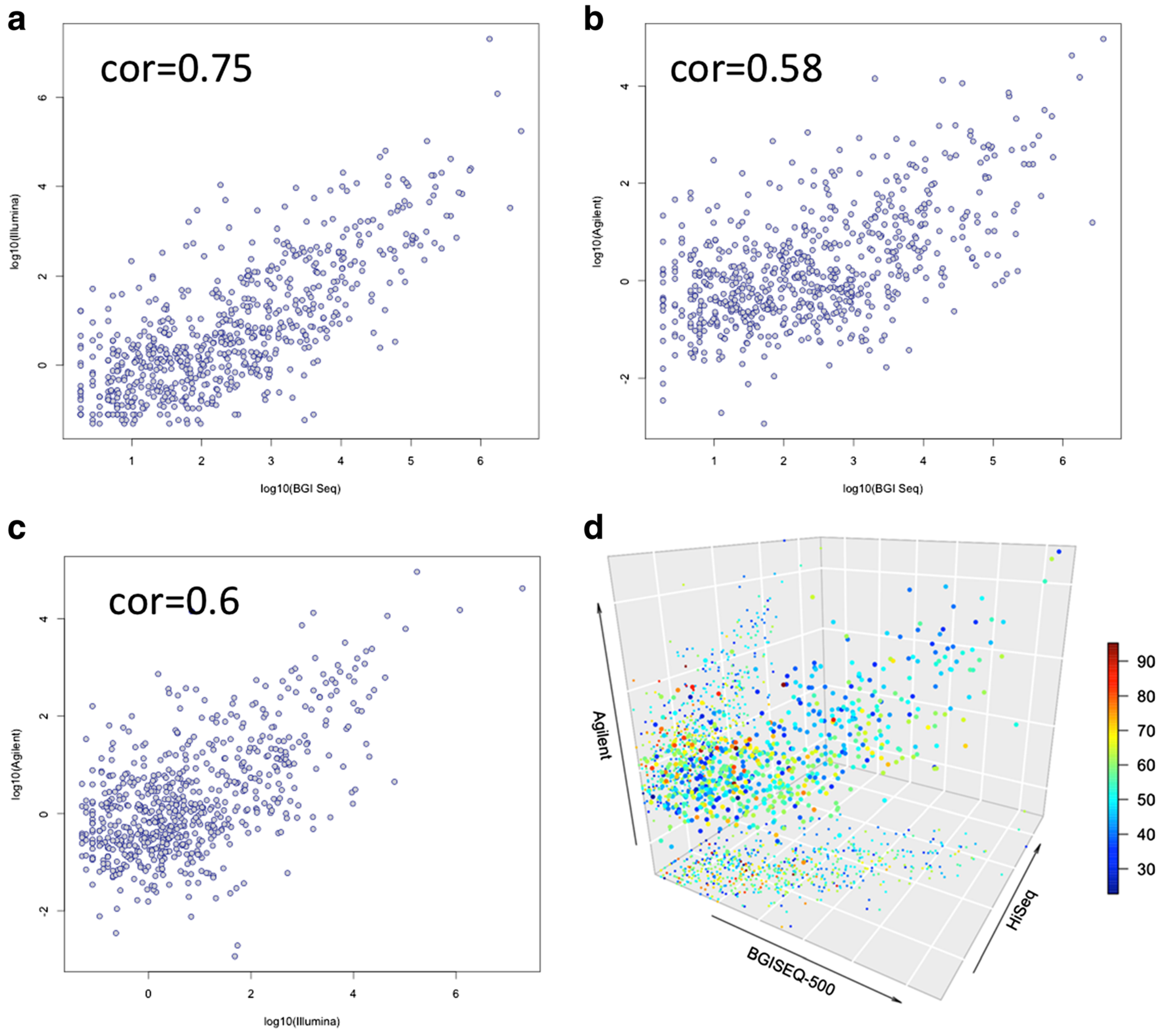


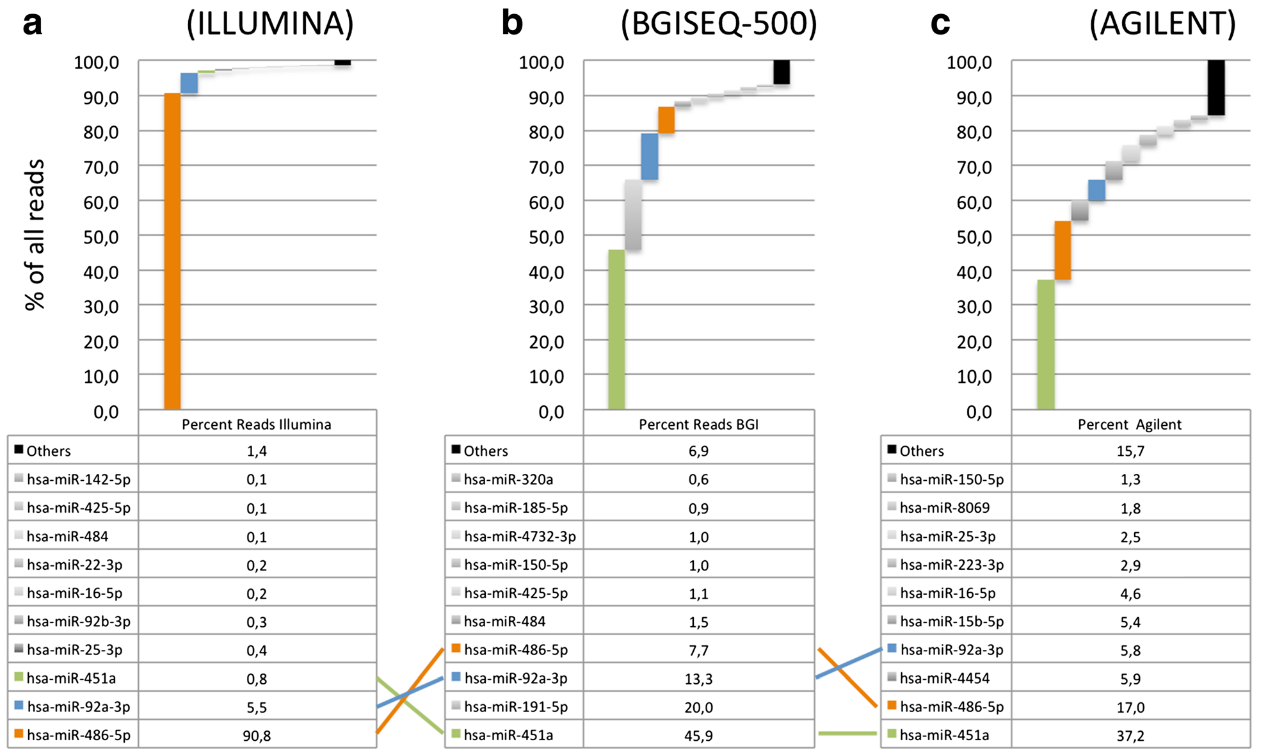
Fig. 3 **a**-**c** Pairwise scatter plots for comparing expression of miRNAs in blood cells on microarrays, HiSeq, and BGISEQ-500. Please note that for HighSeq and Agilent technical replicates were measured, for BGISEQ-500 biological replicates. **d** 3D scatter-plot colored by the GC content of miRNAs

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).



	Percent Reads Illumina
■ Others	1,4
▪ hsa-miR-142-5p	0,1
▪ hsa-miR-425-5p	0,1
▪ hsa-miR-484	0,1
▪ hsa-miR-22-3p	0,2
▪ hsa-miR-16-5p	0,2
▪ hsa-miR-92b-3p	0,3
▪ hsa-miR-25-3p	0,4
🟩 hsa-miR-451a	0,8
🟦 hsa-miR-92a-3p	5,5
🟧 hsa-miR-486-5p	90,8

	Percent Reads BGI
■ Others	6,9
▪ hsa-miR-320a	0,6
▪ hsa-miR-185-5p	0,9
▪ hsa-miR-4732-3p	1,0
▪ hsa-miR-150-5p	1,0
▪ hsa-miR-425-5p	1,1
▪ hsa-miR-484	1,5
🟧 hsa-miR-486-5p	7,7
🟦 hsa-miR-92a-3p	13,3
▪ hsa-miR-191-5p	20,0
🟩 hsa-miR-451a	45,9

	Percent Agilent
■ Others	15,7
▪ hsa-miR-150-5p	1,3
▪ hsa-miR-8069	1,8
▪ hsa-miR-25-3p	2,5
▪ hsa-miR-223-3p	2,9
▪ hsa-miR-16-5p	4,6
▪ hsa-miR-15b-5p	5,4
🟦 hsa-miR-92a-3p	5,8
▪ hsa-miR-4454	5,9
🟧 hsa-miR-486-5p	17,0
🟩 hsa-miR-451a	37,2

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

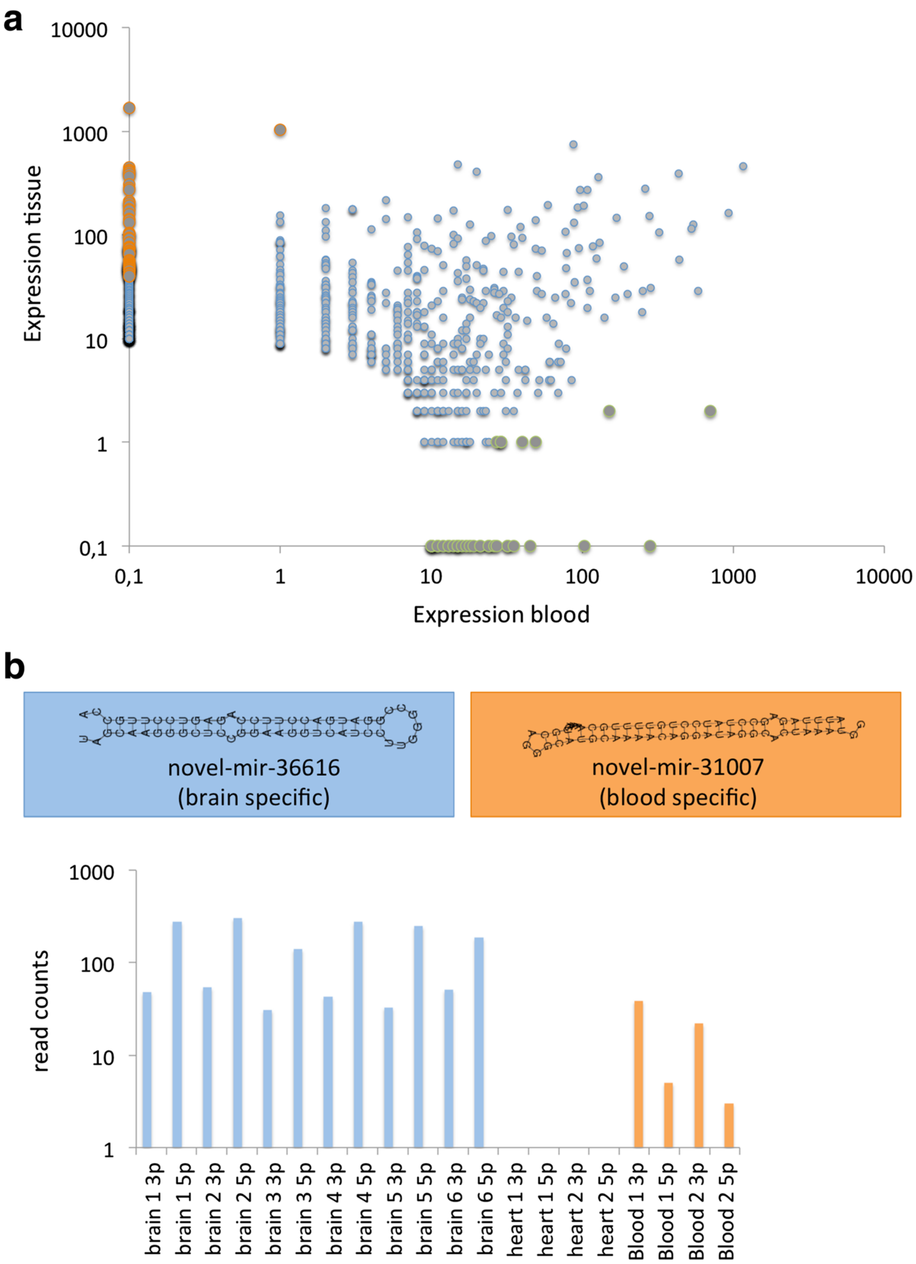


Fig. 5 a Expression of novel miRNAs in blood versus tissue. The *green* miRNAs are specific for blood, the *orange* miRNAs for tissue, and the *blue* miRNAs were detected in both specimens. **b** Bar plot for two different miRNAs. The first miRNA novel-mir-36616 in *blue* is detected only in the brain 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.

doi: 10.1093/nar/gkaa1122

CoolMPS: evaluation of antibody labeling based massively parallel non-coding RNA sequencing

Yongping Li[1,2,†], Tobias Fehlmann[1,†], Adam Borcherding[3], Snezana Drmanac[3], Sophie Liu[3], Laura Groeger[4], Chongjun Xu[2,3,5,6], Matthew Callow[3], Christian Villarosa[3], Alexander Jorjorian[3], Fabian Kern[1], Nadja Grammes[1], Eckart Meese[4], Hui Jiang[2], Radoje Drmanac[2,3,5,6], Nicole Ludwig[4,†] and Andreas Keller[1,7,*,†]

[1]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [2]MGI, BGI-Shenzhen, Shenzhen 518083, China, [3]Complete Genomics Incorporated, San Jose, CA 95134, USA, [4]Department of Human Genetics, Saarland University, 66421 Homburg, Germany, [5]BGI-Shenzhen, Shenzhen 518083, China, [6]China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China and [7]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, CA 94304, USA

Received May 15, 2020; Revised October 02, 2020; Editorial Decision November 02, 2020; Accepted November 26, 2020

ABSTRACT

Results of massive parallel sequencing-by-synthesis vary depending on the sequencing approach. CoolMPS™ is a new sequencing chemistry that incorporates bases by labeled antibodies. To evaluate the performance, we sequenced 240 human non-coding RNA samples (dementia patients and controls) with and without CoolMPS. The Q30 value as indicator of the per base sequencing quality increased from 91.8 to 94%. The higher quality was reached across the whole read length. Likewise, the percentage of reads mapping to the human genome increased from 84.9 to 86.2%. For both technologies, we computed similar distributions between different RNA classes (miRNA, piRNA, tRNA, snoRNA and yRNA) and within the classes. While standard sequencing-by-synthesis allowed to recover more annotated miRNAs, CoolMPS yielded more novel miRNAs. The correlation between the two methods was 0.97. Evaluating the diagnostic performance, we observed lower minimal *P*-values for CoolMPS (adjusted *P*-value of 0.0006 versus 0.0004) and larger effect sizes (Cohen's d of 0.878 versus 0.9). Validating 19 miRNAs resulted in a correlation of 0.852 between CoolMPS and reverse transcriptase-quantitative polymerase chain reaction. Comparison to data generated with Illumina technology confirmed a known shift in the overall RNA composition. With CoolMPS we evaluated a novel sequencing-by-synthesis technology showing high performance for the analysis of non-coding RNAs.

INTRODUCTION

Since the mid 1990′s, massively parallel sequencing approaches have been developed and continuously improved. The first commercial instruments were available on the market around 2005 (1). The rapid development of technology in the first 10 years had a substantial impact on genomic research (2), also leading to a continuous growth of data deposited in resources such as GenBank (3). While one of the most common applications is genome sequencing, RNAs are often analyzed using high-throughput sequencing as well. Even resolution at the single cell level can be reached now (4). A general overview of the different sequencing approaches together with available instruments highlights the diversity of available platforms and applications (5). Most recently, a comparison of Illumina NextSeq 500, NovaSeq 6000 and the BGI MGISEQ-2000 using identical single Cell 3′ libraries generated with the 10× Genomics Chromium platform highlighted comparable performance between the platforms in general (6).

For the high-throughput analyses of small non-coding RNAs (sncRNAs), sequencing has become one of the most frequently used methods (7). This has led to a very deep understanding of the sncRNA expression in humans (8,9) and many other species (10). As a consequence, databases on sncRNAs, especially on microRNAs (miRNAs) are updated regularly with increasing numbers of miRNAs. The miRBase in its most recent release 22 (October 2018 (11)) contains 38 589 entries from 271 species (12). Besides miRBase, MirGeneDB contains 10 899 curated miRNAs from

[*]To whom correspondence should be addressed. Tel: +49 681 302 68611; Fax: +49 681 302 68611; Email: andreas.keller@ccb.uni-saarland.de
[†]The authors wish it to be known that, in their opinion, the first two and last two authors should be regarded as Joint First Authors.

45 different organisms (13) and miRCarta (14) has the ambition to provide a collection of all expressed small RNAs. With 11 000 annual publications on miRNAs, these databases cover particular needs of researchers and provide an important source of information for future miRNA annotations (15). The largest fraction of miRNAs from high-throughput sequencing has been annotated for *Homo sapiens*. For example, as of August 2020, the miRMaster web service (16) has been applied in over 1300 studies. Sequencing data of more than 74 000 human sncRNA samples were evaluated and 1.1 trillion reads (1.1×10^{12}) have been processed using miRMaster. Notably, only a fraction of all available sncRNA sequencing data has been analyzed using the miRMaster tool, e.g. since only one organism is considered. Thus, the total number of sncRNA sequencing data sets exceeds the figures given above substantially. The gold standard sncRNA analysis software miRDeep/miRDeep2 (17,18) for example has been cited almost 2000 times. Constantly decreasing cost and broader availability of sequencing systems will lead to a continuously growing amount of sncRNA datasets in the future.

Many studies, however, indicate a severe influence of sample handling, library preparation and the sequencing technology on the read quantity, composition and quality (19–22). The most commonly applied approach is sequencing-by-synthesis using Illumina systems. These are available in combination with different library preparation approaches (19). We previously evaluated the performance of sequencing-by-synthesis on Illumina systems to combinatorial probe-anchor synthesis (cPAS)-based BGISEQ-500 sequencer (23). As compared to the Illumina system we found a larger variety of sncRNAs in the cPAS data, including twice as much yet unknown microRNAs at that time. Both sequencing approaches however rely on similar sequencing-by-synthesis principles, incorporating labeled nucleotides during each sequencing cycle.

The continuous development of library preparation and sequencing approaches is leading to novel commercially available systems and assay formats. The availability of a new experimental approach however immediately calls questions with respect to the validity of its data and the comparability. Especially for applications in biomarker development a platform change may significantly affect the diagnostic or prognostic performance of tests. Consequently, two questions come up whenever a new experimental approach is available: how does the performance change if technical replicates are compared between platforms and how does it affect biological results?

Recently, a fundamentally novel sequencing approach called CoolMPS has been introduced and made commercially available through MGI Tech Co., Ltd, Shenzhen, China (details are provided in the 'Materials and Methods' section). While it still relies on the sequencing-by-synthesis principle as other methods, no labeled nucleotides are incorporated. In order to measure a signal intensity representative for the incorporated base at each cycle, four specific antibodies, one recognizing each of the four natural bases (A, T, C, G) are used. The approach promises higher data quality by avoiding incorporation and detection interference of base-linked dyes and providing stronger signals by attaching multiple molecules of a dye per antibody. The CoolMPS approach for sequencing non-coding RNAs is described in the 'Materials and Methods' section. More details on the sequencing kits and basic biochemical principles of the methodology and its application are available with the user manual of the commercial kits and as preprint (https://doi.org/10.1101/2020.02.19.953307). It is mandatory to evaluate such new technologies with respect to common application scenarios. Discovering single nucleotide variants or small insertions and deletions pose different challenges as compared to, e.g. the quantification of RNAs in an at least pseudo-quantitative manner. In this study, we set to present the first detailed and direct performance comparison between the novel antibody-based labeling approach in comparison to standard sequencing-by-synthesis using labeled nucleotides for the quantification of small non-coding RNAs.

MATERIALS AND METHODS

RNA sample preparation and quality control

RNA from 2.5 ml whole blood collected in PAXgene tubes was isolated using the PAXgene Blood miRNA Kit (Qiagen, Hilden, Germany) according to the manufacturer's recommendations. RNA concentration and integrity were measured using Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and RNA 6000 Nano Kit for Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), respectively. RNA was aliquoted and used for the four experimental approaches CoolMPS, BGISEQ, Illumina and reverse transcriptase-quantitative polymerase chain reaction (RT-qPCR) as described below. The study was approved by the ethical committee of the Medical Faculty of the University of Tuebingen (Nr. 90/2009BO2). A list of samples included in the study is available as Supplementary Table S1.

CoolMPS™ on the DNBSEQ-G400RS

MiRNA libraries were prepared using the MGIEasy Small RNA Library Prep Kit (MGI Technologies, Shenzhen, China; product number 1000006383) with 800 ng total RNA input according to the manufacturer's recommendations. First, adapter sequences were ligated to the 3′ end of the RNA, followed by ligation of barcoded RT primers. Next, a universal adapter was ligated to the 5′ end. The RNA was then transcribed into cDNA by HiScript II Reverse Transcriptase in the presence of RNAse inhibitor. The primers used for the reverse transcription contained barcodes that allowed the pooling of up to 24 samples per sequencing library. Then cDNA libraries were amplified by 18-cycles of PCR reactions. Amplified PCR products were size selected using 6% TBE gel electrophoresis and the band from 100 to 120 bp was then purified with spin-X centrifuge tube filters followed by ethanol precipitation. The purified PCR products were quantified using Qubit dsDNA HS Assay kit (Invitrogen, Cat No. Q32854). Twelve purified PCR products were pooled with 84 fmol each (total 1 pmol) and circularized using a specific oligo sequence complementary to sequences in both the 3′ and 5′ adaptors provided in the MGIEasy Small RNA Library Prep Kit. The remaining linear DNA was digested.

After purification, the single strand circularized DNA library was quantified using Qubit ssDNA Assay Kit (Invitrogen, Cat No, Q10212). Subsequently, DNA nanoballs (DNBs) were generated using rolling circle amplification from 60 fmol of single stranded, circularized DNA library for 25 min. The DNB concentration was determined using Qubit ssDNA Assay Kit. The DNBs (concentration in the range of 8–20 ng/μl) were mixed with loading buffer by manual pipetting and subsequently loaded onto DNBSEQ-G400RS 4-lane flowcells (product number 1000016985) using the MGIDL-200H DNB loader as described in the CoolMPSTM High-throughput Sequencing Set User Manual provided with the kit. Loaded flow cells were sequenced on the DNBSEQ-G400RS instrument using CoolMPS^{TM} SE50 beta sequencing kits, now available as commercial products (product number 1000019478, MGI Tech Co., Ltd, Shenzhen, China) following manufacturers recommendation. The MGI CoolMPS^{TM} SE50 kits are the standard product for small RNA sequencing. Sequencing was performed by selecting the smallRNA sequencing plan from the application menu on the DNBSEQ-G400RS. Single end sequencing of 50 bp along with 10 bp of barcode was performed. The basic difference between CoolMPS and standard sequencing-by-synthesis, relying on incorporation of labeled nucleotides, is the incorporation of unlabeled, reversibly terminated nucleotides. The fluorescent signal to detect the incorporated bases is generated by using base-specific 3′ block-dependent fluorescently labeled antibodies. After each cycle, the bound antibodies are removed and 3′ blocking moiety on the sugar group of the nucleotide regenerates the natural nucleotides. This procedure has the advantage not leaving a mark on the base and making the current sequencing cycle independent on the previous one. Base calling and generation of FASTQ files on the DNBSEQ-G400RS was performed using the software release for CoolMPS (BasecallLite version_1.0.7.84). An important machine quality control step included the removal of tiles from the FASTQ files that failed at some point in the base calling process leading to 'N' bases for all reads in that respective tile. A detailed description of the CoolMPS method and procedures is available under: https://doi.org/10.1101/2020.02.19.953307. The sequencing has been performed by Complete Genomics Incorporated, San Jose, California. The overall process of library preparation and sequencing on the DNBSEQ-G400 is referred to as 'CoolMPS' through the whole manuscript.

BGISEQ-500 sequencing using standard cPAS

As described above for CoolMPS, the MGIEasy Small RNA Library Prep Kit (product number 1000006383) was used to generate circularized DNA libraries with 800 ng total RNA input according to the manufacturer's recommendations. The library preparation and DNB preparation procedures are exactly the same as the one described in the previous section. DNBs were loaded onto the flow cell using the BGIDL-50 DNB loader and single end 50 bp sequencing was performed using the BGISEQ-500RS High-throughput Sequencing Set SE50 on the BGISEQ-500RS instrument. The sequencing has been carried out in the Human Genetics Department at Saarland University, Germany. This process is referred to as 'BGISEQ' through the whole manuscript.

Illumina library preparation and sequencing

Libraries were prepared according to the protocol of the TruSeq Small RNA Sample Prep Kit (Illumina) with 200 ng of total RNA per sample as starting material as described previously (24). In brief, the concentration of the libraries was assessed using a Bioanalyzer with the DNA 1000 Chip. Before sequencing, libraries were pooled in equal amounts of batches of six samples and clustered with a concentration of 9 pmol in one lane each of a single read flow cell. Sequencing of 50 cycles was performed on a HiSeq instrument (Illumina). Demultiplexing of raw sequencing data and generation of FASTQ files was performed with CASAVA v1.8.2.

RT-qPCR

RT-qPCR experiments are described in detail in the original publication (25). In brief, the miScript PCR system was used with custom miRNA PCR arrays (all reagents from Qiagen, Hilden, Germany). The PCR arrays were designed in 96-well plates to measure the expression of human miRNAs and RNU48 as well as RNU6 as endogenous controls. The RT-qPCR experiments have been performed in the Human Genetics Lab of Saarland University. Reverse transcription was performed using 100 ng total RNA as input using miScriptRT-II kit in 20 μl total volume. PCR reactions with 1 ng cDNA input in a total volume of 20 μl were set up automatically using the miScript SYBR Green PCR system in a Qiagility pipetting robot (Qiagen, Hilden, Germany).

Bioinformatics

The pre-processing of the FASTQ files of CoolMPS, BGISEQ and Illumina has been done using miRMaster 1.1 (16,26). MiRMaster is freely accessible at https://www.ccb.uni-saarland.de/mirmaster/. Briefly, adapters at the 3′ end were trimmed, while allowing an error of maximum one base and requiring a minimum overlap with the read of 10 bases. Reads were quality trimmed when the average quality dropped below 20 in a window of four consecutive bases to ensure a high quality of reads used for the downstream processing. All reads shorter than 17 bases after trimming were discarded from all further analyses. Read duplication levels were computed with FASTQC 0.11.8. The error rate per base was estimated by mapping the trimmed reads to the human genome with bowtie, while allowing up to three mismatches (command line: bowtie -v3 -k 1 –best –fullref) and counting the mismatched bases with Samtools stats (version 1.9, (27)). To further ensure the best comparability, BGISEQ and Illumina data were subsampled to match the CoolMPS distribution that was originally sequenced to a lower extent. In detail, all samples were subsampled to a read depth of 10 Million reads. Reads were mapped to the primary assembly of GRCh38.p10 using bowtie 1.2.2 (28), while allowing no mismatches and discarding reads mapping to over 100 locations (command line: bowtie -v0 -m 100 –best –strata –fullref). Read RNA classes were determined using FeatureCounts 1.5.2 (29) and annotations of

GENCODE v25 (30), piRBase 1 (31), miRBase v22.1 and GtRNAdb 18.1 (32) with the following parameters: -F SAF –O –M –R –f –fracOverlap 0.9, which required an overlap of at least 90% of a read with the annotated region and allowed multimapping reads and overlapping features. MiRBase v22.1 miRNAs were quantified using miRMaster with up to one mismatch and a variability of two bases allowed at the 5′ end and five bases at the 3′ end. Novel miRNA candidates were predicted with miRMaster with a required minimum expression of five reads in at least 75% of all dementia or control samples. Since we expect numerous false positive hits from the next generation sequencing data we performed a quality control of the newly predicted candidates and evaluated them using the NovoMiRank tool (33). NovoMiRank was applied using the default parameters, i.e. miRBase versions 1–7 were used as reference to identify the most reliable candidates. All further downstream analyses have been carried out in R 3.6.1 (https://www.R-project.org/). To test whether miRNAs were normally distributed, Shapiro–Wilk tests were computed per miRNA using the shapiro.test function from the stats package. As hypothesis test, parametric *t*-test and non-parametric Wilcoxon Mann-Whitney (WMW) test were performed using the t.test and wilcox.test functions from the stats package. Statistical tests for group comparisons were carried out as two-tailed and un-paired tests. All *P*-values were subjected to adjustment for multiple testing by using the Benjamini–Hochberg approach through applying the p.adjust function from the stats package. To estimate the effect sizes, the area under the receiver characteristic curve (AUC value) and the Cohen's D effect size were computed using the R pROC package (1.15.0, (34)) and the R effsize package (0.7.4). Plots were generated with ggplot2 (3.1.0), cowplot (0.9.4), complexHeatmap (2.5.3, (35)), ggridges (0.5.1) and vioplot (0.3.5). To compute the most significant overlap between the CoolMPS and BGISEQ technology in terms of dementia biomarkers we employed the dynamic programming based DynaVenn approach (36). DynaVenn is freely accessible at https://www.ccb.uni-saarland.de/dynavenn. Functional categories were analyzed by miRNA set enrichment analysis with default parameters using miEAA 2.0 (37,38) with a list of the miRNAs sorted with respect to their effect sizes as input (with separate adjustment of categories and Benjamini–Hochberg adjustment procedure).

RESULTS

Study setup allowing to evaluate technical and biological aspects

Primary aim of the study was to compare the combinatorial probe-anchor synthesis (cPAS)-based data using labeling of nucleotides to the data generated by the new antibody labeled-based approach on the more recent DNBSEQ-400RS systems. In the context of this manuscript, the former approach is referred to as BGISEQ and the latter as CoolMPS. Secondary aim was to compare the performance and comparability of both approaches in terms of potential liquid biopsy biomarker tests. We thus selected a study setup that allows to address both aims (Figure 1A). We sequenced 240 individual blood samples on both sequencing systems. The 240 samples include 179 controls and 38 patients with dementia. This part of the cohort has been used to evaluate the performance of both technologies to detect dementia biomarkers. Furthermore, the 240 samples include 17 individuals and 6 technical replicates. The latter samples were not used for the biomarker study but to assess the general stability and reproducibility of the technologies. Further, we compared the data to RT-qPCR measurements of a subset of 19 miRNAs in 189 samples and also evaluated the performance in comparison to data generated by Illumina sequencers. A full list of miRNAs and samples together with the respective Delta CT values from the RT-qPCR validation is available in Supplementary Table S2. We first evaluated the general performance of CoolMPS for quantification of RNA and then provide results of CoolMPS as liquid biopsy biomarker for dementia. The cohort was composed of participants with an average age of 67.3 years and a standard deviation of 12.3 years (Figure 1B). Details on the sequencing approaches and data analyses are given in the 'Materials and Methods' section.

Key performance indicators reveal improved quality of CoolMPS

First, we compared the Q30 values for the reads from the two sequencing approaches (Figure 1C). The Q30 value provides the percentage of bases sequenced with a Phred score of at least 30, corresponding to an error rate of 0.1%. The median Q30 of the BGISEQ was 91.8% while the median Q30 of CoolMPS jumped to 94%, representing a significant improved performance of CoolMPS ($P < 10^{-10}$). Intriguingly, we observed the higher per base sequencing accuracy over the complete read length not observing any drop at the beginning or at the end of the read. Moreover, CoolMPS showed lower variability in sequencing performance over the read in general as well as lower variability per base in the read (Figure 1D). While the variation of valid reads per sequencing run still varied for the CoolMPS technology we observed a constantly higher fraction of reads mapping without mismatches to the human genome (84.9% for BGISEQ and 86% for CoolMPS; Figure 1E). We also investigated the GC content of the generated libraries and found a median of 51.10% for BGISEQ and a median of 50.72% for CoolMPS in the unprocessed data, which dropped to a median of 42.38 and 41.60% for BGISEQ and CoolMPS after adapter and quality trimming, respectively (Supplementary Figure S1A and B). The mean quality scores per position varied between 33.95 and 36.35 for CoolMPS and even increased slightly toward the end of the read. In contrast, the BGISEQ reads varied between 27.95 and 36.17 and reached their peak at position 26. Then, the quality of BGISEQ reads decreased until position 50 (Supplementary Figure S1C and D). The mean quality scores for the trimmed files, i.e. those that did not contain any adapters, varied similarly, although the mean quality scores decreased more for longer reads. The estimated error rate was for both technologies similar with a median of 0.74% for BGISEQ and 0.76% for CoolMPS (Supplementary Figure S1E). For both, the raw sequencing files, and the trimmed ones, we observed a close to identical GC content distribution. For both technologies we observed two distinct peaks at 51 and 57% (Supplementary Figure S1F and G). We also found that the

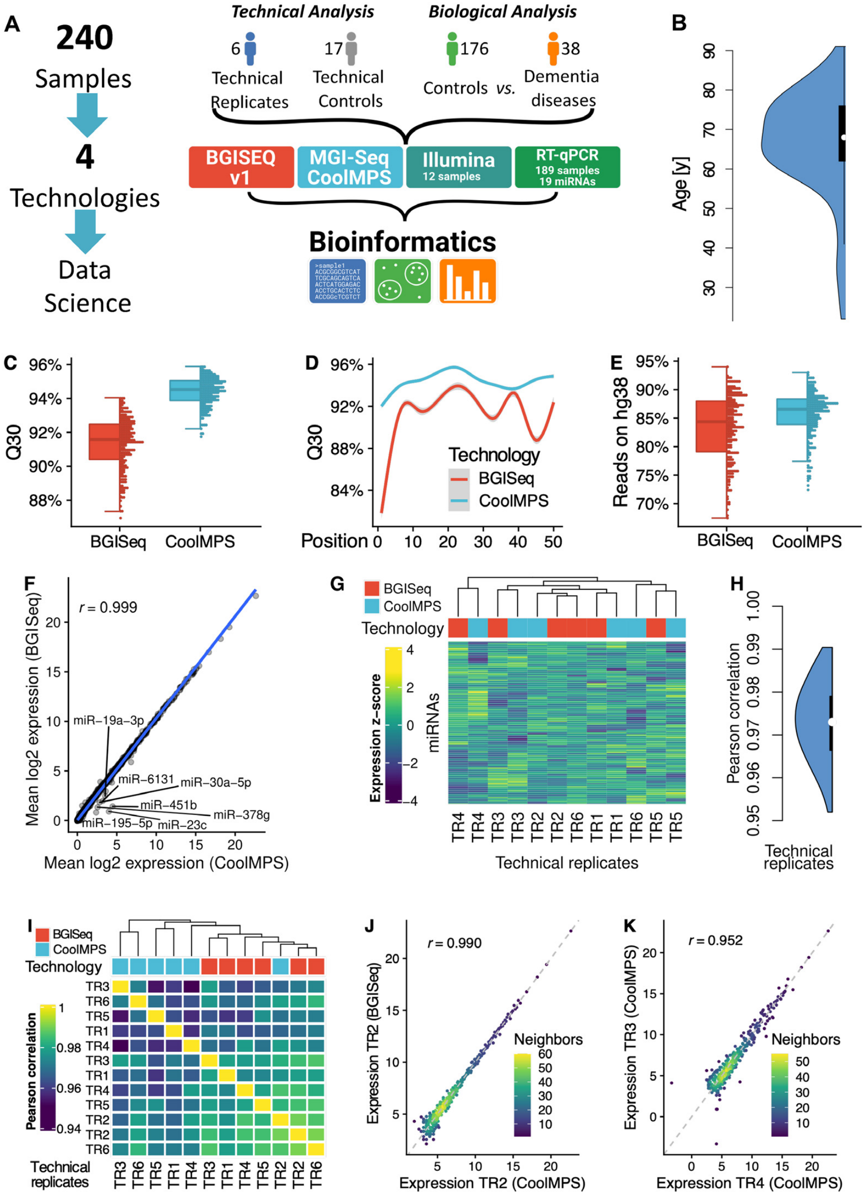


Figure 1. Study setup and quality control. (**A**) In the study we measured 240 individual blood samples using two fundamentally different sequencing approaches and compare the data by bioinformatics approaches before we compute the concordance to RT-qPCR profiles. The 240 samples include one part that has been used only for assessment of technical properties (6 and 17 samples in blue and gray) as well as a second part to evaluate performance related to biomarker discovery (176 controls in green and 38 dementia cases in orange). (**B**) Distribution of the age of the individuals included in the study, shown as violin plot. The black box spans the first to the third quartile and the white dot shows the median. (**C**) Distribution of the average Q30 value per sample for the two technologies, shown as boxplot (left) and dotplot (right). Each sample is shown as one dot. The boxes span the first to the third quartile with the horizontal line inside the box representing the median value. The whiskers show the minimum and maximum values or values up to 1.5 times the interquartile range below or above the first or third quartile if outliers are present. (**D**) Q30 value over all samples per technology as function of the position in the read. The smoothed curve is fitted by a generalized additive model using a cubic regression spline. The gray area represents the confidence interval of the fit. (**E**) Distribution of the percentage of reads mapping to the human reference genome hg38 without mismatch per technology, shown as boxplot (left) and dotplot (right). Each sample is shown as one dot. The boxes span the first to the third quartile with the horizontal line inside the box representing the median value. The whiskers show the minimum and maximum values or values up to 1.5 times the interquartile range below or above the first or third quartile if outliers are present. (**F**) Scatter plot of the average expression of all miRNAs in all samples for the two technologies. The blue line is the regression line. The Pearson correlation is shown in the upper left part of the plot. MiRNAs with a fold change larger than two between both technologies are highlighted. (**G**) Heat map of the clustered expression *z*-scores of miRNAs (rows) and technical replicates (columns). The color code for the columns represents the technology. The dendrogram shows the hierarchical clustering of the samples with Euclidean distance and complete linkage. (**H**) Distribution of all 12*11/2 = 66 pairwise Pearson correlation coefficients, shown as violin plot. The black box spans the first to the third quartile and the white dot shows the median. (**I**) Correlation matrix of the expression values of all miRNAs for all technical replicates. The dendrogram shows the hierarchical clustering of the samples with Euclidean distance and complete linkage. (**J**) Scatter plot of miRNAs for the best correlation between two technical replicates. The dotted line represents the angle bisector. The Pearson correlation is shown in the upper left part of the plot. The points are colored according to the point density in their neighborhood. (**K**) Scatter plot of miRNAs for the worst correlation between two technical replicates. The dotted line represents the angle bisector. The Pearson correlation is shown in the upper left part of the plot. The points are colored according to the point density in their neighborhood.

read length in both libraries after trimming peaked at 22, as we expected from a miRNA enriched library (Supplementary Figure S1H). We further evaluated the duplication levels of the CoolMPS and BGISEQ libraries. In both cases, the distributions were again nearly identical, showing most duplication levels above 10 000 (Supplementary Figure S1I and J). This is expected from miRNA libraries, as often a small number of miRNAs account for most of the reads. Finally, we checked the read base composition and found similar patterns. The first 22 bases reveal the most overrepresented sequence (i.e. the sequence of hsa-miR-451a), followed by the bases of the adapter sequence for the raw reads, and by less sequence specific bases for the trimmed reads (Supplementary Figure S1K and L). For most of the tested relevant key performance indicators (e.g. Q30 and reads mapping to the human genome) that allow to compare the general sequencing performance, CoolMPS yielded an increased performance compared to the classical BGISEQ approach.

Next, we evaluated and compared the reproducibility of the two technologies. When comparing the mean expression of all samples for CoolMPS to BGISEQ we obtained an extremely high correlation of 0.999 (Figure 1F). The scatter plot highlights a set of seven miRNAs, which were measured with higher expression in the CoolMPS data as compared to BGISEQ (miR-19a-3p, miR-30a-5p, miR-6131, miR-451b, miR-378g, miR-195-5p and miR-23c). Next, we considered only the six technical replicates per technology. There, these miRNAs reveal the same pattern as for the complete set of samples, thus excluding variance related to the disease status of the participants as potential cause (Supplementary Figure S2). Sequence and structure properties of these miRNAs are shown in Supplementary Table S3. Neither the length, nor the base composition or secondary structures reveal a joint pattern, arguing against a technological bias. We then asked whether we observe a clustering according to the sequencing approach or whether CoolMPS and BGISEQ samples mix. Indeed, hierarchical clustering indicates that the samples do not cluster by technology (Figure *1G*). The Pearson correlation between all $12 \times 11 / 2 = 66$ pair wise comparisons of technical replicates varied between 0.952 and 0.990 with a median performance of 0.973 (Figure 1H). The correlation matrix revealed marginal differences in the correlation coefficients between all the BGISEQ replicates (median 0.980) in comparison to the ones between the CoolMPS samples (median 0.964) (Figure 1I). Also, the correlation between the two technologies with a coefficient of 0.973 was high. The differences in the correlation lead to a tendency of technologies to cluster together, although CoolMPS Technical Replicate 2 clustered with BGISEQ Technical Replicates 2 and 6. Scatter plots for the best (Figure 1J) and the worst correlation (Figure 1K) demonstrate the generally very high reproducibility between the technologies that is in the same range as technical replicates within the technologies. Most importantly, we did not observe any significant change between the RNAs profiled with BGISEQ compared to CoolMPS after adjustment for multiple testing, both, for the WMW and the *t*-test.

Having understood basic performance of the sequencing technology as well as core aspects on technological reproducibility we next evaluated the content of the different sequencing approaches with respect to quantitative and qualitative aspects.

Composition of different RNA classes is similar between BGISEQ and CoolMPS

The first question related to small non-coding RNA sequencing data is the representation of different RNA classes. Different sample- and library preparation protocols lead to varying results. For example, size selection is applied to enrich-specific populations of sncRNAs. To minimize respective effects and to focus on the performance of the sequencing technique, we used the same libraries for sequencing and purified small non-coding RNAs by gel electrophoresis (see 'Materials and Methods' section). This protocol has been optimized to enrich for miRNAs, however, leaving also reads to evaluate other RNA classes. The distribution to the different classes matched generally very well between BGISEQ and CoolMPS (Figure 2A). Especially, we observed the intended enrichment for miRNAs. For BGISEQ, 91.7% of all mappable reads matched to miRNAs, for CoolMPS we even reached a higher mapping of 92.7%. The second most abundant RNA class was the Ensembl's misc RNA category, containing among others yRNAs and signal recognition particle RNAs (SRP RNAs). This category contains 5.1% of all BGISEQ and 4.5% of all CoolMPS reads. All other categories were covered by less than 1% of reads in both technologies. The scatter plot contrasting the $\log_{10}$ percentages for both technologies highlights the very reproducible distribution of reads to the different RNA classes (Figure 2B). Since the protocol was optimized to enrich for miRNAs and our results demonstrate that this enrichment was successful, we focused on comparing the performance for this class of sncRNAs.

CoolMPS yields more novel miRNA candidates

With respect to different technologies a bias in sncRNA-seq data is known. Especially for specimen types such as whole blood where already an enrichment of selected miRNAs exist, additional technological bias can further impair the data analysis. In whole blood, miRNA expression is not uniformly distributed but few miRNAs are significantly higher expressed than others. Technology bias further overamplifies the respective miRNA reads. These circumstances complicate the discovery of new miRNAs with the aim of completing the repertoire of annotated miRNAs (8). We thus evaluated and compared the distribution of reads to different miRNAs using the two sequencing technologies and asked how many novel miRNA candidates could be obtained. As expected, we observed an uneven distribution, which is however highly concordant between the technologies (Figure 3A). At the same time, we discovered 124 novel miRNA candidates using BGISEQ while CoolMPS based results highlight 134 novel miRNA candidates (Figure 3B and Supplementary Figure S3A). These findings suggest a higher sensitivity in terms of discovering low abundant yet unknown miRNA molecules. Remarkably, a large fraction of all new microRNA candidates, in total 88, have been detected by both technologies. To assess the quality of those

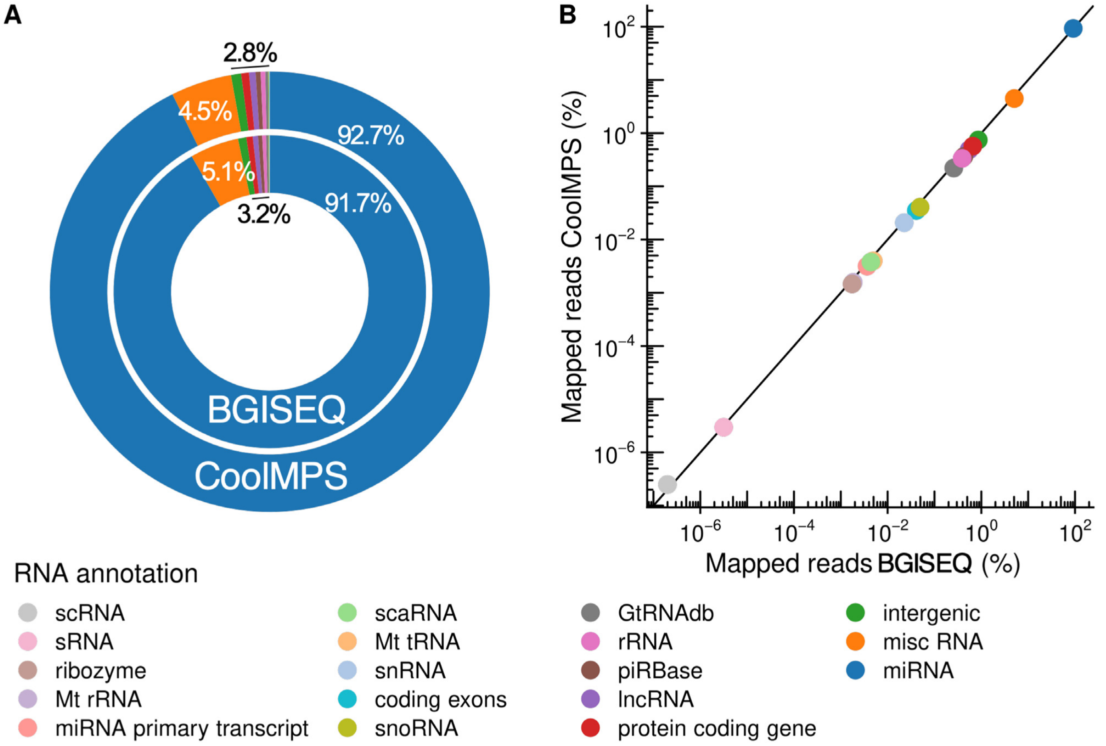


Figure 2. Distribution to the different sncRNAs classes. (**A**) Donut plot comparing the distribution of all RNA classes and intergenic regions that were covered by reads from CoolMPS and BGISEQ. (**B**) Scatter plot that shows the percentage of reads mapping to the RNA classes and intergenic regions for BGISEQ (*x*-axis) and CoolMPS (*y*-axis) on a logarithmic scale.

miRNA candidates we scored them using NovoMiRank. The score computed by NovoMiRank considers sequence and structural features and describes the average distance of the new candidates to a reference set, which is per default miRBase v1-7. The median score obtained for the common candidates was 1.12, while the technology specific candidates obtained median scores of 1.05 for CoolMPS and 1.00 for BGISEQ. As highlighted by the distribution shown in Supplementary Figure S3B, the score ranges are similar between the approaches and only few candidates (four detected by both CoolMPS and BGISEQ, three CoolMPS specific and one for BGISEQ specific) showed scores above 1.5. The score of 1.5 has been set since it is the maximum score observed for miRBase v1-7 miRNAs i.e. the reference set of NovoMiRank. In summary, both technologies do not reveal quantitative differences in the quality of reported miRNAs but only in the quantity, with remarkable advantages of CoolMPS.

In comparing the distribution of miRNAs annotated in the miRBase we observe 76.7% of all BGISEQ reads mapping to the most abundant miRNA (miR-451a; Figure 3C). Using CoolMPS, 78.5% of all reads matched to this miRNA (Figure 3D). The second most abundant miRNA is represented by 9 and 8.4% of all reads, respectively (miR-92a-3p). In sum, the top five miRNAs are covered by 93.8% of all reads in the BGISEQ and by 92.6% of all reads in the CoolMPS approach. A more detailed breakdown by excluding the most abundant miR-451a demonstrates that the order of the 10 most abundant miRNAs matches perfectly between the two technologies (Figure 3E and F). At the same time, the data reinforces that especially for biospecimens with an uneven distribution of miRNA molecules, deep sequencing with the least possible bias is required to profile known and to discover new miRNAs.

Comparing biomarker profiles shows high reproducibility between the different approaches

One of the most important question in introducing new technologies is not only whether general performance improves but also whether previous biological results can be reproduced. One core example are biomarker tests. Often, biomarker sets change substantially when a new quantification approach is introduced. This might be an expected and even desired result, e.g. if a new technology generation with higher technical sensitivity is introduced. But if a new technology has the main task to support translation of biomarkers to care by facilitating better integration into clinical workflows or lower experimental costs, original biomarker profiles should not be compromised. We thus evaluated the diagnostic performance of miRNA biomarkers using BGISEQ and CoolMPS and used a liquid biopsy dementia test as validation example. We sequenced cases with dementia as well as controls with similar age distribution (Figure 1A and B). As performance criteria we considered the result of two commonly used hypothesis tests, the *t*-test and the WMW test. Since not all miRNAs were normally distributed according to the Shapiro Wilk test, we here focus on the results of the WMW test and provide the t-test *P*-values only in the supplement (Supplemen-

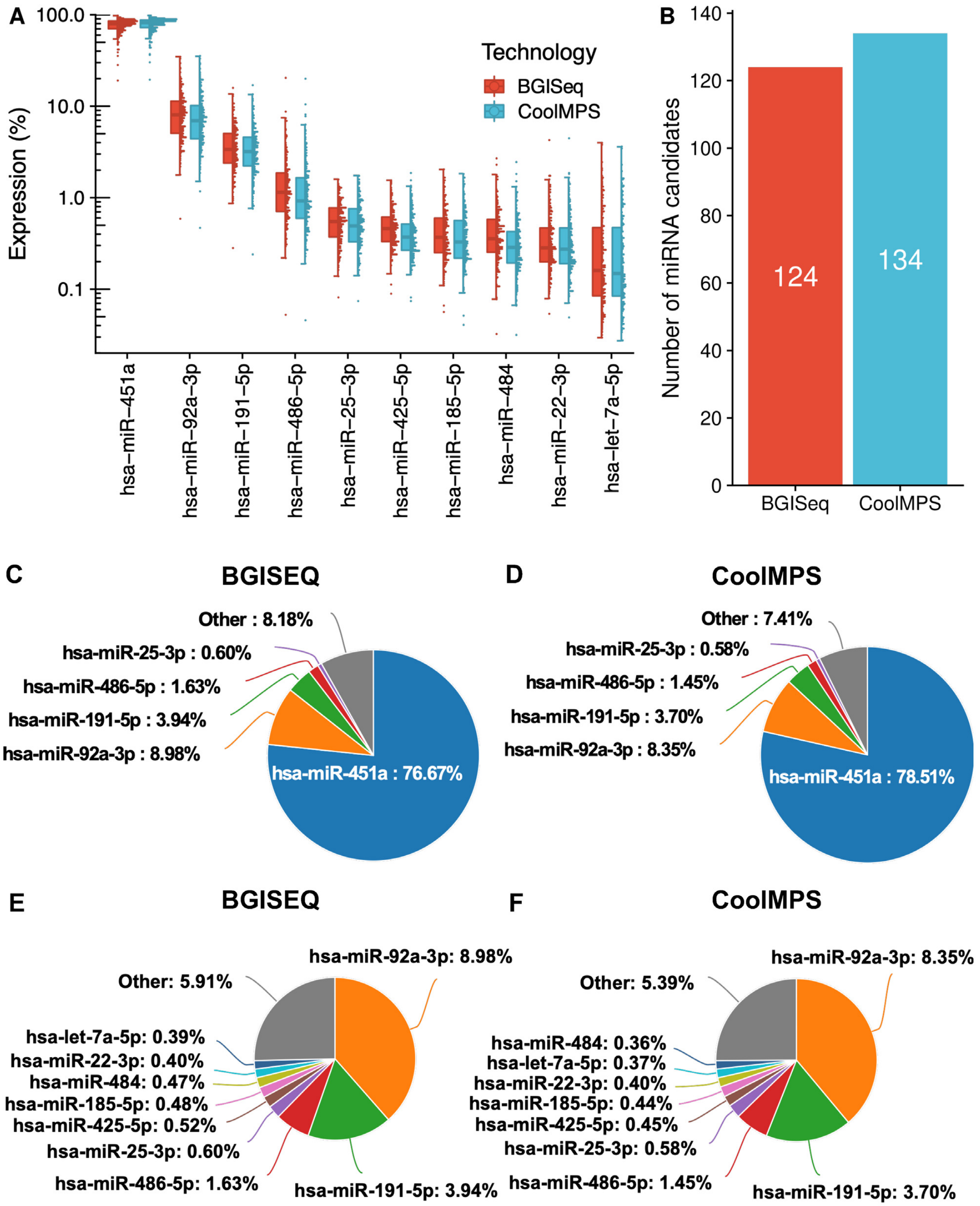


Figure 3. Distribution to microRNAs. (**A**) Distribution of the read percentage of the 10 most abundant miRNAs in the CoolMPS and BGISEQ data, shown as boxplot (left) and dotplot (right). Each sample is shown as one dot. The boxes span the first to the third quartile with the horizontal line inside the box representing the median value. The whiskers show the minimum and maximum values or values up to 1.5 times the interquartile range below or above the first or third quartile if outliers are present. (**B**) Number of novel microRNA candidates for both technologies. (**C**) Pie chart for the top five miRNAs on the BGISEQ. (**D**) Pie chart for the top five miRNAs on the CoolMPS. (**E**) Pie chart for the top ten miRNAs on the BGISEQ after exclusion of the most abundant miR-451a. (**F**) Pie chart for the top ten miRNAs using CoolMPS after exclusion of the most abundant miR-451a.

tary Table S4 and 5). Because of known challenges with *P*-values and the controversial discussion on this topic (39), we also computed effect sizes, namely Cohen's D and the area under the receiver characteristics curve AUC. Detailed results for each miRNA and each of the different metrics are provided for both BGISEQ (Supplementary Table S4) and CoolMPS (Supplementary Table S5). In terms of AUC, BGISEQ and CoolMPS showed an almost identical distribution (Figure 4A). The scatter plot displays a very high degree of reproducibility (Pearson correlation coefficient of 0.905) between the two technologies considering the diagnostic performance (Figure 4B). As consequence, also the volcano plots for the two technologies were very similar (Figure 4C and D). Given the general concordance of

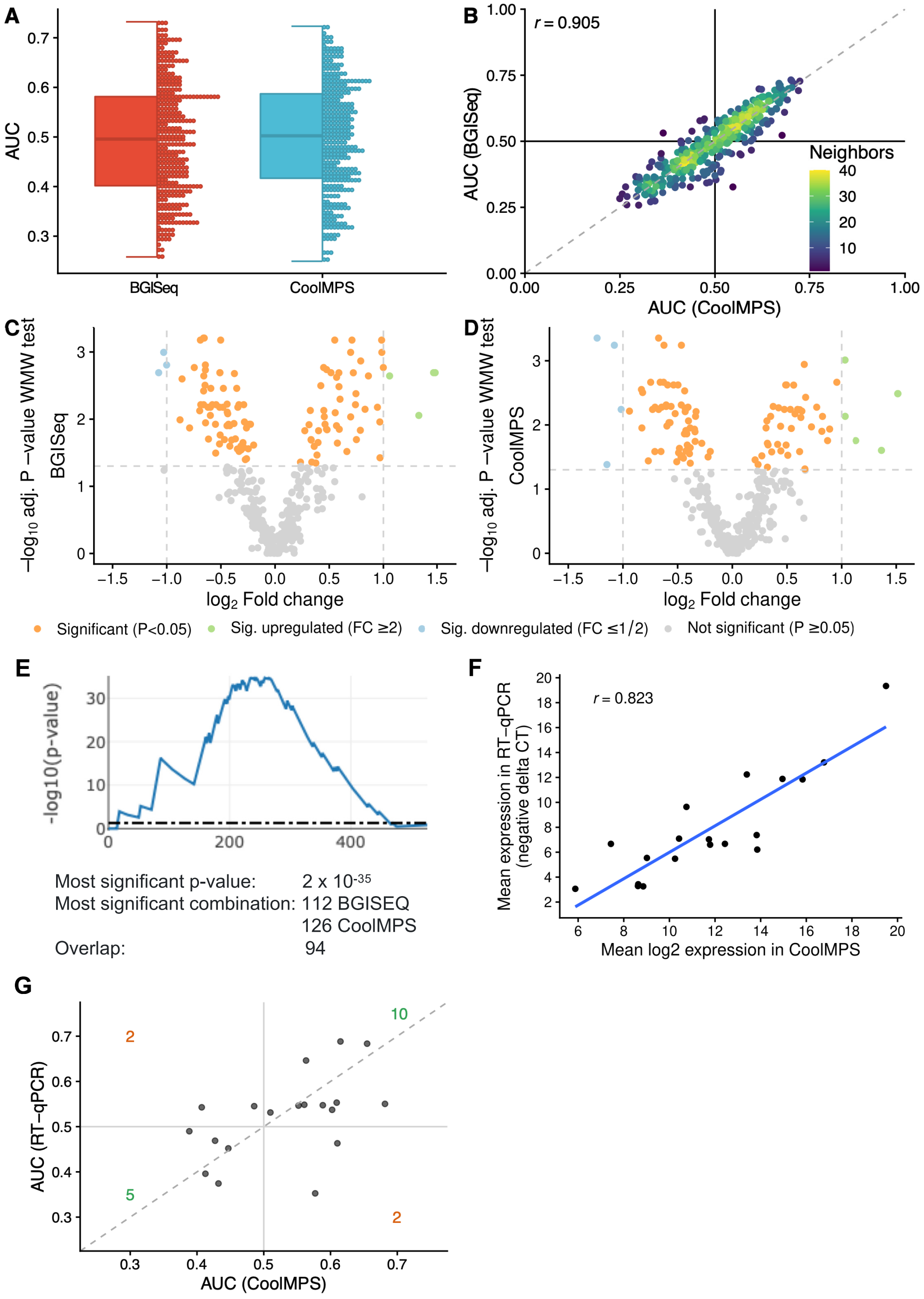


Figure 4. Diagnostic performance on dementia patients. (**A**) Distribution of the AUC values to differentiate between dementia and controls obtained for both technologies. An AUC of 0.5 means no dys-regulation. A deviation from 0.5 toward one means an upregulation and toward zero a downregulation of the biomarkers. The distribution is shown as boxplot (left) and dotplot (right). Each miRNA is shown as one dot. The boxes span the first to the third quartile with the horizontal line inside the box representing the median value. The whiskers show the minimum and maximum values or values up to 1.5 times the interquartile range below or above the first or third quartile if outliers are present. (**B**) Scatter plot of the AUC values to differentiate between dementia and controls in CoolMPS (*x*-axis) versus BGISEQ (*y*-axis). The black horizontal and vertical line represent the AUC value of 0.5, respectively. The Pearson correlation is shown in the upper left part of the plot. The points are colored according to the point density in their neighborhood. (**C**) Volcano plot showing the $\log_2$ fold change on the x-axis and the FDR adjusted negative $\log_{10}$ of the Wilcoxon–Mann–Whitney (WMW) *P*-value on the *y*-axis for BGISEQ. Orange dots are located above the horizontal line and are significant. Blue and green dots above the horizontal and on the left / right of the vertical lines are significant and have a fold-change above 2. (**D**) Same volcano plot as in Figure 4C, but for CoolMPS. (**E**) Result of DynaVenn that presents the negative $\log_{10}$ of the overlap between the two miRNA sets dependent on how many miRNAs are included. The peak of the curve represents the most significant overlap. (**F**) Scatter plot of the $\log_2$ CoolMPS expression (*x*-axis) and the negative delta CT value for the 19 miRNAs included in the validation study. The Pearson correlation coefficient is shown in the upper left part of the plot. (**G**) Scatter plot of the AUC values to differentiate between dementia and controls in CoolMPS (*x*-axis) and BGISEQ (*y*-axis). The dashed line represents the angle bisector.

the results we speculated that also the ranks of biomarkers were consistent between the two technologies. For the top-10 markers of BGISEQ and CoolMPS we thus compared the ranks and absolute values (Table 1). First, we recognized that the top marker performed better in CoolMPS as compared to BGISEQ in all metrics, the raw *P*-value, the adjusted *P*-value, the Cohen's D and the AUC. The adjusted *P*-values were for example 0.0006 in BGISEQ data and 0.0004 in CoolMPS data. Second, we observed that four miRNAs were among the top 10 markers in both technologies (miR-3200-3p, let-7e-5p, miR-15b-5p, miR-19b-3p). For other markers we computed partially very different ranks. One of the most extreme examples is miR-3335-5p, which is ranked 9th most significant in CoolMPS and 117th in BGISEQ. Nonetheless, this miRNA was significant in both approaches. On the one hand we observed a very high correlation, on the other hand, we also noticed substantial differences in the ranks, most likely related to the close range of the *P*-values, challenging the concept of fixed thresholds. To overcome the bias of selecting fixed rank ranges, we developed the DynaVenn approach that computes the most significant overlap between two biomarker sets containing technical or biological replicates. DynaVenn computed the best overlap in selecting the best 112 miRNAs from BGISEQ and the best 126 miRNAs from CoolMPS, yielding an overlap of 94 miRNAs and a *P*-value of 2×10^{-35} (Figure 4E). Thus, the two biomarker sets show a highly significant overlap which might have remained hidden if only the top 10 markers would have been considered.

Illumina sequencing data shows differently biased but comparable measurements

In addition to BGISEQ we also compared the performance of CoolMPS to standard Illumina sequencing for small non-coding RNAs on a subset of 12 samples (24). As part of our quality control we filter reads shorter than 17 nucleotides. We thus compared the fraction of filtered reads for the three technologies on the subset of samples sequenced by the three technologies. For BGISEQ, 2.76% (SD of 0.56) of reads, for CoolMPS 3.90% (SD of 0.54) of reads and for Illumina 3.95% (SD of 3.24%) of reads were excluded. In a first analysis step we evaluated the Q30 values obtained by both approaches and found a median Q30 of 94.95% for Illumina in comparison to 93.10% for CoolMPS (Supplementary Figure S4A). The quality profile revealed Q30 values going up to a median of 99.43% for Illumina in the first 20 positions, whereas a strong drop could be observed afterwards, going down to a Q30 of 87.33% at position 50 (Supplementary Figure S4B). In comparison, the CoolMPS quality remained more stable for the complete read length with an average Q30 of 93.12% (SD: 1.79%) and even showed an increased quality toward the end of the reads. For the fraction of reads that can be used in further analyses, i.e. the ones mapping to the human genome, we observed for CoolMPS a median of 90.74%, while for Illumina only 77.85% could be mapped (Supplementary Figure S4C). In the next step, we inspected the expression similarity of both technologies and found a general agreement of both with a Pearson correlation of 0.873 (Supplementary Figure S4D). Nevertheless, we could observe 52 miRNAs with expression values differing by fold changes above 10, showing the technological specific biases (e.g. hsa-miR-486-5p was expressed 76 times higher in the Illumina samples). In addition, we confirmed that the samples of both technologies clustered separately according to their miRNA profiles (Supplementary Figure S4E) and showed a much higher intra-technology expression correlation (Pearson correlation of 0.960 for CoolMPS on median, 0.955 for Illumina) than between technologies (median Pearson correlation of 0.742) (Supplementary Figure S4F and G). We then asked if the RNA class distribution between both technologies show similar patterns. We found that the CoolMPS samples showed a higher diversity of RNA classes, whereas the Illumina samples contained a higher percentage of reads mapping to piRNAs (0.70 versus 0.17% in CoolMPS) and miRNAs (97.76 versus 94.98% in CoolMPS) (Supplementary Figure S5). Next, we focused on the composition of the detected miRNAs and found that of the top 10 most expressed miRNAs of both technologies, six overlapped. The largest differences could be observed for hsa-miR-486-5p and hsa-miR-451a, which are both the most expressed miRNAs in Illumina and CoolMPS and differ by a fold change of 76 and 87, respectively (Supplementary Figure S6A). For the Illumina samples, thus only 9.48% of the reads could be mapped to other miRNAs and after excluding the top 5 miRNAs, only 3.15% of the reads mapped to others (Supplementary Figure S6B). For the CoolMPS samples, we observed slightly increased mapping rates to the top five miRNAs on this subset of samples, with 5.60% of the reads mapping to the other miRNAs (Supplementary Figure S6C). Supplementary Figure S6D and E show a detailed breakdown of the top expressed miRNAs, after excluding the most abundant one. We also found that some miRNAs that were detected with low abundance in one technology (e.g. hsa-miR-223-3p and hsa-miR-185-5p for Illumina and hsa-miR-142-5p for CoolMPS) were among the 10 most expressed miRNAs in the other. This reinforces the necessity of deep sequencing, especially for the Illumina libraries, to quantify a larger range of miRNAs.

RT-qPCR data largely fit to the CoolMPS measurements

Finally, it is important to understand whether a third and independent technology validates the biomarker profiles. Since we previously already validated the BGISEQ approach using RT-qPCR (23) and demonstrate in the present work that CoolMPS is concordant to BGISEQ we can speculate that the RT-qPCR data would also match the CoolMPS profiles. To evaluate this hypothesis, we compared the expression values of 19 miRNAs that have been measured for 189 samples from the present study by RT-qPCR (25). Between the mean $\log_2$ CoolMPS expression and the negative delta CT values computed from RT-qPCR we observed a high correlation of 0.823 (Figure 4F). To validate how well this translates into biomarker patterns we again computed the difference between controls and dementia patients (Figure 4G). In this comparison we observed 10 miRNAs that were upregulated in both technologies, 5 miRNAs that were downregulated in both technologies and four miRNAs that were discordantly regulated between the technologies. According to Fishers Exact test this corresponds to a significant overlap ($P = 0.022$).

Table 1. For the top 10 most significant miRNAs with both technologies the rank in each technology is provided, followed by nominal and adjusted *P*-value, the effect size (Cohen's D) and AUC

miRNA	Rank BGISEQ	Rank CoolMPS	WMW raw *P*-value	WMW adj *P*-value	Cohen's D	AUC
hsa-miR-3688-3p	1	16	2.89E-06	0.0006	– 0.88	0.26
hsa-miR-3200-3p	**2**	**4**	**3.23E-06**	**0.0006**	**– 0.87**	**0.26**
hsa-let-7d-5p	3	17	6.98E-06	0.0007	0.83	0.73
hsa-miR-589-5p	4	51	9.26E-06	0.0007	– 0.79	0.27
hsa-miR-550a-3-5p	5	NA	9.39E-06	0.0007	0.78	0.73
hsa-let-7e-5p	**6**	**6**	**1.06E-05**	**0.0007**	**0.82**	**0.73**
hsa-miR-193a-3p	7	69	1.21E-05	0.0007	– 0.76	0.27
hsa-miR-4448	8	55	2.21E-05	0.0010	0.57	0.72
hsa-miR-15b-5p	**9**	**8**	**2.55E-05**	**0.0010**	**0.77**	**0.72**
hsa-miR-19b-3p	**10**	**1**	**2.64E-05**	**0.0010**	**– 0.77**	**0.28**
hsa-miR-181c-5p	21	2	2.38E-06	0.0004	– 0.80	0.26
hsa-miR-185-5p	48	3	4.84E-06	0.0006	– 0.75	0.26
hsa-miR-5695	12	5	7.70E-06	0.0006	– 0.76	0.27
hsa-miR-363-3p	33	7	2.15E-05	0.0011	0.75	0.72
hsa-miR-335-5p	117	9	5.55E-05	0.0022	– 0.44	0.29
hsa-miR-30b-5p	83	10	5.83E-05	0.0022	– 0.72	0.29

Bold miRNAs are in the top 10 for both technologies.

BGISEQ and CoolMPS AD miRNAs are matching known AD miRNAs and correlated to functional categories

As described in the previous sections, the miRNAs identified by the CoolMPS and BGISEQ approach have a significant diagnostic potential from a statistical perspective. We asked whether the signatures matched previously published results and which functional categories are enriched. To this end, we employed a miRNA set enrichment analysis using miEAA (37,38). As input the miRNAs were sorted with respect to their CoolMPS effect sizes. Downregulated miRNAs were most significantly associated to the miEAA disease category 'Downregulated in Alzheimer's Disease' (raw and adjusted *P*-value of 2.3×10^{-5} and 6.88×10^{-4}) while upregulated miRNAs were most strongly correlated to glioma (raw and adjusted *P*-value of 0.002 and 0.025, respectively). With respect to Gene Ontology and pathway databases we computed two significant categories. Upregulated AD miRNAs were enriched in chromosome condensation (raw and adjusted *P*-value of 3.3×10^{-6} and 0.018) as well as response to magnesium ion (raw and adjusted *P*-value of 1.3×10^{-5} and 0.036).

DISCUSSION

Whenever new technologies emerge in a field it is mandatory to test the fit to former technologies. The more disruptive a technological change is, the more the results differ from previous ones. An extreme example is the step from microarrays to RNA sequencing for analyzing expression profiles. If a novel technology aims to improve a previous one in a rather evolutionary manner by adapting and improving a specific step, the research results should generally be more aligned with previous findings. In biomedicine, such improvements can aim at an improved translational aspect of research in making workflows easier to use or in reducing the cost of assays. With CoolMPS we evaluated such an evolutionary improvement. Still, the main principle is sequencing-by-synthesis and also the detection and evaluation approach stay the same. The main difference is in using labeled antibodies instead of incorporating labeled nucleotides. While theoretical advantages of this approach, e.g. a potential re-use of the sequencing chemistry, are obvious we don't expect disruptive new findings. It is essential to benchmark CoolMPS to related high-throughput approaches, in our case standard cPAS sequencing-by-synthesis and Illumina sequencing, but also to a gold standard technology, in our case RT-qPCR. As primary comparison high-throughput technology we selected cPAS on the BGISEQ since we already previously performed a detailed benchmarking to the Illumina sequencing-by-synthesis approach, highlighting the advantages and disadvantages of both approaches (23). As biospecimens we intentionally selected whole blood. Not only because whole blood samples can be used to screen for minimally invasive biomarkers but also because of their challenging characteristics. The repertoire of small noncoding RNAs varies between different blood cell types and sncRNAs have a very high dynamic range. In fact, this means that few high abundant molecules are sequenced often whereas low abundant molecules are hardly observed. In whole blood small non-coding RNA sequencing data generated by Illumina sequencers, partially over 90% of the reads belong to miR-486-5p. While this miRNA is certainly highly abundant in red blood cells, this extreme distribution does not seem to match reality. In both, the BGISEQ and CoolMPS data we still observe an extreme distribution with around $\frac{3}{4}$ of all reads matching to the most abundant miRNA, miR-451a. This can also be recognized in Supplementary Figure S1K and L. Still, this distribution is less extreme than for the previously investigated Illumina sequencing data. The less extreme overrepresentation in the BGISEQ and CoolMPS data thus facilitates the discovery of yet unknown and less abundant non-coding RNA molecules.

Among the top 10 markers that we discovered by CoolMPS (Table 1), eight miRNAs (miR-19b-3p, miR-181c-5p, miR-185-5p, miR-3200-3p, let-7e-5p, miR-15b-5p, miR-335-5p and miR-30b-5p) were already described in the literature to be correlated to Alzheimer's disease or dementia. For example, miR-19b-3p prevents amyloid β-induced injury by targeting BACE1 in SH-SY5Y cells (40) and is altered in CSF exosomes of AD patients (41). Similarly, miR-185-5p is known as exosomal AD biomarker (42). Also, let-7e-5p and miR-3200-3p were previously identified as blood biomarkers (43). Interestingly, the same manuscript also lists miR-30c-5p, miR-30d-5p and miR-15a-5p. For these miRNAs we report differential expression in related miRNA family members (miR-30b-5p and miR15b-5p respectively). The latter miRNA has also been reported in other studies a circulating AD biomarker (44,45) and targets the amyloid precursor protein (46). Similarly, miR-335-5p inhibits β-Amyloid in AD (47). Already for the 10 most significant miRNAs we thus found substantial evidence for their role in AD, both as biomarker but also linked to a potential pathogenic function.

One step in our analysis pipeline is to filter out short reads (below 17 nucleotides), that might add noise to the data. For BGISEQ, the lowest number of reads was filtered out in this step followed by CoolMPS and Illumina sequencing data. While the percentages overall were similar, we observed a higher standard deviation in Illumina data (3.24%) as compared to BGISEQ (0.54) and CoolMPS (0.56) data. In comparing CoolMPS data to Illumina data we observed a slightly better averaged Q30 value for the Illumina data. This advantage could be observed however mostly in the beginning of the read. Toward the end of the 50 base reads, Illumina Q30 values dropped more as compared to the stable performance of CoolMPS. This resulted in a higher mapping rate of the CoolMPS data. One explanation for a drop of quality is in the small size of miRNAs that are usually shorter than 25 nucleotides but 50 bases are sequenced. This effect might be more pronounced for Illumina as compared to the BGISEQ and CoolMPS data. In consequence, we can expect that this factor is likely less relevant for longer RNAs or sequencing of DNA. Also, the composition of the RNA classes was different between the technologies. Illumina data revealed higher percentages of piRNAs and miRNAs while CoolMPS shows a higher diversity also including other non-coding RNA classes. A difference between the BGISEQ/CoolMPS and Illumina protocols was the amount of starting material. For BGISEQ and CoolMPS, 800 ng was used while the Illumina data have been generated from 200 ng input material. This might pretend that a higher input amount is required for CoolMPS as compared to Illumina. We used this higher input amount however only during the exploratory phase of the CoolMPS protocol. Even with lower amount of input material down to 100ng we did not observe significant changes (data not shown). Indeed, the manufacturer's instruction would even allow input from 10ng RNA only. Thus, the input volume seems not to be a limiting factor for the CoolMPS technology.

In sum, both of the technologies have their advantages and disadvantages and the best systems should be chosen dependent on the application. Our data thus clearly suggest that small RNA sequencing results from Illumina data should not be directly compared to sequencing results from BGISEQ since the technical differences between identical samples are statistically highly significant. With respect to comparing between BGISEQ and CoolMPS datasets we observed generally very similar performance. The most striking advantage of CoolMPS is a significantly improved single base call quality. This led to marginal improvements in the biomarker patterns but did not improve the performance of any biomarker in a substantial manner. Interpreting the results, we have to bear in mind that the BGISEQ technology and chemistry have already matured over at least five years while we used prototype beta testing chemistry for CoolMPS. Since already this chemistry lead to improved performance we can expect further improvements with revised kits of CoolMPS. Finally, one big advantage is the potential to recover the used labeled antibodies for a second sequencing run.

DATA AVAILABILITY

All sequencing data have been deposited in the Sequence Read Archive with the accession SRP271972.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We appreciate the support of all participants contributing to the research in providing blood samples.

FUNDING

Saarland University; Michael J. Fox Foundation for Parkinson's Research [14446]; Complete Genomics. Funding for open access charge: Michael J. Fox Foundation.

Conflict of interest statement. The authors with affiliations (2, 3, 5, 6: Y.L., A.B., S.D., S.L., C.X., M.C., C.V., A.J., H.J. and R.D.) are employed by BGI and MGI group, manufacturing the CoolMPS kits and distributing these kits commercially.

REFERENCES

1. Ronaghi,M., Karamohamed,S., Pettersson,B., Uhlen,M. and Nyren,P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.*, **242**, 84–89.
2. Koboldt,D.C., Steinberg,K.M., Larson,D.E., Wilson,R.K. and Mardis,E.R. (2013) The next-generation sequencing revolution and its impact on genomics. *Cell*, **155**, 27–38.
3. Benson,D.A., Cavanaugh,M., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2013) GenBank. *Nucleic Acids Res.*, **41**, D36–D42.
4. Saliba,A.E., Westermann,A.J., Gorski,S.A. and Vogel,J. (2014) Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.*, **42**, 8845–8860.
5. Slatko,B.E., Gardner,A.F. and Ausubel,F.M. (2018) Overview of next-generation sequencing technologies. *Curr. Protoc. Mol. Biol.*, **122**, e59.
6. Senabouth,A., Andersen,S., Shi,Q., Shi,L., Jiang,F., Zhang,W., Wing,K., Daniszewski,M., Lukowski,S.W., Hung,S.S.C. *et al.* (2020) Comparative performance of the BGI and Illumina sequencing technology for single-cell RNA-sequencing. *NAR Genomics Bioinform.*, **2**, lqaa034.
7. Mathew,R., Mattei,V., Al Hashmi,M. and Tomei,S. (2020) Updates on the current technologies for microRNA profiling. *Microrna*, **9**, 17–24.

8. Alles,J., Fehlmann,T., Fischer,U., Backes,C., Galata,V., Minet,M., Hart,M., Abu-Halima,M., Grasser,F.A., Lenhof,H.P. *et al.* (2019) An estimate of the total number of true human miRNAs. *Nucleic Acids Res.*, **47**, 3353–3364.
9. Fehlmann,T., Backes,C., Alles,J., Fischer,U., Hart,M., Kern,F., Langseth,H., Rounge,T., Umu,S.U., Kahraman,M. *et al.* (2018) A high-resolution map of the human small non-coding transcriptome. *Bioinformatics*, **34**, 1621–1628.
10. Fehlmann,T., Backes,C., Pirritano,M., Laufer,T., Galata,V., Kern,F., Kahraman,M., Gasparoni,G., Ludwig,N., Lenhof,H.P. *et al.* (2019) The sncRNA Zoo: a repository for circulating small noncoding RNAs in animals. *Nucleic Acids Res.*, **47**, 4431–4441.
11. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
12. 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.
13. Fromm,B., Domanska,D., Hoye,E., Ovchinnikov,V., Kang,W., Aparicio-Puerta,E., Johansen,M., Flatmark,K., Mathelier,A., Hovig,E. *et al.* (2020) MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res.*, **48**, D132–D141.
14. Backes,C., Fehlmann,T., Kern,F., Kehl,T., Lenhof,H.P., Meese,E. and Keller,A. (2018) miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res.*, **46**, D160–D167.
15. Fromm,B., Keller,A., Yang,X., Friedlander,M.R., Peterson,K.J. and Griffiths-Jones,S. (2020) Quo vadis microRNAs? *Trends Genet*, **36**, 461–463.
16. Fehlmann,T., Backes,C., Kahraman,M., Haas,J., Ludwig,N., Posch,A.E., Wurstle,M.L., Hubenthal,M., Franke,A., Meder,B *et al.* (2017) Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic Acids Res.*, **45**, 8731–8744.
17. Friedlander,M.R., Chen,W., Adamidi,C., Maaskola,J., Einspanier,R., Knespel,S. and Rajewsky,N. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, **26**, 407–415.
18. Friedlander,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
19. Heinicke,F., Zhong,X., Zucknick,M., Breidenbach,J., Sundaram,A.Y.M., S,T.F., Leithaug,M., Dalland,M., Farmer,A., Henderson,J.M. *et al.* (2020) Systematic assessment of commercially available low-input miRNA library preparation kits. *RNA Biol.*, **17**, 75–86.
20. Meistertzheim,M., Fehlmann,T., Drews,F., Pirritano,M., Gasparoni,G., Keller,A. and Simon,M. (2019) Comparative analysis of biochemical biases by ligation- and template-switch-Based small RNA library preparation protocols. *Clin. Chem.*, **65**, 1581–1591.
21. Ludwig,N., Fehlmann,T., Galata,V., Franke,A., Backes,C., Meese,E. and Keller,A. (2018) Small ncRNA-Seq results of human Tissues: Variations depending on sample integrity. *Clin. Chem.*, **64**, 1074–1084.
22. Baroin-Tourancheau,A., Jaszczyszyn,Y., Benigni,X. and Amar,L. (2019) Evaluating and correcting inherent bias of microRNA expression in Illumina sequencing analysis. *Front. Mol. Biosci.*, **6**, 17.
23. Fehlmann,T., Reinheimer,S., Geng,C., Su,X., Drmanac,S., Alexeev,A., Zhang,C., Backes,C., Ludwig,N., Hart,M. *et al.* (2016) cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin. Epigenet.*, **8**, 123.
24. Keller,A., Backes,C., Haas,J., Leidinger,P., Maetzler,W., Deuschle,C., Berg,D., Ruschil,C., Galata,V., Ruprecht,K. *et al.* (2016) Validating Alzheimer's disease micro RNAs using next-generation sequencing. *Alzheimers Dement*, **12**, 565–576.
25. Ludwig,N., Fehlmann,T., Kern,F., Gogol,M., Maetzler,W., Deutscher,S., Gurlit,S., Schulte,C., von Thaler,A.K., Deuschle,C. *et al.* (2019) Machine learning to detect Alzheimer's disease from circulating Non-coding RNAs. *Genomics Proteomics Bioinform.*, **17**, 430–440.
26. Fehlmann,T., Meese,E. and Keller,A. (2017) Exploring ncRNAs in Alzheimer's disease by miRMaster. *Oncotarget*, **8**, 3771–3772.
27. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
28. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
29. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
30. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
31. Wang,J., Zhang,P., Lu,Y., Li,Y., Zheng,Y., Kan,Y., Chen,R. and He,S. (2019) piRBase: a comprehensive database of piRNA sequences. *Nucleic Acids Res.*, **47**, D175–D180.
32. Chan,P.P. and Lowe,T.M. (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.*, **44**, D184–D189.
33. Backes,C., Meder,B., Hart,M., Ludwig,N., Leidinger,P., Vogel,B., Galata,V., Roth,P., Menegatti,J., Grasser,F. *et al.* (2016) Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic Acids Res.*, **44**, e53.
34. Robin,X., Turck,N., Hainard,A., Tiberti,N., Lisacek,F., Sanchez,J.C. and Muller,M. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77.
35. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.
36. Amand,J., Fehlmann,T., Backes,C. and Keller,A. (2019) DynaVenn: web-based computation of the most significant overlap between ordered sets. *BMC Bioinformatics*, **20**, 743.
37. Kern,F., Fehlmann,T., Solomon,J., Schwed,L., Grammes,N., Backes,C., Van Keuren-Jensen,K., Craig,D.W., Meese,E. and Keller,A. (2020) miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. *Nucleic Acids Res.*, **48**, W521–W528.
38. Backes,C., Khaleeq,Q.T., Meese,E. and Keller,A. (2016) miEAA: microRNA enrichment analysis and annotation. *Nucleic Acids Res.*, **44**, W110–W116.
39. Amrhein,V., Greenland,S. and McShane,B. (2019) Scientists rise up against statistical significance. *Nature*, **567**, 305–307.
40. Zhang,N., Li,W.W., Lv,C.M., Gao,Y.W., Liu,X.L. and Zhao,L. (2020) miR-16-5p and miR-19b-3p prevent amyloid beta-induced injury by targeting BACE1 in SH-SY5Y cells. *Neuroreport*, **31**, 205–212.
41. Gui,Y., Liu,H., Zhang,L., Lv,W. and Hu,X. (2015) Altered microRNA profiles in cerebrospinal fluid exosome in Parkinson disease and Alzheimer disease. *Oncotarget*, **6**, 37043–37053.
42. Lugli,G., Cohen,A.M., Bennett,D.A., Shah,R.C., Fields,C.J., Hernandez,A.G. and Smalheiser,N.R. (2015) Plasma exosomal miRNAs in persons with and without Alzheimer Disease: Altered expression and prospects for biomarkers. *PLoS One*, **10**, e0139233.
43. Satoh,J., Kino,Y. and Niida,S. (2015) MicroRNA-seq data analysis pipeline to identify blood biomarkers for Alzheimer's disease from public data. *Biomark Insights*, **10**, 21–31.
44. Kumar,P., Dezso,Z., MacKenzie,C., Oestreicher,J., Agoulnik,S., Byrne,M., Bernier,F., Yanagimachi,M., Aoshima,K. and Oda,Y. (2013) Circulating miRNA biomarkers for Alzheimer's disease. *PLoS One*, **8**, e69807.
45. Wu,H.Z.Y., Thalamuthu,A., Cheng,L., Fowler,C., Masters,C.L., Sachdev,P., Mather,K.A. and and the Australian Imaging, B. and Lifestyle Flagship Study of, A. (2020) Differential blood miRNA expression in brain amyloid imaging-defined Alzheimer's disease and controls. *Alzheimers Res. Ther.*, **12**, 59.
46. Liu,H.Y., Fu,X., Li,Y.F., Li,X.L., Ma,Z.Y., Zhang,Y. and Gao,Q.C. (2019) miR-15b-5p targeting amyloid precursor protein is involved in the anti-amyloid eflect of curcumin in swAPP695-HEK293 cells. *Neural. Regen. Res.*, **14**, 1603–1609.
47. Wang,D., Fei,Z., Luo,S. and Wang,H. (2020) MiR-335-5p Inhibits beta-Amyloid (Abeta) accumulation to attenuate cognitive deficits through targeting c-jun-N-terminal kinase 3 in Alzheimer's disease. *Curr. Neurovasc. Res.*, **17**, 93–101.

doi: 10.1093/nar/gkab297

miRTargetLink 2.0—interactive miRNA target gene and target pathway networks

Fabian Kern[1,†], **Ernesto Aparicio-Puerta**[1,†], **Yongping Li**[1,†], **Tobias Fehlmann**[1], **Tim Kehl**[2], **Viktoria Wagner**[1], **Kamalika Ray**[1], **Nicole Ludwig**[3], **Hans-Peter Lenhof**[2], **Eckart Meese**[3] **and Andreas Keller**[1,4,*]

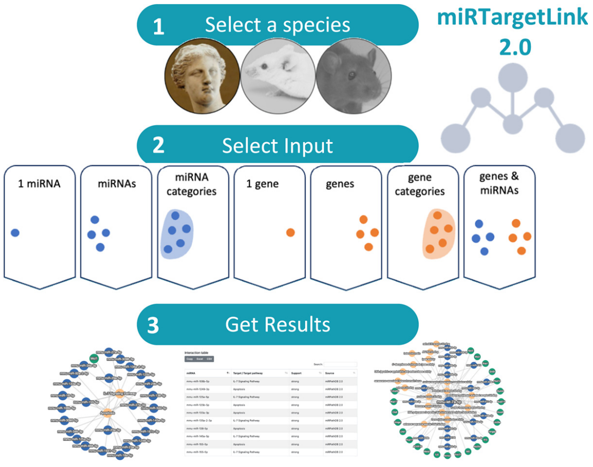
[1]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [2]Center for Bioinformatics, Saarland Informatics Campus, Saarland University, 66123 Saarbrücken, Germany, [3]Center for Human and Molecular Biology, Institute of Human Genetics, Saarland University, 66421 Homburg, Germany and [4]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford 94304, CA, USA

Received February 24, 2021; Revised April 05, 2021; Editorial Decision April 08, 2021; Accepted April 12, 2021

ABSTRACT

Which genes, gene sets or pathways are regulated by certain miRNAs? Which miRNAs regulate a particular target gene or target pathway in a certain physiological context? Answering such common research questions can be time consuming and labor intensive. Especially for researchers without computational experience, the integration of different data sources, selection of the right parameters and concise visualization can be demanding. A comprehensive analysis should be central to present adequate answers to complex biological questions. With miRTargetLink 2.0, we develop an all-in-one solution for human, mouse and rat miRNA networks. Users input in the unidirectional search mode either a single gene, gene set or gene pathway, alternatively a single miRNA, a set of miRNAs or an miRNA pathway. Moreover, genes and miRNAs can jointly be provided to the tool in the bidirectional search mode. For the selected entities, interaction graphs are generated from different data sources and dynamically presented. Connected application programming interfaces (APIs) to the tailored enrichment tools miEAA and GeneTrail facilitate downstream analysis of pathways and context-annotated categories of network nodes. MiRTargetLink 2.0 is freely accessible at https://www.ccb.uni-saarland.de/mirtargetlink2.

GRAPHICAL ABSTRACT



INTRODUCTION

A central question in biomedical and life science research studies is how the expression of genes is modulated in physiological and pathophysiological processes (1). In this context, miRNAs play a central role in orchestrating gene expression. A frequent mechanism is that miRNAs bind to a specific sequence at the 3′ untranslated region (UTR) of a target mRNA to induce translational repression (2). In addition to this common mode of action, miRNA-binding sites in other mRNA regions such as the 5′ untranslated region, coding sequence or promoter regions exist (3). Moreover, RNA-binding proteins have an important role in the regulation of miRNA activity (4). An overview on canonical and noncanonical miRNA targeting (including aspects upstream such as miRNA biogenesis) has been provided by O'Brien *et al.* (5) and Erkeland (6). Moreover, Bartel provided a comprehensive overview on the current understand-

[*]To whom correspondence should be addressed. Tel: +49 681 302 68611; Email: andreas.keller@ccb.uni-saarland.de
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

ing of the defining features for miRNA biogenesis and related genomics (7).

It has become evident that miRNAs target genes in a systematic manner and exhibit a targetome specificity up to the pathway level (8). On the molecular level this effect manifests in different aspects. For example, the 3′ UTR of mRNAs often harbors multiple binding sites of the same miRNA, and the binding to these target sites has a cooperative effect. Likewise, the binding of different miRNAs to the same UTR can have cooperative effects. To facilitate the systematic analysis of miRNAs in the context of target genes or vice versa, i.e., in one-to-many or many-to-many relationships, we implemented miRTargetLink (9). The analysis of putative target genes and target pathways in a systems biology context is an essential step in non-coding RNA studies, including contrasts of experimental groups in disease research. Thus, several research tools with varying scope and functionality, as both stand-alone and web-based tools, have been proposed. Tools4mirs, a popular meta-repository for miRNA analysis methods (10), currently lists 60 target prediction tools and 26 toolboxes for the functional analysis of targets. In the following, we introduce the tools with an application scope closest to miRTargetLink 2 (11,12). For instance, miRTarVis is a tool specifically developed to display co-expression networks of paired miRNA and mRNA data (13). Second, MIENTURNET generates interaction networks by estimating the statistical significance of paired lists of miRNA and mRNA identifiers provided (14). An advantage of this approach is to reduce potentially large input lists to likely core interactions of the putative biological network, while a disadvantage is that weak informative edges might be removed due to statistical instability, potentially introducing a bias to the network. The tool miRViz allows to quickly visualize precomputed networks for multiple species, based on preselected features such as shared seed region identities between related miRNAs from the same or similar families (15). Furthermore, miRNet is a web-based tool supporting statistical analysis and functional interpretation miRNA studies (16,17). Also, it facilitates exploring the results in miRNA–target interaction networks. To analyse miRNA function in a more tissue-specific manner, miTALOS has been developed (18,19). Furthermore, tools that analyse miRNA and gene expression data in an integrated manner are available. One example of such is MMIA (miRNA and mRNA integrated analysis) that processes miRNA and gene expression experiments (20). With a similar application focus, TaLasso (21) and miRTrail were developed (22). Other more specialised web servers such as FFLtool are designed for transcription factor and miRNA feed-forward loop analysis (23).

In general, many methods are based on gene and miRNA expression data to find putative new regulatory edges or integrate known edges from miRNA target gene association databases. To test the significance of putative interactions in a statistical framework, several tools perform miRNA and gene set enrichment analyses to annotate biological function. One research goal in our studies was to provide evidence that miRNAs target genes in a systematic manner. To this end, we released miRPathDB 2.0, indexing thousands of enriched pathways for known miRNAs and miRNA candidates using validated and predicted target genes from the literature (24). Following up on our observations, we published the first comprehensive experimental validation of miRNA target pathway regulation (8).

In 2016, we presented miRTargetLink Human (9), a tool that hierarchically builds miRNA regulatory networks, containing validated and predicted target genes. Here, we present miRTargetLink 2.0, a novel version of our interactive tool for systems biology applications in miRNA research by a dynamic presentation of miRNA target gene and pathway networks. We provide a large set of miRNA gene associations from published repositories [miRTarBase (25), mirDIP (26), miRDB (27) and miRATBase (8)] and extend it by the pathway data from the recent release of miRPathDB 2.0 (24). Besides new analysis-centric features that are described in this manuscript, we want to highlight the new multi-species support, as *Mus musculus* and *Rattus norvegicus* are now available for analysis.

MATERIALS AND METHODS

Data selection and processing

The new version of miRTargetLink supports miRNAs and targets for *Homo sapiens*, *M. musculus* and *R. norvegicus*. MiRNA identifiers and annotation records were obtained from the latest release of miRBase (v.22.1) (28), and validated targets were downloaded from miRTarBase (v.8) (25) and miRATBase (8).

As for predicted targets, top 5% predictions (high and very high confidence) from mirDIP (v.4.1) (27) were used for human whereas we used miRDB (v.6) (27) for mouse and rat. miRTargetLink also supports target pathways from miRPathDB 2.0 (24).

Annotations for functional or categorical miRNA sets were obtained from miEAA 2.0 (29) and from GeneTrail 3 (30) for gene sets. MiEAA sets can be used for all three species, but target pathways are only available for human and mouse. Mygene python package (31) was used to translate RefSeq names to gene symbols where required. An overview on the different tools used throughout this work is given in Table 1 along with the type, the task we used the resources for and the respective version.

Web server implementation

The web server was implemented using Django v2.2 in a docker environment with a PostgreSQL database, celery for job scheduling and execution and Redis as message broker backend. The frontend was built using common HTML, CSS and Javascript libraries, including the Bootstrap framework (v.4) for the styling, dataTables for the network node and interaction tables, and jQuery and Cytoscape.js (32) to create the interactive network visualizations. Typing-ahead suggestions are generated using the autoComplete JS library.

RESULTS

Data input

We implemented a convenient workflow for miRTargetLink 2 (Figure 1). We observed that most research questions re-

Table 1. Overview of in-house and third-party resources included in miRTargetLink 2.0

Database	Task	Own/third-party	Type	Version
miRBase	miRNA annotation	Third-party	Database	22.1
miRTarBase	miRNA target database	Third-party	Database	8.0 (2020)
mirDIP	miRNA target database	Third-party	Database	4.1
miRDB	miRNA target database	Third-party	Database	6.0
miRATBase	miRNA target database	Own	Database	1.0
miRPathDB	miRNA target pathway database	Own	Database	2.0
miEAA	miRNA set enrichment analysis	Own	Tool	2.0
GeneTrail	Gene set enrichment analysis	Own	Tool	3.0

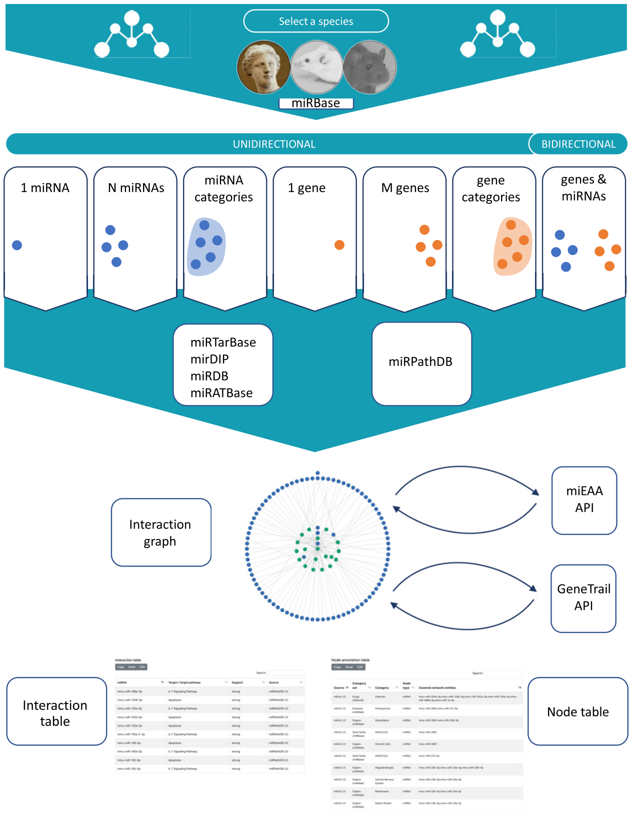


Figure 1. Workflow of miRTargetLink 2.0. The user selects the input species and the data input option. All information is extracted from incorporated databases automatically, and the interaction graph is immediately generated and visualized. APIs to gene and miRNA set enrichment facilitate the interpretation of more complex interaction graphs. Nodes and edges are further annotated in interactive tables.

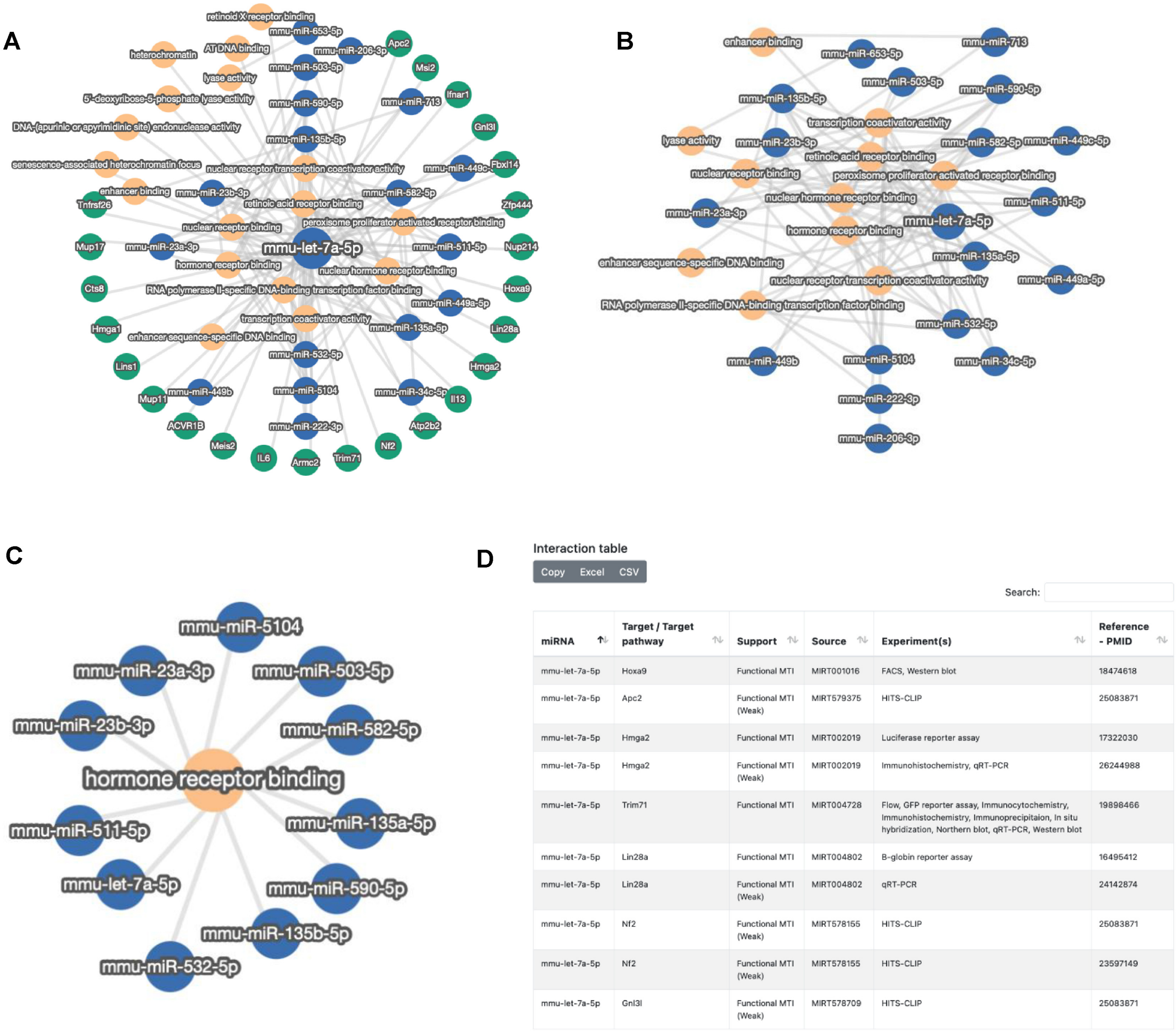
lated to miRNAs and genes, respectively, and their interactions can be addressed from a simple yet powerful selection of input types. In the unidirectional mode, the user can decide from six upload options, i.e. whether to select a single miRNA, a single gene, a list of either miRNAs or genes and lastly a predefined miRNA set or a gene set, as for instance, disease-associated miRNAs. In addition, a bidirectional query (paired miRNA gene lists) can be initiated. To prevent unintended results that may occur if the organism is selected automatically (e.g. genes can share the same name in mouse and human), we also ask the user to select the organism. From the input data, a comprehensive network on miRNA targets and target pathways is generated. As background data sets, we integrate four miRNA–target databases (miRTarBase, mirDip, miRDB and miRATBase) complemented with pathway interactions from miRPathDB. Altogether, the miRTargetLink knowledge base hosts ∼553 000 entries from miRTarBase, ∼1 519 000 entries from mirDIP, ∼1 173 000 entries from miRDB, ∼300 targets from miRATBase and ∼13 000 entries from miRPathDB. The detailed distribution of data records per organism and category such as validated or predicted targets, or pathways is presented on the miRTargetLink statistics page.

Representation of results

Based on the input and the information in the knowledge database, the interaction graphs with all edges between miRNAs and targets, and miRNAs and target pathways are generated and visualized. Edges between genes and pathways have been omitted from the graphs, since those would introduce more complexity to the graphs without adding information content for the miRNA-centered application. The network can be shifted, zoomed and node positions and colors can be adjusted. To edit the network, the user can choose from a range of options, e.g. select whether weak/strong evidence or also predicted targets are shown or whether pathways should be added. If pathways are available, then also other miRNAs regulating these pathways can be revealed. Further, the number of shared targets for each miRNA can be increased to highlight the key regulators. Finally, six different layout options for the network are available. The network can be downloaded as JSON file or as image in jpg and png format. Below the network view, interactions are presented as the table where miRNAs, target evidence and sources are given in detail. In addition, available node annotations based on biological categories such as known tissue expression of miRNAs and genes are shown in a separate view. Both tables can be downloaded in either xls or csv format and directly copied to the clipboard.

APIs facilitate gene and miRNA set enrichment analysis

Once the miRNA and target gene network has been generated by miRTargetLink 2, the interpretation of the results becomes important. For small networks and field experts, this task can often be achieved manually. But especially for larger networks with several dozens of genes and miRNAs, it is frequently not obvious where to focus on. Here, the pathway information from miRPathDB supports interpretability, but it focuses on miRNA pathways that are annotated with experimental evidence. To guide researchers and to draw their attention to relevant hits, we integrated



Interaction table

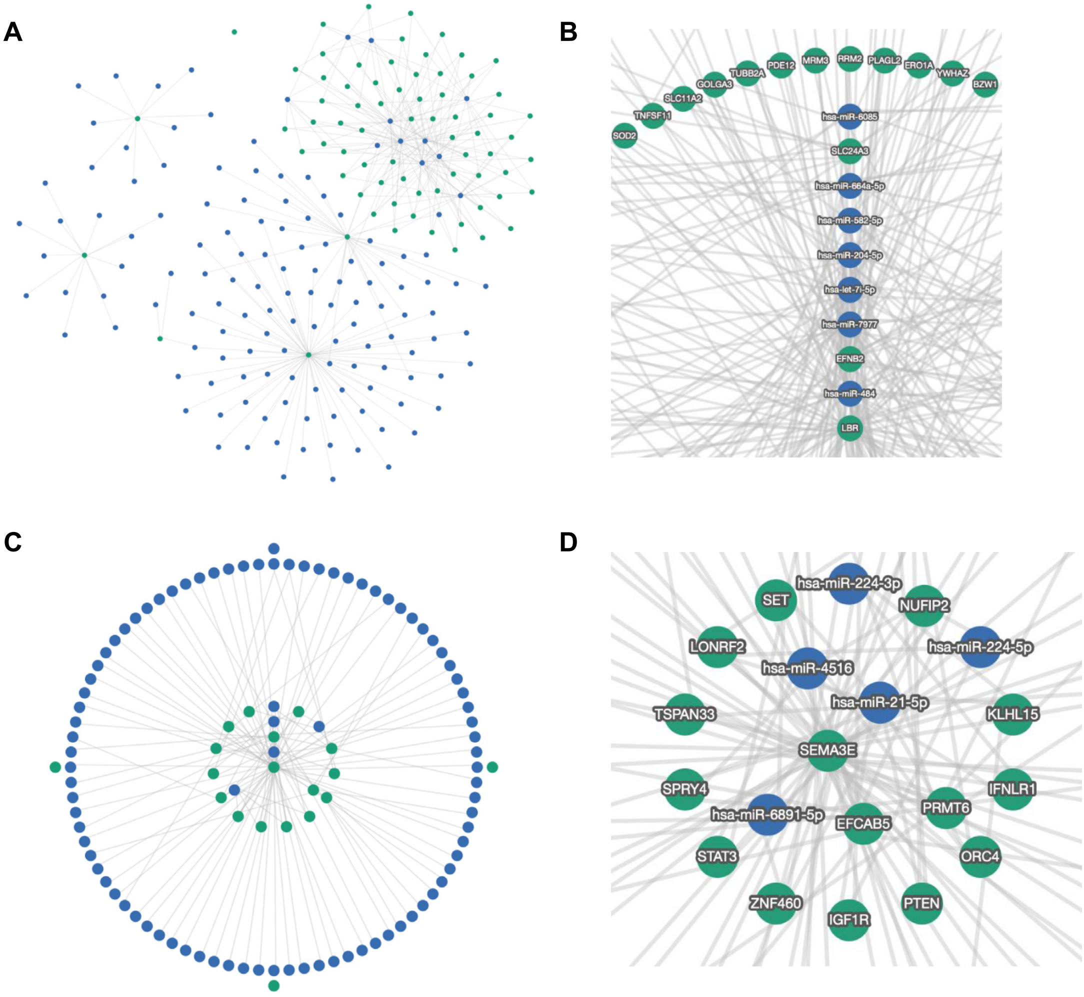
miRNA	Target / Target pathway	Support	Source	Experiment(s)	Reference - PMID
mmu-let-7a-5p	Hoxa9	Functional MTI	MIRT001016	FACS, Western blot	18474618
mmu-let-7a-5p	Apc2	Functional MTI (Weak)	MIRT579375	HITS-CLIP	25083871
mmu-let-7a-5p	Hmga2	Functional MTI	MIRT002019	Luciferase reporter assay	17322030
mmu-let-7a-5p	Hmga2	Functional MTI (Weak)	MIRT002019	Immunohistochemistry, qRT-PCR	26244988
mmu-let-7a-5p	Trim71	Functional MTI	MIRT004728	Flow, GFP reporter assay, Immunocytochemistry, Immunohistochemistry, Immunoprecipitaion, In situ hybridization, Northern blot, qRT-PCR, Western blot	19898466
mmu-let-7a-5p	Lin28a	Functional MTI	MIRT004802	B-globin reporter assay	16495412
mmu-let-7a-5p	Lin28a	Functional MTI (Weak)	MIRT004802	qRT-PCR	24142874
mmu-let-7a-5p	Nf2	Functional MTI (Weak)	MIRT578155	HITS-CLIP	25083871
mmu-let-7a-5p	Nf2	Functional MTI (Weak)	MIRT578155	HITS-CLIP	23597149
mmu-let-7a-5p	Gnl3l	Functional MTI (Weak)	MIRT578709	HITS-CLIP	25083871

Figure 2. Use case 1—mouse let-7a-5p targetome. (**A**) For the input option of a single miRNA (mmu-let-7a-5p), the interaction graph is presented. (**B**) After increasing the number of required target interactions per gene, only the pathways and miRNAs remain. (**C**) As the second input, the hormone receptor binding was selected, and miRNAs targeting this pathway are shown. (**D**) Interaction graph for mmu-let-7a-5p as browsable table adjacent to the network representation. The first 10 of 129 entries are displayed.

miRNA and gene set enrichment through miEAA 2.0 (29) and GeneTrail 3 (30) via external APIs. By selecting the relevant enrichment analysis method, these tools are automatically executed using their standard parameters and the aggregated download of result tables is initiated after completion.

Context dependency

miRNA expression is known to depend on the physiological or pathophysiological contexts. For example, the age of unaffected individuals or patients with different diseases can affect miRNA expression (33). Cholinergic-targeting noncoding RNAs, miRNAs or lncRNAs can also modulate sex-specific- and age-related acetylcholine signals (34). Especially in the context of age-related disorders such as Parkinson's disease, miRNAs seem to be differentially expressed in specific age windows (1). Moreover, the tissue or different body fluids can confound miRNA expression (35), and regulatory events between miRNAs and target genes seem to depend on the tissue context (36). As the first step to make this information visible, miRTargetLink 2 contains and displays available context information where available. Specifically, we added metadata information from miRTarBase about the experimental setup around which the listed interactions were obtained, e.g. which type of experiment was used along with the source PubMed ID. Moreover, miRTargetLink specifically allows to search for age- and sex-related miRNAs on the unidirectional search. Similarly, users can spawn a network by searching for biological pathways or categories from GeneTrail 3. For instance, users can test systematically for a bias in a current network toward sex- or age-associated miRNAs using the connected miEAA or GeneTrail APIs, respectively. Finally, we added a one-click



E

Category	P-value	P-adjusted	Expected	Observed
SEMA3E	3,22E-98	2,95E-94	2,02654	59
chemorepellent activity GO0045499	5,03E-35	3,95E-31	14,0484	65
branching involved in blood vessel morphogenesis GO0001569	6,25E-35	3,95E-31	13,4988	64
negative regulation of axon extension involved in axon guidance GO0048843	3,79E-34	1,60E-30	13,8767	64
regulation of actin cytoskeleton reorganization GO2000249	7,76E-34	2,45E-30	16,5902	68
negative regulation of cell-matrix adhesion GO0001953	3,05E-33	7,71E-30	16,2467	67
semaphorin receptor binding GO0030215	1,51E-32	3,19E-29	15,3193	65
synapse organization GO0050808	2,70E-32	4,88E-29	16,1093	66
neural crest cell migration GO0001755	3,08E-30	4,86E-27	17,3115	66
organic acidsodium symporter activity GO0005343	4,38E-30	6,15E-27	1,54567	27
SLC5A12	4,38E-30	2,01E-26	1,54567	27
negative chemotaxis GO0050919	9,12E-29	1,15E-25	19,7158	68
sprouting angiogenesis GO0002040	3,66E-28	4,20E-25	19,3724	67

Figure 3. Use case 2—aniridia genes and miRNAs. (**A**) Interactions between downregulated miRNAs (blue nodes) and upregulated genes (green nodes) with a minimum of three shared targets. (**B**) The enlarged core part of the interaction network shown in (A). (**C**) The opposite case upregulated miRNAs and downregulated genes. (**D**) The enlarged central part of the interaction graph shown in (C). (**E**) Top 10 significant miEAA pathways for the network in panel (C). Following the category name, the raw and adjusted *P*-value are provided, followed by the expected number of miRNAs by chance and the actual observed number.

literature search functionality. Users can initiate a search for a selected miRNA or gene in the context of age, sex or function using PubMed.

Use case 1—mouse miRNA let-7a-5p target network

As the first use case, we studied the target gene and target pathway network of mouse miRNA let-7a-5p. This miRNA has previously been described in *M. musculus* with various functions (37–39). The quick-search functionality highlighted 24 target genes, including *IL6* and *IL13*. Additionally, 17 pathways are targeted by this miRNA (Figure 2A). To examine whether other miRNAs are also targeting these genes, we increased the number of minimal targets. As a result, we got a target network that only contains miRNAs and the aforementioned pathways (Figure 2B). To display all miRNAs in mouse targeting the hormone receptor binding, we applied the third input option presented above, resulting in a condensed set of 11 miRNAs (Figure 2C). The total interaction set for mmu-let-7a-5p contains 129 entries that are displayed below the interaction graph (Figure 2D). The results presented in this use case focus on a single miRNA or a single pathway with an interaction graph of only a few dozens of nodes and edges, which can be well interpreted manually, not yet requiring gene and miRNA set enrichment analysis. To further showcase this application, we next evaluated human aniridia paired miRNA and gene expression data.

Use case 2—integrated miRNA/gene analysis in human aniridia

As second use case, we explored miRNA and mRNA expression patterns that were generated from the same case and control samples, facilitating paired analysis of the two RNA classes (40). From this study, we extracted both dysregulated genes and miRNAs. We then performed two analyses, where we computed the network of upregulated genes and downregulated miRNAs and, vice versa, the network of upregulated miRNAs and downregulated genes. This opposite direction of regulation was selected due to the dominant biological role of miRNAs repressing the translation of target mRNAs. In all cases, we limited the input to the top 10 genes and miRNAs. In the first scenario, the interaction graph contained 4609 edges. Even after requiring a minimum shared number of three targets, the graph contains 422 edges (Figure 3A), however, revealing a central component and the genes *SLC24A3*, *EFNB2* and *LBR* with high node degrees (Figure 3B). The opposite use case still highlighted 2474 entries in the initial interaction graph. Here, 228 edges remained in the collapsed network with a minimal number of three shared targets (Figure 3C). The genes central to the network were *SEMA3E*, *EFCAB5*, *PRMT6* and several others (Figure 3D). To simplify the analysis of the complex network, we performed miRNA set enrichment analysis. The 10 most significant pathways and categories are provided as results table (Figure 3E). The top hit with an adjusted *P*-value of 3×10^{-94} was the gene *SEMA3E*, exactly validating a statistically significant coverage by multiple miRNAs in the interaction graph, more than one would expect for a random enrichment. Second most significant was the Gene Ontology category chemorepellent activity with an adjusted *P*-value of 4×10^{-31}, followed by neuronal pathways, cell adhesion and others. This use case demonstrates how the practical application of miRTargetLink 2 guides researchers to focus on potentially more relevant biological findings.

DISCUSSION AND CONCLUSION

We present a significant update of our web server miRTargetLink 2 for the integrative analysis of miRNA, target gene and target pathway interaction networks. While the original version was focused on human data, we now offer support for other highly relevant model organisms. Altogether, the integrated knowledge base contains over 3 million entries of regulatory events between miRNAs and genes, and miRNAs and pathways across the three supported species. By adding the layer of validated pathways to the network view and providing quick access to frequently used gene / miRNA set enrichment tools, we lower the boundaries for potential users from life science to generate new insights into driving questions in fundamental biology and biomedicine. Due to the largely increased number of miRNA–target interactions and to prevent major performance issues, we were required to completely reimplement both the front- and backend of the original web server.

One current limitation of miRTargetLink is the restriction of its scope to miRNAs. Other small RNA classes are emerging and should be taken into account. Our tool miRMaster (41) already includes all previously characterized noncoding RNA classes. Among those most similar to miRNAs, tRNA fragments play a remarkable role. For example, tRNA fragments can replace miRNA regulators in diseases, as demonstrated for the cholinergic poststroke immune blockade (42). As such, tRNA fragments will add to a complete picture of how small RNAs regulate genes. Before such information is added to available integrative miRNA tools like miRTargetLink, the development of comprehensive databases containing detailed and experimentally validated regulatory events of tRNAs is mandatory.

In addition to expanding miRTargetLink to other noncoding RNA classes in the future, we will continue to add new features requested by the community. One extension could be to support uploading of expression or fold-change scores together with the identifiers such as to dynamically modify node and edge strength in the inferred network. This way, one could rank and select connected components in the graph according to their importance in a particular research context.

DATA AVAILABILITY

Not applicable.

FUNDING

Internal funds of Saarland University; Instituto de Salud Carlos III [IFI16/00041, MV19/00058 to E.A.]. Funding for open access charge: Internal funds of Saarland University.

Conflict of interest statement. None declared.

REFERENCES

1. Kern,F., Fehlmann,T., Violich,I., Alsop,E., Hutchins,E., Kahraman,M., Grammes,N.L., Guimarães,P., Backes,C., Poston,K.L. *et al.* (2021) Deep sequencing of sncRNAs reveals hallmarks and regulatory modules of the transcriptome during Parkinson's disease progression. *Nat. Aging*, **1**, 309–322.
2. Huntzinger,E. and Izaurralde,E. (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat. Rev. Genet.*, **12**, 99–110.
3. Xu,W., San Lucas,A., Wang,Z. and Liu,Y. (2014) Identifying microRNA targets in different gene regions. *BMC Bioinform.*, **15**, S4.
4. Loffreda,A., Rigamonti,A., Barabino,S.M. and Lenzken,S.C. (2015) RNA-binding proteins in the regulation of miRNA activity: a focus on neuronal functions. *Biomolecules*, **5**, 2363–2387.
5. O'Brien,J., Hayder,H., Zayed,Y. and Peng,C. (2018) Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Front. Endocrinol. (Lausanne)*, **9**, 402.
6. Stavast,C.J. and Erkeland,S.J. (2019) The non-canonical aspects of microRNAs: many roads to gene regulation. *Cells*, **8,** 1465 .
7. Bartel,D.P. (2018) Metazoan microRNAs. *Cell*, **173**, 20–51.
8. Kern,F., Krammes,L., Danz,K., Diener,C., Kehl,T., Kuchler,O., Fehlmann,T., Kahraman,M., Rheinheimer,S., Aparicio-Puerta,E. *et al.* (2021) Validation of human microRNA target pathways enables evaluation of target prediction tools. *Nucleic Acids Res.*, **49**, 127–144.
9. Hamberg,M., Backes,C., Fehlmann,T., Hart,M., Meder,B., Meese,E. and Keller,A. (2016) miRTargetLink—miRNAs, genes and interaction networks. *Int. J. Mol. Sci.*, **17**, 564.
10. Lukasik,A., Wojcikowski,M. and Zielenkiewicz,P. (2016) Tools4miRs: one place to gather all the tools for miRNA analysis. *Bioinformatics*, **32**, 2722–2724.
11. Paraskevopoulou,M.D., Georgakilas,G., Kostoulas,N., Vlachos,I.S., Vergoulis,T., Reczko,M., Filippidis,C., Dalamagas,T. and Hatzigeorgiou,A.G. (2013) DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res.*, **41**, W169–W173.
12. Vlachos,I.S., Zagganas,K., Paraskevopoulou,M.D., Georgakilas,G., Karagkouni,D., Vergoulis,T., Dalamagas,T. and Hatzigeorgiou,A.G. (2015) DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res.*, **43**, W460–W466.
13. Jung,D., Kim,B., Freishtat,R.J., Giri,M., Hoffman,E. and Seo,J. (2015) miRTarVis: an interactive visual analysis tool for microRNA–mRNA expression profile data. *BMC Proc.*, **9**, S2.
14. Licursi,V., Conte,F., Fiscon,G. and Paci,P. (2019) MIENTURNET: an interactive web tool for microRNA-target enrichment and network-based analysis. *BMC Bioinform.*, **20**, 545.
15. Giroux,P., Bhajun,R., Segard,S., Picquenot,C., Charavay,C., Desquilles,L., Pinna,G., Ginestier,C., Denis,J., Cherradi,N. *et al.* (2020) miRViz: a novel webserver application to visualize and interpret microRNA datasets. *Nucleic Acids Res.*, **48**, W252–W261.
16. Chang,L., Zhou,G., Soufan,O. and Xia,J. (2020) miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. *Nucleic Acids Res.*, **48**, W244–W251.
17. Fan,Y., Siklenka,K., Arora,S.K., Ribeiro,P., Kimmins,S. and Xia,J. (2016) miRNet: dissecting miRNA–target interactions and functional associations through network-based visual analysis. *Nucleic Acids Res.*, **44**, W135–W141.
18. Kowarsch,A., Preusse,M., Marr,C. and Theis,F.J. (2011) miTALOS: analyzing the tissue-specific regulation of signaling pathways by human and mouse microRNAs. *RNA*, **17**, 809–819.
19. Preusse,M., Theis,F.J. and Mueller,N.S. (2016) miTALOS v2: analyzing tissue specific microRNA function. *PLoS One*, **11**, e0151771.
20. Nam,S., Li,M., Choi,K., Balch,C., Kim,S. and Nephew,K.P. (2009) MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Res.*, **37**, W356–W362.
21. Muniategui,A., Nogales-Cadenas,R., Vazquez,M., Aranguren,X.L., Agirre,X., Luttun,A., Prosper,F., Pascual-Montano,A. and Rubio,A. (2012) Quantification of miRNA–mRNA interactions. *PLoS One*, **7**, e30766.
22. Laczny,C., Leidinger,P., Haas,J., Ludwig,N., Backes,C., Gerasch,A., Kaufmann,M., Vogel,B., Katus,H.A., Meder,B. *et al.* (2012) miRTrail–a comprehensive webserver for analyzing gene and miRNA patterns to enhance the understanding of regulatory mechanisms in diseases. *BMC Bioinform.*, **13**, 36.
23. Xie,G.Y., Xia,M., Miao,Y.R., Luo,M., Zhang,Q. and Guo,A.Y. (2020) FFLtool: a web server for transcription factor and miRNA feed forward loop analysis in human. *Bioinformatics*, **36**, 2605–2607.
24. Kehl,T., Kern,F., Backes,C., Fehlmann,T., Stöckel,D., Meese,E., Lenhof,H.-P. and Keller,A. (2019) miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database. *Nucleic Acids Res.*, **48**, D142–D147.
25. Huang,H.-Y., Lin,Y.-C.-D., Li,J., Huang,K.-Y., Shrestha,S., Hong,H.-C., Tang,Y., Chen,Y.-G., Jin,C.-N., Yu,Y. *et al.* (2019) miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic Acids Res.*, **48**, D148–D154.
26. Tokar,T., Pastrello,C., Rossos,A.E.M., Abovsky,M., Hauschild,A.-C., Tsay,M., Lu,R. and Jurisica,I. (2017) mirDIP 4.1—integrative database of human microRNA target predictions. *Nucleic Acids Res.*, **46**, D360–D370.
27. Chen,Y. and Wang,X. (2020) miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res.*, **48**, D127–D131.
28. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
29. Kern,F., Fehlmann,T., Solomon,J., Schwed,L., Grammes,N., Backes,C., Van Keuren-Jensen,K., Craig,D.W., Meese,E. and Keller,A. (2020) miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. *Nucleic Acids Res.*, **48**, W521–W528.
30. Gerstner,N., Kehl,T., Lenhof,K., Muller,A., Mayer,C., Eckhart,L., Grammes,N.L., Diener,C., Hart,M., Hahn,O. *et al.* (2020) GeneTrail 3: advanced high-throughput enrichment analysis. *Nucleic Acids Res.*, **48**, W515–W520.
31. Xin,J., Mark,A., Afrasiabi,C., Tsueng,G., Juchler,M., Gopal,N., Stupp,G.S., Putman,T.E., Ainscough,B.J., Griffith,O.L. *et al.* (2016) High-performance web services for querying gene and variant annotation. *Genome Biol.*, **17**, 91.
32. Franz,M., Lopes,C.T., Huck,G., Dong,Y., Sumer,O. and Bader,G.D. (2016) Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*, **32**, 309–311.
33. Fehlmann,T., Lehallier,B., Schaum,N., Hahn,O., Kahraman,M., Li,Y., Grammes,N., Geffers,L., Backes,C., Balling,R. *et al.* (2020) Common diseases alter the physiological age-related blood microRNA profile. *Nat. Commun.*, **11**, 5958.
34. Madrer,N. and Soreq,H. (2020) Cholino-ncRNAs modulate sex-specific- and age-related acetylcholine signals. *FEBS Lett.*, **594**, 2185–2198.
35. Ludwig,N., Leidinger,P., Becker,K., Backes,C., Fehlmann,T., Pallasch,C., Rheinheimer,S., Meder,B., Stahler,C., Meese,E. *et al.* (2016) Distribution of miRNA expression across human tissues. *Nucleic Acids Res.*, **44**, 3865–3877.
36. Palmieri,V., Backes,C., Ludwig,N., Fehlmann,T., Kern,F., Meese,E. and Keller,A. (2018) IMOTA: an interactive multi-omics tissue atlas for the analysis of human miRNA-target interactions. *Nucleic Acids Res.*, **46**, D770–D775.
37. Chen,Z., Wang,H., Zhong,J., Yang,J., Darwazeh,R., Tian,X., Huang,Z., Jiang,L., Cheng,C., Wu,Y. *et al.* (2019) Significant changes in circular RNA in the mouse cerebral cortex around an injury site after traumatic brain injury. *Exp. Neurol.*, **313**, 37–48.
38. Martyniuk,C.J., Martinez,R., Kostyniuk,D.J., Mennigen,J.A. and Zubcevic,J. (2020) Genetic ablation of bone marrow beta-adrenergic receptors in mice modulates miRNA-transcriptome networks of neuroinflammation in the paraventricular nucleus. *Physiol. Genomics*, **52**, 169–177.
39. Seifer,B.J., Su,D. and Taylor,H.S. (2017) Circulating miRNAs in murine experimental endometriosis. *Reprod. Sci.*, **24**, 376–381.
40. Latta,L., Ludwig,N., Krammes,L., Stachon,T., Fries,F.N., Mukwaya,A., Szentmary,N., Seitz,B., Wowra,B., Kahraman,M. *et al.* (2021) Abnormal neovascular and proliferative conjunctival phenotype in limbal stem cell deficiency is associated with altered

microRNA and gene expression modulated by PAX6 mutational status in congenital aniridia. *Ocul. Surf.*, **19**, 115–127.

41. Fehlmann,T., Backes,C., Kahraman,M., Haas,J., Ludwig,N., Posch,A.E., Wurstle,M.L., Hubenthal,M., Franke,A., Meder,B. *et al.* (2017) Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic Acids Res.*, **45**, 8731–8744.
42. Winek,K., Lobentanzer,S., Nadorp,B., Dubnov,S., Dames,C., Jagdmann,S., Moshitzky,G., Hotter,B., Meisel,C., Greenberg,D.S. *et al.* (2020) Transfer RNA fragments replace microRNA regulators of the cholinergic poststroke immune blockade. *Proc. Natl. Acad. Sci. USA*, **117**, 32606–32616.

Next Generation Sequencing Analysis of Total Small Noncoding RNAs from Low Input RNA from Dried Blood Sampling

PMID: 30198258 DOI: 10.1021/acs.analchem.8b03557

Next Generation Sequencing Analysis of Total Small Noncoding RNAs from Low Input RNA from Dried Blood Sampling

PMID: 30198258 DOI: 10.1021/acs.analchem.8b03557

Next Generation Sequencing Analysis of Total Small Noncoding RNAs from Low Input RNA from Dried Blood Sampling

PMID: 30198258 DOI: 10.1021/acs.analchem.8b03557

Next Generation Sequencing Analysis of Total Small Noncoding RNAs from Low Input RNA from Dried Blood Sampling

PMID: 30198258 DOI: 10.1021/acs.analchem.8b03557

Next Generation Sequencing Analysis of Total Small Noncoding RNAs from Low Input RNA from Dried Blood Sampling

PMID: 30198258 DOI: 10.1021/acs.analchem.8b03557

Next Generation Sequencing Analysis of Total Small Noncoding RNAs from Low Input RNA from Dried Blood Sampling

PMID: 30198258 DOI: 10.1021/acs.analchem.8b03557

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-19665-1 OPEN

Check for updates

Common diseases alter the physiological age-related blood microRNA profile

Tobias Fehlmann[1], Benoit Lehallier[2], Nicholas Schaum[2], Oliver Hahn[2], Mustafa Kahraman[1], Yongping Li[1], Nadja Grammes[1], Lars Geffers[3], Christina Backes[1], Rudi Balling[3,4,5], Fabian Kern[1], Rejko Krüger[3,4,5], Frank Lammert[6], Nicole Ludwig[7], Benjamin Meder[8], Bastian Fromm[9], Walter Maetzler[10], Daniela Berg[10], Kathrin Brockmann[11], Christian Deuschle[11], Anna-Katharina von Thaler[11], Gerhard W. Eschweiler[12], Sofiya Milman[13], Nir Barziliai[13], Matthias Reichert[6], Tony Wyss-Coray[2], Eckart Meese[7] & Andreas Keller[1,2,14]✉

Aging is a key risk factor for chronic diseases of the elderly. MicroRNAs regulate post-transcriptional gene silencing through base-pair binding on their target mRNAs. We identified nonlinear changes in age-related microRNAs by analyzing whole blood from 1334 healthy individuals. We observed a larger influence of the age as compared to the sex and provide evidence for a shift to the 5′ mature form of miRNAs in healthy aging. The addition of 3059 diseased patients uncovered pan-disease and disease-specific alterations in aging profiles. Disease biomarker sets for all diseases were different between young and old patients. Computational deconvolution of whole-blood miRNAs into blood cell types suggests that cell intrinsic gene expression changes may impart greater significance than cell abundance changes to the whole blood miRNA profile. Altogether, these data provide a foundation for understanding the relationship between healthy aging and disease, and for the development of age-specific disease biomarkers.

[1] Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany. [2] Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA 94305, USA. [3] Luxembourg Center for Systems Biomedicine, 4362 Esch-sur-Alzette, Luxemburg. [4] Transversal Translational Medicine, Luxembourg Institute of Health (LIH), 1445 Strassen, Luxembourg. [5] Parkinson Research Clinic, Centre Hospitalier de Luxembourg, 1210 Luxembourg, Luxemburg. [6] Internal Medicine, Saarland University, 66421 Homburg, Germany. [7] Human Genetics, Saarland University, 66421 Homburg, Germany. [8] Internal Medicine, University Hospital Heidelberg, 69120 Heidelberg, Germany. [9] Department of Molecular Biosciences, Stockholm University, 11418 Stockholm, Sweden. [10] Department of Neurology, Christian-Albrechts-Universität zu Kiel, 24105 Kiel, Germany. [11] TREND study center Tübingen, Tübingen, Germany. [12] Geriatric Center and the Department of Psychiatry and Psychotherapy, University Hospital Tübingen, 72076 Tübingen, Germany. [13] The Institute for Aging Research, Albert Einstein College of Medicine, New York, NY 10461, USA. [14] Center for Bioinformatics, Saarland Informatics Campus, Saarland University, 66123 Saarbrücken, Germany. ✉email: andreas.keller@ccb.uni-saarland.de

Aging is the leading risk factor for cardiovascular disease, diabetes, dementias including Alzheimer's disease, and cancer, together accounting for the majority of debilitating illnesses worldwide[1]. Uncovering common therapeutic targets to prevent or treat these diseases simultaneously could convey enormous benefits to quality of life. It is therefore essential to model the cellular processes culminating in these diverse maladies through an understanding of the molecular changes underlying healthy and pathological aging[2]. Accordingly, a variety of molecular studies have been conducted in humans, including whole genome analysis of long-lived individuals[3], transcriptomic analyses of tissues[4], plasma proteomic profiling[5], and the exploration of epigenetic control of aging clocks[6]. Recent organism-wide RNA-sequencing data of whole organs and single cells across the mouse lifespan provide an important and complementary database from which to build models of molecular cascades in aging[7,8].

Functional improvement of aged tissues has been achieved by an expanding number of techniques, ranging from dietary restriction[9] to senescent cell elimination and partial cellular reprogramming. This also includes heterochronic parabiosis, in which an old mouse is exposed to a young circulatory system. These experiments point to systemic factors in the blood of young mice that modulate organ function in aged animals[10,11]. Indeed, the list of individual plasma proteins with beneficial or detrimental effects on different tissues is growing. It is likely, however, that each plasma protein interacts with complex intracellular regulatory networks, and that alterations to such networks are a key component of aging and rejuvenation.

Non-coding ribonucleic acids like microRNAs (miRNAs) represent essential players governing these molecular cascades, and they show a highly complex spectrum of biological actions[12–14]. MicroRNAs are a family of short single stranded non-coding RNA molecules that regulate post-transcriptional gene silencing through base-pair binding on their target mRNAs[13], thereby regulating most if not all cellular and biological processes[15]. Yet, their involvement in the aging process and rejuvenation of aged tissues is often ignored by transcriptomic studies and is thus largely uncharacterized. A single microRNA targets not only untranslated regions (UTRs) of numerous genes, but it can also bind multiple sites within a single UTR[16]. Similarly, a UTR of a specific gene can contain target sites for dozens or even hundreds of miRNAs. Since their discovery, miRNA changes have been reported for almost all cancers and many non-cancer diseases like Alzheimer's disease[17,18], multiple sclerosis[19], or heart failure[20]. And although relatively sparse, several studies have measured aging miRNA expression in different human and primate tissues[21]. For example, Somel and co-workers analyzed miRNA, mRNA, and protein expression linked to development and aging in the prefrontal cortex of humans and rhesus macaques over the lifespan[22]. Likewise, changes of miRNA levels in aging human skeletal muscle have been characterized[23], as have miRNA levels in body fluids such as serum[24,25]. In whole blood, we previously reported a significant number of age-related miRNAs[26], and Huan and co-workers measured a selection of miRNAs by RT-qPCR in whole blood from over 5000 individuals from the Framingham Heart Study[27]. While these initial studies are intriguing, they can be limited by the use of discrete time points, incomplete lifespan coverage, limited cohort sizes, and incomplete miRNA panels.

Here, we performed a comprehensive characterization of all 2549 annotated miRNAs (miRBase V21) in 4393 whole blood samples from both sexes across the lifespan (30–90 years). To understand the relationship between healthy aging and disease, we included 1334 healthy controls (HC), 944 patients with Parkinson's disease (PD), 607 with heart diseases (HD), 586 with non-tumor lung diseases (NTLD), 517 with lung cancer (LC), and 405 with other diseases (OD) (Fig. 1a, b; Supplementary Data 1).

Results

miRNA profiles are stronger associated with the age as compared to the sex. We first sought to model healthy aging as a baseline for understanding disease. As males have shorter lifespans than females, and each sex suffers a different array of age-related diseases, we investigated the interplay between age and sex on blood miRNA profiles. Confirming our previous observation in a cohort of 109 individuals[26], we found that age has a more pronounced influence than sex. In fact, 1568 miRNAs significantly correlated with age, but only 362 correlated with sex according to Benjamini–Hochberg adjusted p-values of the Wilcoxon Mann–Whitney test (Fig. 2a, b). While 231 miRNAs overlapped between these groups, this number was not significant (two-sided Fisher's exact test p-value of 0.35; Pearson's Chi-squared Test of 0.36), suggesting that, in general, those miRNAs changing with age are shared by both sexes, and those specific to one sex do not change with age. In consequence, the Spearman correlation coefficient (SC) of age-related changes between males and females was high (SC of 0.884, $p < 10^{-16}$, Fig. 2c).

We next sorted miRNAs by their correlation with age, regardless of their significance, and assigned each to one of 5 groups: strongly decreasing with age (cluster 1: 174 miRNAs, SC < -0.2), moderately decreasing (cluster 2: 382 miRNAs; $-0.2 < \mathrm{SC} < -0.1$), unaltered (cluster 3: 1451 miRNAs; $-0.1 < \mathrm{SC} < 0.1$), moderately increasing (cluster 4: 368 miRNAs; $0.1 < \mathrm{SC} < 0.2$), and strongly increasing (cluster 5: 174 miRNAs, $\mathrm{SC} > 0.2$) (Supplementary Data 2). As miRNAs regulate a diverse array of critical pathways[28], we performed microRNA enrichment analysis and annotation (miEAA) on this sorted list, thereby calculating a running sum of miRNAs associated with each of ~14,000 biochemical categories and pathways. We revealed a remarkable disequilibrium between the number of pathways related to downregulated miRNAs (76 pathways) and upregulated miRNAs (620 pathways; adjusted p-value < 0.05; Supplementary Data 3). This is even more striking considering the number of miRNAs increasing or decreasing did not differ significantly (556 with $\mathrm{SC} < -0.1$; 542 with $\mathrm{SC} > 0.1$), and suggests that miRNAs increasing with age have a higher functional relevance. Reassuringly, for miRNAs decreasing with age we found "Negative Correlated with Age" ($p = 4 \times 10^{-10}$) among the most significant categories (Fig. 2d). A large fraction of the top pathways regardless of the miRNA direction were enriched for brain function and neurodegeneration, including "Downregulated in Alzheimer's Disease" ($p = 10^{-5}$), "regulation of synaptic transmission" ($p = 0.028$), and "APP catabolic processes" ($p = 0.032$) (Fig. 2e, Supplementary Fig. 1a–l).

Although such linear correlation analyses can reveal meaningful biological features, the importance of nonlinear aging changes, such as those found for plasma proteins[5] and tissue gene expression, is becoming increasingly evident. We therefore aimed to use the high temporal resolution of the dataset to more thoroughly understand whole blood miRNA dynamics across the lifespan. We first plotted miRNA trajectories for each of the 5 clusters (Supplementary Fig. 2), confirming many miRNAs exhibit non-linear patterns. By comparing linear and nonlinear correlations for each, we uncovered nonlinear changes in 116 of the 1098 miRNAs altered with age, of which 90 decreased and 26 increased (Fig. 2f, g, Supplementary Data 4). A miEAA analysis highlighted a significant enrichment of miRNAs following nonlinear trajectories with aging in basically all human tissues[29] (Fig. 2h). This finding stands out considering the high degree of tissue specificity of miRNAs. We thus speculate that diseases

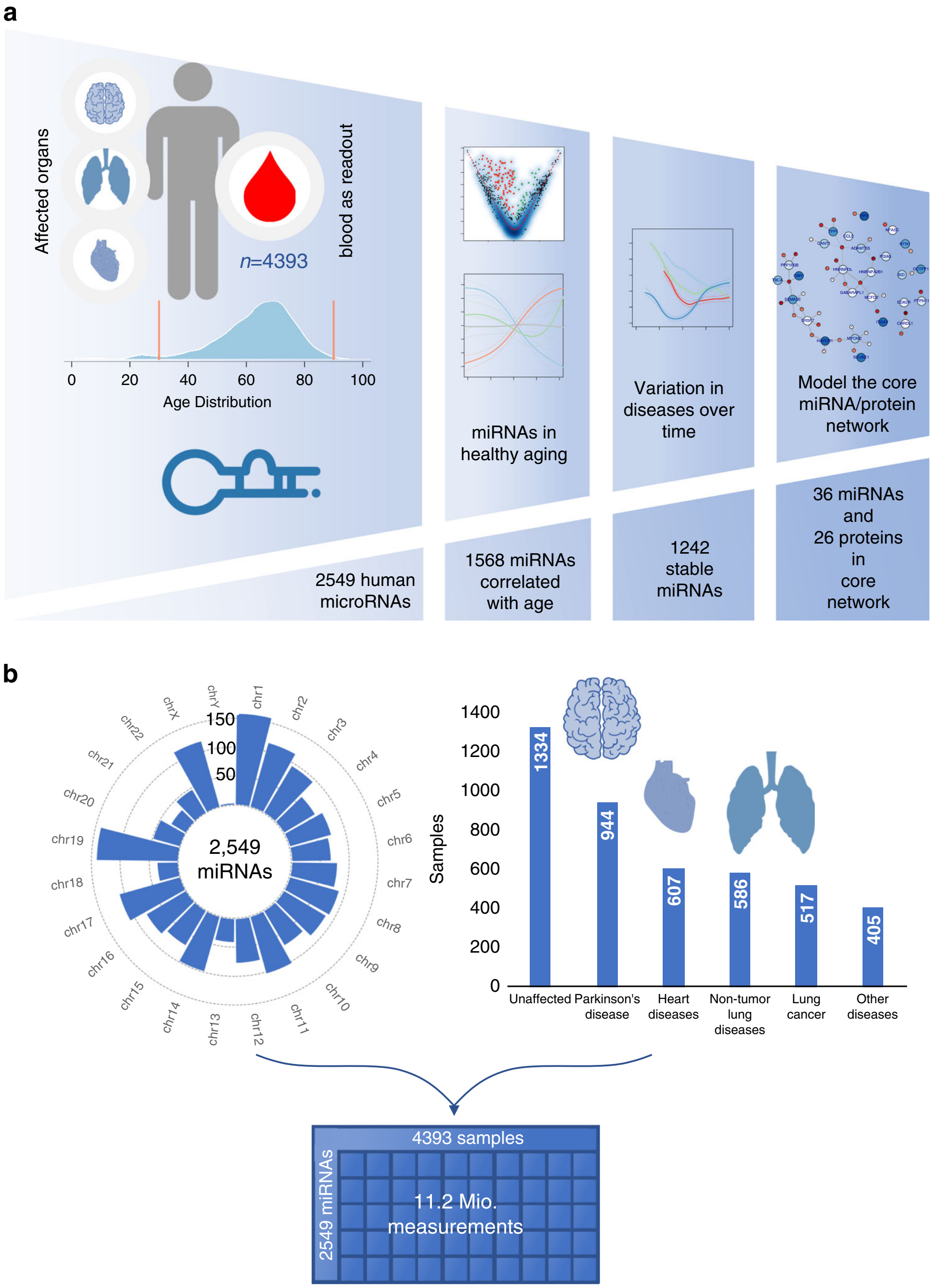
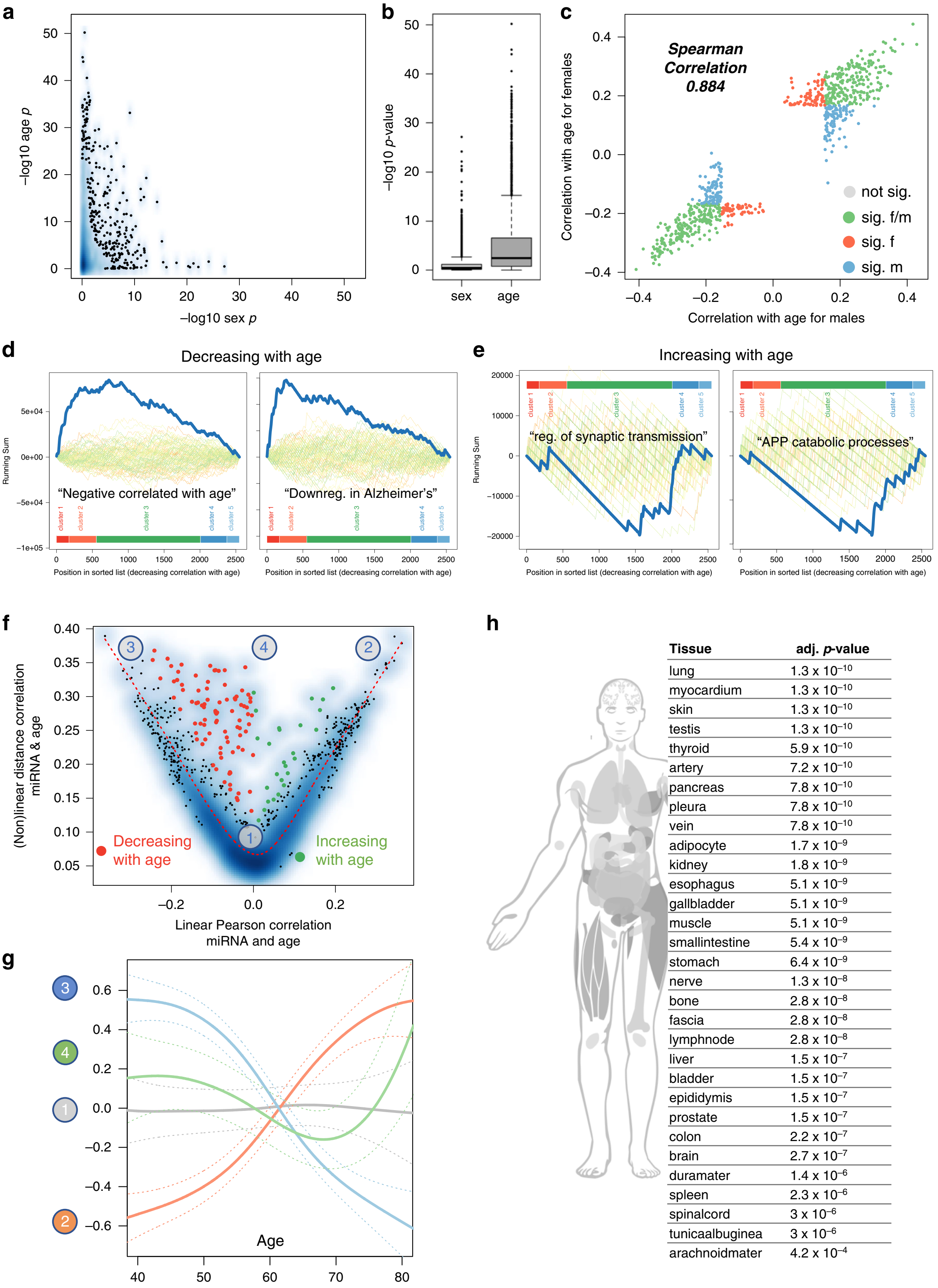


Fig. 1 Study characteristics. a Study set up and analysis workflow from high-throughput data to a specific aging network. The cohort consist of 4393 samples of which the age distribution is provided. For the 4393 samples genome wide miRNA screening using microarrays has been performed. The first analysis describes 1568 miRNAs that are correlated to age in healthy individuals. In the second step we identified disease specific miRNA changes with aging and finally define a set of 1242 miRNAs that are not affected by diseases. Finally, to model regulatory cascades in healthy aging we related the miRNA data to plasma proteins and identified a core aging network. **b** The circular plot shows the genome wide nature of our miRNA approach, all miRNAs from miRBase V21 were included in the experimental analysis. We measured 4393 samples for the abundance of these miRNAs, resulting in a 2549 times 4393 data table containing 11.2 million miRNA measurements that correspond to over 2×10^8 spots on the arrays.

affecting these organs might be associated with changes in blood miRNA profiles.

miRNA arm shifts are associated with aging. A shift in the expression of the 3' and 5' mature arm of miRNAs is observed between different tissues[30] tissues but also in healthy and diseased conditions such as cancer[31]. We speculated that likewise aging may affect the arm distribution and searched for respective arm shift events. Indeed, we observed a correlation of the arm specific expression in 40 cases (Supplementary Data 5). For 27 miRNAs (67.5%) we observed increasing 5' mature expression and decreasing 3' expression over age while in 13 cases 32.5% of cases the 3' form increased and the 5' form decreased. These results indicate a generally increasing 5' mature miRNA expression with aging. The largest absolute increase of 5' mature expression was identified for miR-6786. A miRSwitch analysis highlighted that usually the 3' form is dominating in H. sapiens



Tissue	adj. *p*-value
lung	1.3×10^{-10}
myocardium	1.3×10^{-10}
skin	1.3×10^{-10}
testis	1.3×10^{-10}
thyroid	5.9×10^{-10}
artery	7.2×10^{-10}
pancreas	7.8×10^{-10}
pleura	7.8×10^{-10}
vein	7.8×10^{-10}
adipocyte	1.7×10^{-9}
kidney	1.8×10^{-9}
esophagus	5.1×10^{-9}
gallbladder	5.1×10^{-9}
muscle	5.1×10^{-9}
smallintestine	5.4×10^{-9}
stomach	6.4×10^{-9}
nerve	1.3×10^{-8}
bone	2.8×10^{-8}
fascia	2.8×10^{-8}
lymphnode	2.8×10^{-8}
liver	1.5×10^{-7}
bladder	1.5×10^{-7}
epididymis	1.5×10^{-7}
prostate	1.5×10^{-7}
colon	2.2×10^{-7}
brain	2.7×10^{-7}
duramater	1.4×10^{-6}
spleen	2.3×10^{-6}
spinalcord	3×10^{-6}
tunicaalbuginea	3×10^{-6}
arachnoidmater	4.2×10^{-4}

with 5' dominance mostly in plasma samples. For the miRNA with the most decreasing 5' expression ratio (miR-4423) we found dominating 3' expression mostly in breast milk, the heart, testis, stem cells and blood cells. Our results thus suggest an altered ratio of the 3' to 5' mature expression ratio that might be attributed to or effect different tissues.

The association between age and miRNA expression is partially lost in diseases. Although the cellular and molecular degeneration of aging often instigates age-related disease, there are nonetheless elderly individuals who have lived entirely disease-free lives. We therefore asked what differentiates such healthy aging from aging resulting in disease. For each disease and healthy controls, we

Fig. 2 miRNAs dependency on age and gender. a Smoothed scatter plot of the two-tailed age and gender association *p*-value for 2549 miRNAs. *P*-values for the sex are computed using Wilcoxon Mann–Whitney test and for the Spearman Correlation via the asymptotic t approximation. The *p*-values are Benjamini–Hochberg adjusted. **b** Boxplot of the age and gender *p*-value from **a** for 2549 miRNAs. The box spans the 25% and 75% quantile, the solid horizontal line represents the median and the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. **c** Correlation of miRNAs with age in males and females. Gray dots: not significant; orange and blue dots: miRNAs significantly correlated with age only in males or females; green dots: miRNAs significantly correlated with age in males and females. **d** Results of the miRNA enrichment analysis. Colored curves in the background represent random permutations of miRNAs. The cluster membership is projected next to the order of miRNAs. The category "negative correlated with age" is highly significant and confirms our data in general. Also, the category "downregulated in AD" is enriched with miRNAs decreasing over age. **e** Regulation of synaptic transmission is among the categories being enriched in miRNAs going up with age. Moreover, APP catabolic processes is another category being enriched in miRNAs going up with age. **f** Linear Pearson correlation versus non-linear distance correlation for the association of age to miRNAs. Orange dots have a high non-linear correlation that is not explained by linear correlation and are decreasing with age, green dots have a high non-linear correlation that is not explained by linear correlation and are increasing with. The orange dotted line represents a smoothed spline and the four numbers in gray circles represent the position of miRNAs where examples are provided in **g**. **g** Examples of correlation for miRNAs with age. (1) gray: no correlation; (2) orange dominantly positive linear correlation; (3) blue dominantly negative linear correlation; (4) non-linear correlation. Each solid line is a smoothing spline. **h** Tissue enrichment for the miRNAs that are correlated with age in a non-linear fashion. The human model has all organs highlighted in gray that are significantly enriched. The table on the right lists the organs with corresponding *p*-values. *P*-values have been computed using the hypergeometric distribution and were adjusted for multiple testing using the Benjamini–Hochberg approach.

computed the Spearman correlation (SC) with age for all 2549 miRNAs (Fig. 3a, Supplementary Data 6). Overall, healthy controls reached the largest absolute SC, greater than twice that of the pooled disease cohort, and larger than any individual disease. Using an Analysis of variance, we found highly significant differences ($p < 2.2 \times 10^{-16}$) and a non-parametric Wilcoxon Mann–Whitney test confirmed the significant differences of absolute Spearman correlation in healthy versus diseased samples ($p < 2.2 \times 10^{-16}$). In line with these findings, samples from healthy individuals showed far more miRNAs with significant age correlations (Fig. 3b), suggesting that the presence of an age-related disease may disrupt healthy aging miRNA profiles (Wilcoxon Mann–Whitney test $p < 2.2 \times 10^{-16}$). For example, lung cancer patients were enriched for a positive correlation with age, while miRNAs in patients with heart disease were enriched for negative correlation with age. We then compared the miRNA trajectories from the 5 clusters of healthy individuals to the matched clusters in diseased patients (Supplementary Fig. 2), and similarly, miRNAs from diseased individuals show far weaker aging patterns. This held true both when each disease was analyzed separately, or pooled.

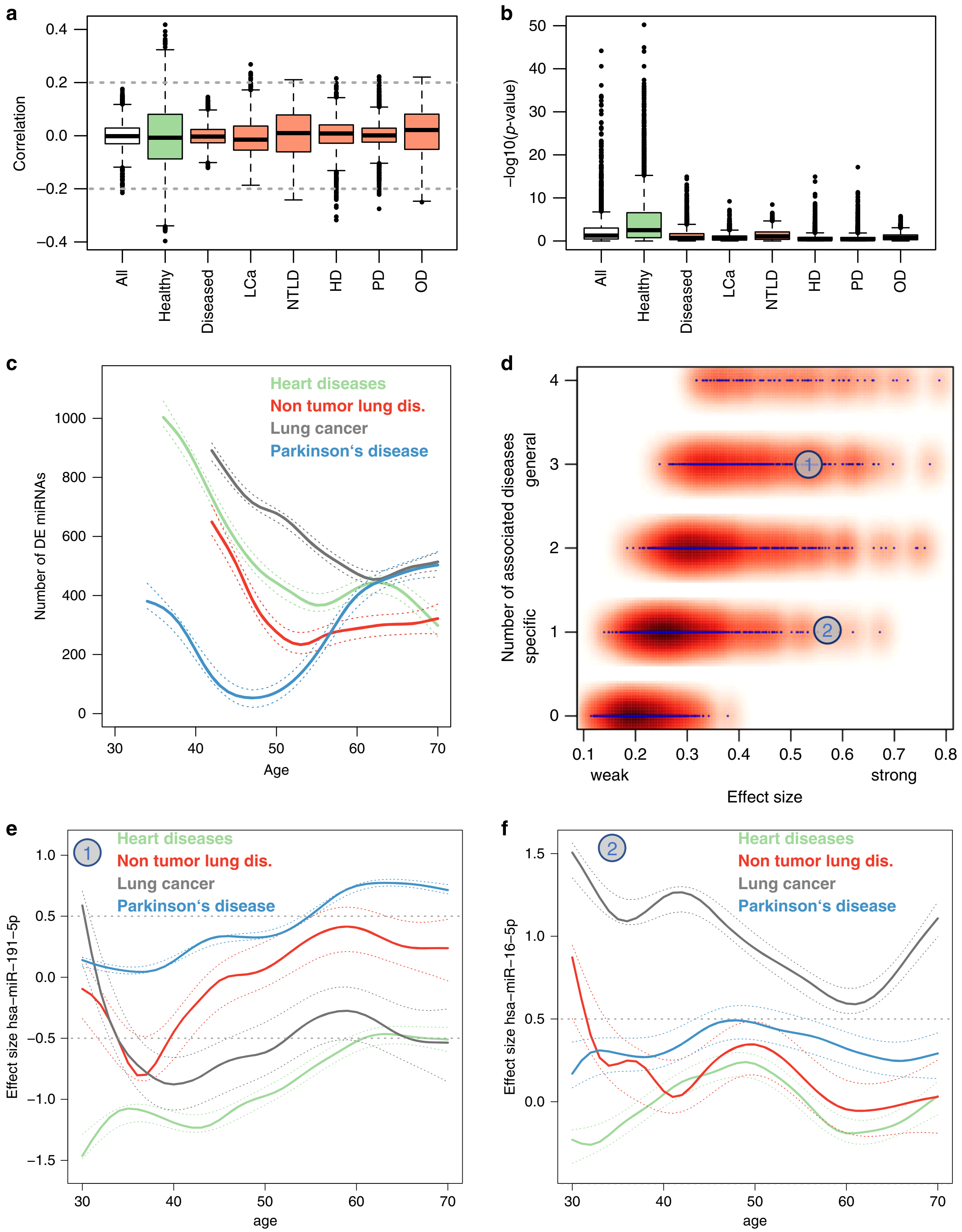
To determine the extent to which diseases affect miRNA abundance compared to healthy controls, we computed the number of differentially expressed miRNAs between cases and controls using a sliding window analysis. That is, we first compared diseased individuals aged 30–39 years to healthy individuals aged 30–39 years, then increased the window in increments of one year (31–40 years, 32–41 years, etc.) to the final window of 70–79 years (Fig. 3c, Supplementary Fig. 3a, b). As the age distribution varied between these groups, we excluded any window in which there were fewer than 20 disease cases and 20 healthy controls. Interestingly, for all diseases the number of differentially expressed miRNAs was high in young adults but decreased sharply into middle age, plateauing around age 60 for lung cancer and 50 for non-tumor lung diseases. Heart diseases largely plateaued by the early 50s. Parkinson's disease (PD), on the other hand, reached a minimum around age 47 before sharply increasing. With the exception of PD, these data show that aged healthy and diseased individuals are more similar than younger healthy and diseased individuals, perhaps suggesting that aged healthy individuals share some phenotypic characteristics of heart and lung disease.

We next asked if these diseases shared any miRNA alterations, and surprisingly we found that those miRNAs most commonly dysregulated were also those with the largest effect size (Fig. 3d). These pan-disease miRNAs included miR-191-5p (Fig. 3e), which targets mRNAs involved in cellular senescence[28]. We also observed disease-specific miRNAs like miR-16-5p, which targets the PI3K-Akt signaling pathway and microRNAs involved in lung cancer[28]. In summary, miRNA expression seems to be orchestrated in healthy aging with a loss of regulation in disease. In addition to disease-specific miRNAs, there appears to be a group of pan-disease miRNAs that change in a distinct manner. We thus asked on the specificity of biomarkers for diseases, especially in an age dependent context.

Distinct miRNA biomarker sets exist in young and old patients. The previous analyses of biomarkers in diseases were largely quantitative, i.e., we computed the number of dysregulated miRNAs in diseases for young and old patients. Here, we set to evaluate changes in the miRNA sets for young and old patients in the diseases. In this context we made use of the dimension reduction and visualization capabilities of self-organizing maps (SOMs). First, we considered the effect sizes of miRNAs for the two most global comparisons, i.e., healthy controls versus diseases and old (60–79 years) versus young (30–59 years) individuals. The heat map representation for the healthy versus disease comparison (Fig. 4a) and for young versus old individuals (Fig. 4b) highlights distinct patterns for the two comparisons and indicates that the aging miRNAs are different from the general disease miRNAs. This analysis however calls for a disease specific consideration. To this end we computed for each of the four diseases biomarkers in old and young patients using again the effect size as performance indicator and the self-organizing map analysis followed by a hierarchical clustering (Fig. 4c). While the cluster heat maps identify larger differences between the disease biomarker sets as compared to young and old biomarkers, also the sets within the diseases vary greatly (Fig. 4c). In line with the previous analyses we observe larger effects for all diseases but PD in young patients (middle row of Fig. 4c). In old patients, the respective biomarkers are partially lost. Only in few cases new biomarkers emerge in old patients that are not present in young patients. As the full annotation of the SOM grid shows, each SOM cell has an average of 8 cluster members with a standard deviation of 3.5 miRNAs (Supplementary Data 7). The distribution largely corresponds to a normal distribution, only four cells (24, 62, 81, and 82 in Supplementary Data 7) contain more than 15 miRNAs (mean + two times the standard deviation).

The previous analyses suggest distinct biomarker sets for young and old patients in the different diseases. As a consequence, future biomarker test based on miRNAs may not only be established for a disease but for a specific age range of patients with that disease.

Given the results from this and the previous section we computed for each miRNA in each disease and each age window



the effect size (Supplementary Data 8). The respective supplementary data provides detailed insights in how specific certain miRNAs are for specific diseases and age ranges and can support ongoing biomarker studies significantly.

All results obtained so far argue for a strong immunological component of the miRNAs, and as a consequence of miRNA target networks. Since our experimental system profiles whole blood miRNAs, we set out to determine the cellular origin by computational deconvolution.

White blood cells are the major repository of miRNAs in whole blood. Circulating immune cells have been implicated in aging and a variety of age-related diseases, and one of the most

Fig. 3 Diseases miRNAs are affected by age effects. a Boxplot of the Spearman correlation coefficient for each miRNA to all samples, healthy individuals, and patients. Group sizes: $n_{HC} = 1334$, $n_{PD} = 944$, $n_{HD} = 607$, n_{NTLD}, $n_{LC} = 517$, $n_{OD} = 405$. The box spans the 25% and 75% quantile, the solid horizontal line represents the median and the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. **b** Boxplot of *p*-values for the Spearman correlation coefficient of each miRNA to all samples, healthy individuals, and patients from **a**. Group sizes: $n_{HC} = 1334$, $n_{PD} = 944$, $n_{HD} = 607$, n_{NTLD}, $n_{LC} = 517$, $n_{OD} = 405$. The box spans the 25% and 75% quantile, the solid horizontal line represents the median and the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. The *p*-values have been computed via the asymptotic t approximation. **c** Number of deregulated miRNAs in disease groups depending on different ages in a sliding window analysis. Each solid line is a smoothing spline (green–heart diseases; red–non tumor lung diseases; gray–lung cancer; blue–Parkinson's disease). The areas represent the 95% confidence intervals. For all disease groups, the number of deregulated miRNAs decreases with age while it increases for Parkinson's Disease. **d** Smoothed scatterplot showing the average effect size per miRNA dependent on the number of diseases where the miRNA is associated with. In the lower right corner (the *y*-axis value of 1) the specific miRNAs with high effect sizes can be found. In the upper right corner, miRNAs with high effect sizes independent of the disease are located. The two numbers represent the location of the examples provided in **e** and **f**. **e** Example of a miRNA that is downregulated in heart diseases of younger patients, upregulated in older Parkinson's patients and not deregulated in lung diseases. Each solid line is a smoothing spline (green–heart diseases; red–non tumor lung diseases; gray–lung cancer; blue–Parkinson's disease). The areas represent the 95% confidence intervals. **f** Example of a miRNA from the lower right part of Fig. 3d. The miRNA is significant upregulated in lung cancer independent of age but basically not associated with other diseases. Color codes of panels **c**, **e**, and **f** are matched. Each solid line is a smoothing spline (green–heart diseases; red–non tumor lung diseases; gray–lung cancer; blue–Parkinson's disease). The areas represent the 95% confidence intervals.

common diagnostic tests for disease is blood cell profiling. Since miRNAs are known to be enriched in different blood cell types[32], we performed computational deconvolution of the whole blood miRNA profile, thereby grouping miRNAs by their predicted cell type(s) of origin (Fig. 5a). A total of 196 miRNAs were attributed to one specific cell type, including 127 miRNAs arising from monocytes. Most others derive from three or more types. For example, the largest group of 139 miRNAs stems from a combination of white and red blood cells (WBCs, RBCs), exosomes, and serum. And the third largest group of 119 is restricted to six types of WBCs. We also observed 31 miRNAs specific for NK cells, 19 specific for T-helper cells, 11 specific for B cells, and 8 specific for cytotoxic T cells. Overall, for those miRNAs for which we could assign a prospective origin, we found WBCs as the main contributor, even though they represent a substantially smaller volume of whole blood relative to RBCs and serum (Fig. 5b).

We then applied this analysis to those miRNAs changing with age, and found that those increasing appear to largely originate from B cells, monocytes, NK cells, cytotoxic T cells, and serum (Fig. 5c). In contrast, miRNAs decreasing with age are those enriched in neutrophils, T helper cells, and RBCs. These data indicate shifts in aging miRNA trajectories of specific blood cell types (Supplementary Fig. 4). Interestingly, for the above cell types, known age-related abundance changes largely follow opposite trends: lymphocytes generally decrease with age while neutrophils increase with age[33]. This suggests that cell-intrinsic gene expression changes age may significantly contribute to the observed whole blood miRNA profiles.

miRNAs associated with healthy aging regulate the expression of plasma proteins. An increasing body of evidence points to functional roles of systemic plasma proteins in aging and disease[5]. These proteins may represent downstream targets of blood-borne miRNAs. We thus compared our data to a recent dataset of plasma proteins associated with age in healthy individuals[5]. Because miRNAs regulate genes/proteins in a complex network, miRNAs increasing with age do not necessarily lead to downregulation of all target genes/proteins, and vice versa. Accordingly, we observed only one tendency: miRNAs decreasing with age (cluster 1 and 2) showed a slight enrichment for regulating proteins increasing with age (Fig. 6a). Considering such complexity, we employed a network-based analysis. Using all pairwise interactions of miRNAs with plasma proteins, we first computed a regulatory network (Fig. 6b). From this, we extracted a core network containing the top 5% downregulated miRNAs and the top 5% upregulated proteins, which was then further refined by including only experimentally validated miRNA/target genes mined from the literature[34], as well as miRNA/target pairs with an absolute Spearman correlation of at least 0.6. This stringent core network consists of 36 miRNAs targeting 26 genes (proteins) and splits into two larger and six smaller connected components (Fig. 6c). The densest part of the core network contains the axon guidance related semaphorin 3E (SEMA3E) and serine and arginine rich splicing factor 7 (SRSF7), which were targeted by 8 miRNAs including miR-6812-3p (Fig. 6d, Supplementary Fig. 5, Supplementary Fig. 6). Intriguingly, there exist no studies of this miRNA, but it targets SEMA3E in an age dependent manner with a Spearman correlation of −0.89.

Finally, we investigated the possible cell type of origin of these core miRNAs with deconvolution, which showed enrichment for neutrophils, monocytes, and B cells (Fig. 6e). We then used single-cell PBMC transcriptomic data to determine if SEMA3A or SRSF7 were expressed in these same cell types. While SEMA3E was not detectable, we did observe SRSF7 expression widely across cell types, including neutrophils, monocytes, and B cells (Fig. 6f, g). SRSF7 plays a role in alternative RNA processing and mRNA export, but has no known role in aging or neurodegeneration. Further research will be required to determine if miRNAs like miR-6812-3p do indeed target SRSF7 in these specific cell types, and to uncover if this process contributes to the global decline of transcription observed with age.

Discussion

Our analysis of blood derived microRNAs provides insights into changes in microRNA abundance dependent on age, sex, and disease. While age clearly contributes to expression changes, sex has a more modest effect. In fact, most miRNAs show a similar behavior over the lifespan in males and females. This is generally in-line with recent results in transcriptomic mouse tissue aging[7,8]. Generally, our results compare well to other studies of miRNAs in aging[27], especially regarding miRNAs increasing with age, for which we observe high concordance. There are, however, miRNAs decreasing with age reported in the previous study for which we did not find evidence. The most extreme examples are miR-30d-5p and miR-505-5p, both increasing with age in our study in the healthy individuals. Nonetheless, given different cohorts with different ethnicity, varying age range, and distinct profiling technologies, we observed remarkable concordance between the studies.

Here, we observed that diseases globally disturb the normal aging progression of blood-borne miRNAs. While linear

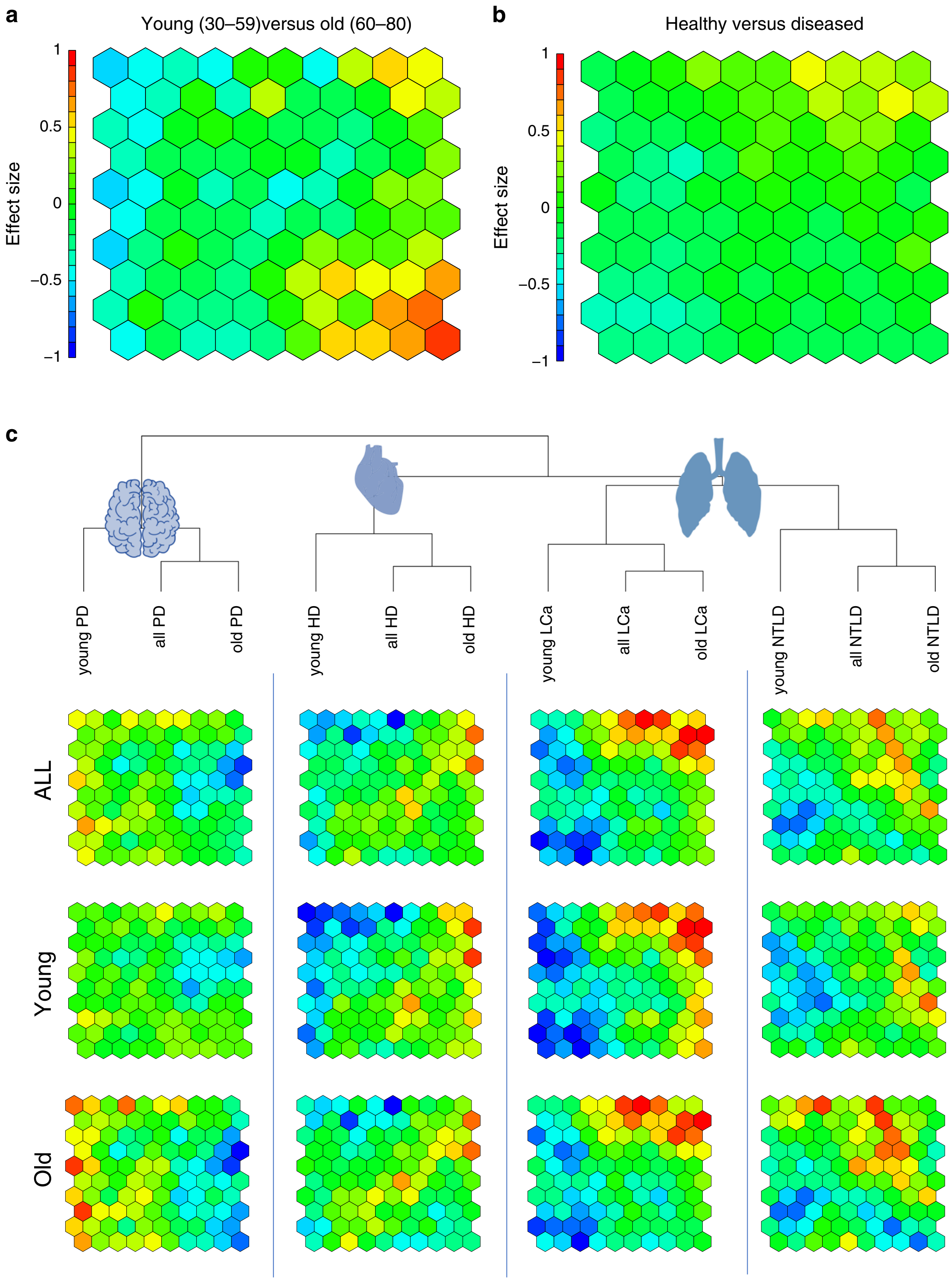


Fig. 4 Disease specificity of miRNA biomarkers. a Heat map representation of the SOM analysis as a 10 × 10 grid with 100 entries. Each cell contains at least one miRNA and up to 20 miRNAs. The full annotation of miRNAs to cells are provided in Supplementary Data 7). The cells are colored by the effect size of miRNAs for the comparison in old versus young. Red cells contain miRNAs with effect sizes >0.5 that are upregulated and in blue miRNAs that are downregulated with effect sizes <−0.5. **b** Same heat map as in **a** but colored for the difference in young versus old. The scale for the effect size has been kept the same as **a**. Thus fewer yellow/red, as well as blue spots indicate overall lower effect sizes. **c** Clustering of the SOM results in biomarkers for the four diseases and in all biomarkers independently of age, biomarkers for young patients and biomarker for old patients. The dendrogram has been computed from hierarchical clustering (complete linkage on the Euclidean distance). In all cases the biomarkers cluster by disease and not by age and the old biomarker set is closest to the all biomarker set while the young biomarker set has larger distances. Overall, NTLD and LCa markers are closest to each other, second closest are heart biomarkers and most different PD biomarkers. The SOM cells clearly highlight differences between biomarkers for diseases in young and old patients.

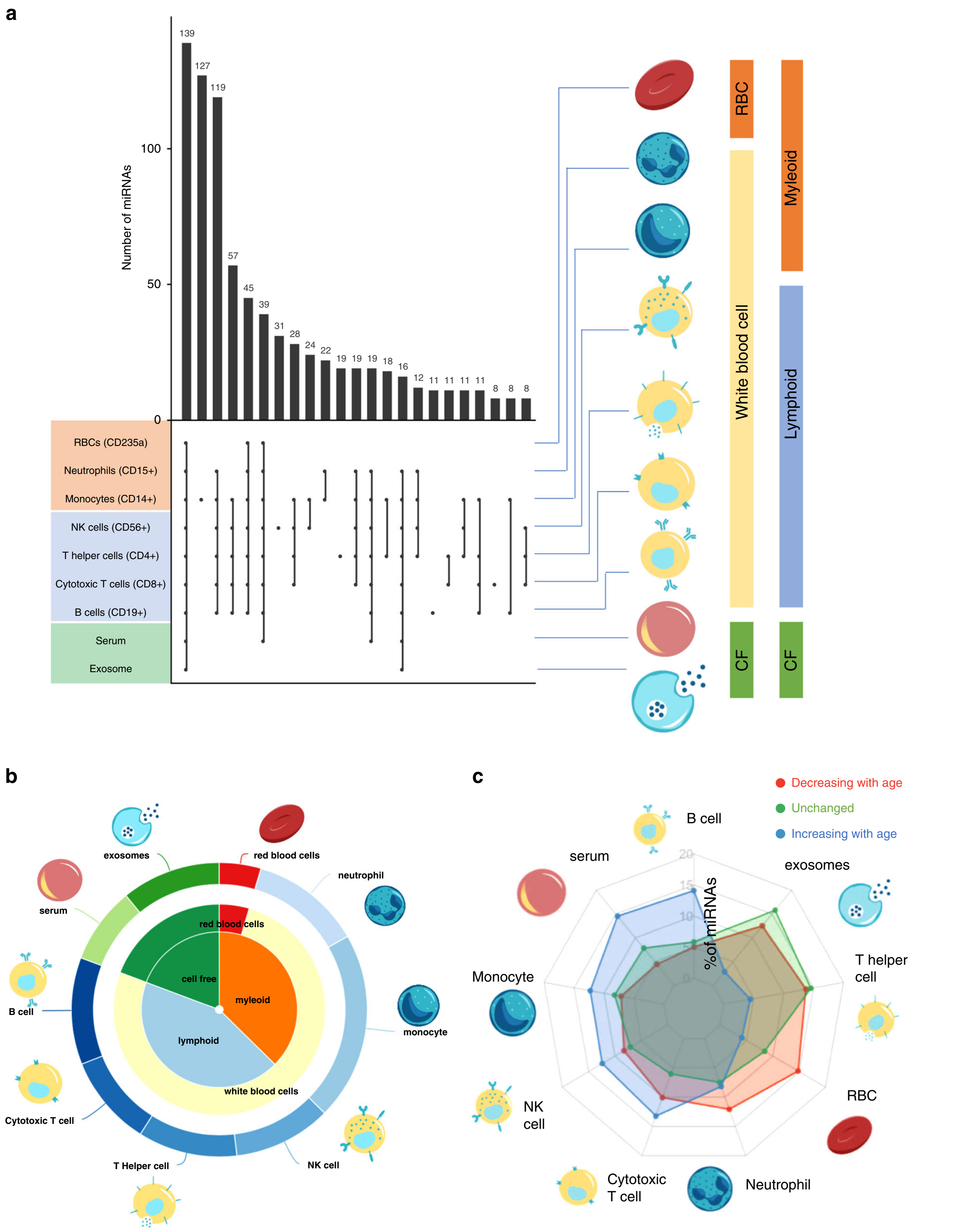
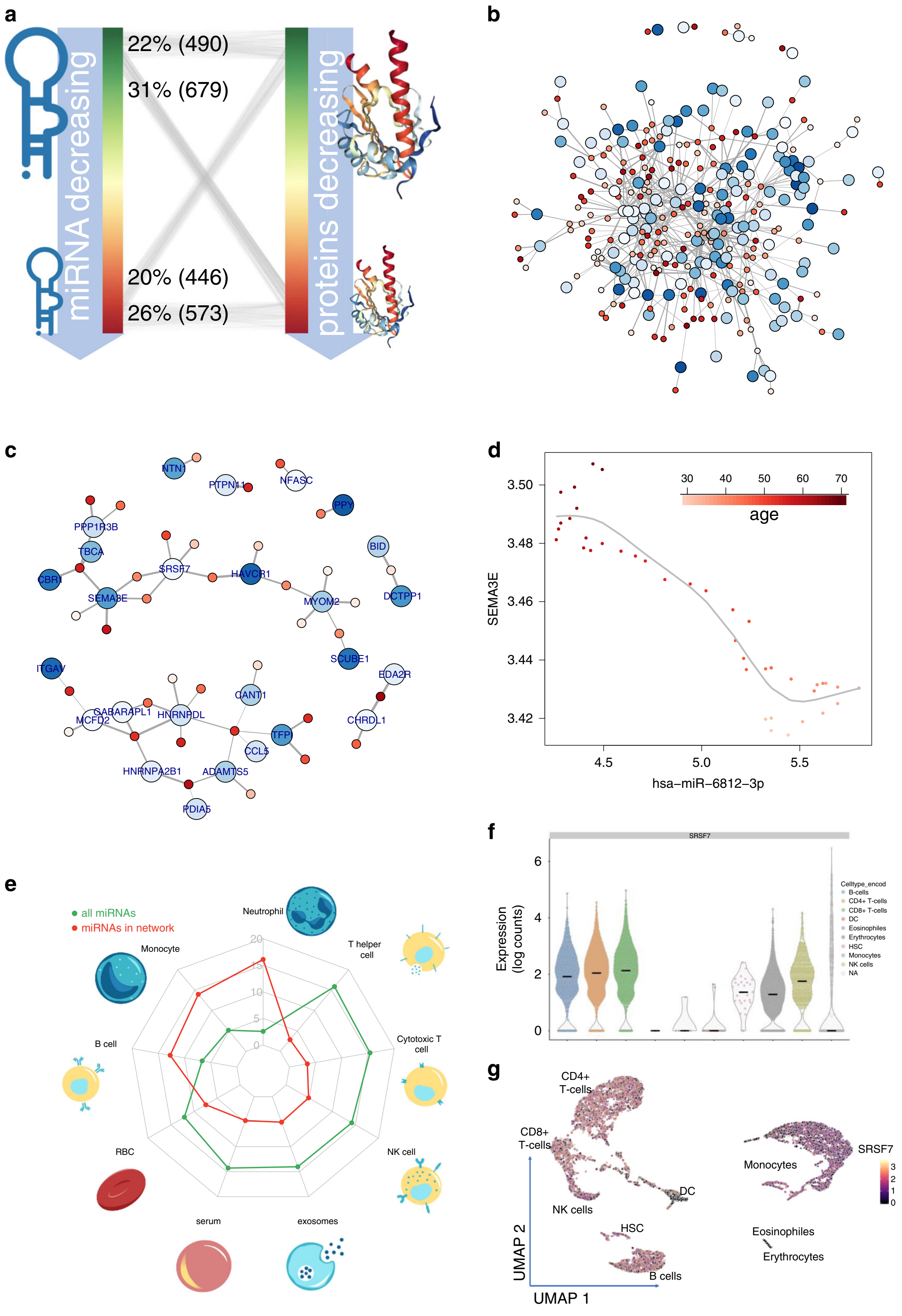


Fig. 5 Blood cell deconvolution. a The distribution of miRNAs in the different blood compounds. The rows are sorted by the blood compounds given on the right (RBC: red blood cell; CF: cell free), the columns are sorted according to a decreasing number of miRNAs. **b** Relative abundance of all miRNAs in the different blood compounds. **c** Distribution of miRNAs in cell types. The green distribution is the background and presents the relative composition of 1451 miRNAs in cluster 3. The blue distribution represents miRNAs increasing by age (cluster 4&5) and are enriched e.g., in B cells and serum. The red distribution represents miRNAs decreasing by age (cluster 1&2) and are enriched e.g., in neutrophils and RBCs.



modeling insufficiently explained changes with aging, distance correlation analysis identified 90 miRNAs that were decreasing and 26 that were increasing with age in a non-linear manner. These effects are, however, frequently not disease specific. If disease specific effects occur, they appear to establish themselves in given time windows throughout live. For example, lung and heart diseases show the largest effect sizes in the 4th to 5th decade of life, and Parkinson's disease showed the largest effect size in the 6th to 7th decade. All known biological factors including age, sex, and disease status together only explained part of the overall data variance. Thus, unknown biological variables and technical factors also contribute to miRNA abundance.

Our results underline not only the importance of age as a confounder in biomarker studies, but they show that age needs to be

Fig. 6 Age related miRNAs are correlated to age related proteins. a Correlation of miRNAs to proteins. miRNAs and proteins are sorted by increasing correlation with age. Thin lines are miRNA/gene interactions between top/bottom 10% of miRNAs and proteins. Numbers represent actual count of edges. **b**, **c** Core network. Proteins (larger nodes) are targeted by miRNAs (smaller nodes). Edge width correspond to the correlation. Blue nodes represent increase with age, red nodes decrease with age. The outer circles of the protein nodes indicate an expected an influence of the miRNAs leading to an increase with age. Panel **c** represents a more stringent version of the network from panel **b**. **d** One representative example of an edge from the network in **b**, **c**: SEMA3E and miR-6812-3p. Each dot represents all individuals in a time interval of 10 years, shifted between 30 and 70 years. SEMA3E is high expressed in older individuals while miR-6812-3p is low expressed (dark red points in the upper right corner). In young individuals the pattern is opposite (tale points in the lower right corner). **e** Blood cell compound distribution. miRNAs from the core network come from neutrophils, monocytes and B cells. **f** Violin plot of expression of SRSF7 in human blood cells. **g** UMAP embedding of human blood cells colored by expression of SRSF7.

incorporated into the definition of disease biomarkers. The age dependency of miRNA biomarkers may be even more prominent for acute diseases that are accompanied by drastic molecular changes. Furthermore, the influence of a disease on healthy aging miRNA patterns suggests that it is conceivable to define "negative biomarkers", i.e., biomarkers that reflect the degree of disturbance of a given time-dependent pattern typically found in healthy individuals.

miRNAs comprise complex gene regulatory networks, and it is essential to identify the miRNA-targets that are regulated by a given miRNA network. However, this is already a demanding task for static networks, and it becomes even more challenging when considering how entire networks change with age. We attempted to overcome this complexity and identify a core miRNA network by implementing several stringent criteria: (i) the inclusion of miRNA-gene pairs only if experimental evidence exists, (ii) limiting the analysis to the top 5% of miRNAs decreasing with age, and (iii) the top 5% of proteins increasing with age and with pairwise absolute correlation of at least 0.6. This stringent parameter set identified a core network of 36 miRNAs and 26 proteins organized in two larger hubs with eight miRNAs targeting the axon guidance related semaphorin 3E (SEMA3E) and serine and arginine rich splicing factor 7 (SRSF7). Semaphorines play crucial roles during the development of the nervous system, especially in the hippocampal formation[35]. SEMA3E suppresses endothelial cell proliferation and angiogenic capacity, and in complex with PlexinD1 it inhibits recruitment of pericytes in endothelial cells[36]. Since we did not detect SEMA3E mRNA expression in single blood cell data we also explored other sources such as the Genotype-Tissue Expression (GTEx) project[37]. But also in the GTEx data no expression for the gene was reported in bulk sequencing data. It thus remains unclear how or if these miRNAs directly or indirectly impact SEMA3E protein levels in plasma. In this context, low abundant fractions of the blood such as exosomes might play a role. However, SRSF7, which belongs to a protein family linking alternative RNA processing to mRNA export[38], is expressed across a variety of circulating immune cells. This is intriguing as no role in aging or neurodegeneration is known.

Often, different technologies are available for high-throughput studies. To characterize the complete miRNome, usually microarrays or high-throughput sequencing are used. The choice of the best technology depends both, on technical factors and on the underlying biological question to be addressed. We decided to use microarray technology mostly because of the high dynamic range of blood miRNAs. In whole blood, the majority of reads (90–95%) are matching to few (2–5) miRNAs[39]. While generally a depletion is feasible[40], it bears the risk to alter the profile of other miRNAs especially since it has to be tailored for the respective sequencing technology. To use microarrays has however also disadvantages. MicroRNAs are often modified and build so-called isomiRs and basically all human miRNAs express different isoforms[41]. Likewise, data from the Rigoutsos lab demonstrate the importance and presence of isomiRs[42]. To address the age specific expression of isomiRs, single nucleotide resolution is required. Improved library preparation and sequencing methods together with increasing read numbers per sample will likely allow for an in-depth characterization of isomiRs in challenging specimens such as whole blood.

Another aspect for respective studies is the underlying specimen type. A literature search reveals that for human miRNA biomarker studies mostly plasma, serum, and blood cells (either PBMCs or whole blood) are considered with a more recent trend towards exosomes. Since we are interested in the connection of miRNA expression and the immune system by analyzing multiple diseases[43] we measured blood cells. Different aspects can be used to provide an even more comprehensive systemic picture of miRNAs and aging. First, the cell free part of the blood is also correlated to miRNA aging[44,45]. One important aspect are vesicles. Cellular senescence for example contributes to age-dependent changes in circulating extracellular vesicle cargo[46]. Moreover, the differential loading of vesicles is correlated to different human diseases[47–49]. Likewise, for the cellular part, resolution can be increased. For example, the miRNomes could be investigated per blood cell type[50]. One challenge is in that the purification of the different cell types by different isolation techniques potentially alters the miRNA content. Positive and negative selection, as well as Fluorescence-activated cell sorting (FACS) have a highly significant influence on the physiological miRNA content[32]. Here, single cell miRNA profiling might help to improve our understanding of age-related miRNA patterns in the future. At best, single cell miRNA data and cell free miRNA profiles are combined in the future using advancing sequencing technologies. Finally, such data might further our understanding of miRNAs in aging, diseases and their interplay with organ patterns that are only partially understood[29,51].

Over recent years, numerous studies have emerged highlighting systemic molecular aging factors detected with different omics technologies, including epigenetics, transcriptomics, and proteomics. Our study specifically extends our knowledge of blood and plasma-based miRNA patterns in aging. In our study we observe non-linear miRNA aging patterns. Moreover, the high degree of age-related biomarker patterns challenges the concept of age independent miRNA biomarker profiles, calling for different statistical models in aged and younger individuals. The changes with aging are not only attributed to one mature form, we also provide detailed insights into changes of the usage of the 3' and 5' mature arms in aging.

Furthering our understanding of age-related miRNA changes in healthy individuals and diseased patients will not only increase our understanding of age-related blood-borne gene regulation, but also improve miRNA-based biomarker development, and aid the development of RNA-based therapies.

Methods

Cohort. In this study, we processed data from $n_{\mathrm{total}} = 4433$ whole blood samples. We excluded 40 individuals (0.9%) because of insufficient data quality or missing clinical or demographic information. The final cohort consists thus of 4393 samples. These include unaffected controls ($n_{\mathrm{HC}} = 1{,}334$), Parkinson's Disease ($n_{\mathrm{PD}} = 944$), heart diseases ($n_{\mathrm{HD}} = 607$), non-tumor lung diseases ($n_{\mathrm{NTLD}} = 586$), lung cancer ($n_{\mathrm{LC}} = 517$), and other diseases ($n_{\mathrm{OD}} = 405$). The diseases can be split further in sub-classes. For lung cancer, we included non-small cell, as well as small cell lung cancer. For non-small cell lung cancer, we can further divide them in adenocarcinoma and squamous cell carcinoma. These split in low grade and high-grade tumors according to the TNM grading. The lung cancer cohort has been previously described in more detail[52]. The heart diseases include coronary artery disease, dilated cardiomyopathies and acute coronary syndrome. The non-tumor lung diseases include mostly chronic obstructive pulmonary diseases, the other diseases include sepsis, liver cirrhosis, breast cancer, endometriosis, and melanoma patients. We aggregate the diseases to an organ level (heart, brain and lung). Only for the lung we split the cohort in cancer and non-cancer samples. This aggregation level has been selected in a manner to be able to distinguish between healthy and diseased aging by having sufficient cohort sizes. Detailed diagnoses for each sample are provided in Supplementary Data 1. All participants gave informed consent. The local ethics committee of Saarland University approved the study. The study has been conducted in compliance with all relevant ethical regulations regarding the use of human study participants.

RNA extraction and measurement of miRNAs. RNA from 4433 whole blood samples in PAXgeneTubes (BD Biosciences, Franklin Lakes, NJ, USA) was isolated using the PAXgene Blood miRNA Kit (Qiagen, Hilden, Germany) using manufacturers recommendation. The extractions were done manually or semi-automatically on the Qiacube robot. The RNA was quantified using Nanodrop (Thermo Fisher Scientific, Waltham, MA, USA) and the RNA integrity was checked using a bioanalyzer with the RNA Nano Kit (Agilent Technologies, Santa Clara, CA, USA). The genome-wide expression profiles of human mature miRNAs was determined with Human miRNA microarrays and the miRNA Complete Labeling and Hyb Kit (Agilent Technologies). The labeled RNA was hybridized to the arrays for 20 h at 55 °C with 20 rpm rotation. The microarrays were subsequently washed twice, dried and scanned with 3 µm resolution in double-path mode (Agilent Technologies). The raw data were extracted using the manufacturers Feature Extraction software (Agilent Technologies). Details on the RNA extraction and microarray measurement procedure have been also previously described[53,54]. In difference to our previous studies we tried to further minimize any variability. In this study, we thus only included genome wide miRNA profiles that have been measured using the Agilent miRBase V21 biochip.

Blood cell deconvolution. To analyze the miRNA blood cell composition, we made use of our previous study that presented a high-resolution representation of human miRNAs in different blood compounds[50]. From the data, we asked which miRNAs are present in at least one sample of the respective blood compound and generated an upset plot from the data. In some detail, we included serum, microvesicles, red blood cells, CD15, CD19, CD8, CD56, CD4, and CD14 cells.

Correlation of age and sex to miRNAs. To find associations between the sex and the miRNA expression we applied 2-tailed non-parametric Wilcoxon Mann–Whitney tests. To compute linear correlation values between the age and miRNA expression values we computed the Pearson Correlation Coefficient (PC) and Spearman Correlation (SC). Further, to detect potentially non-linear relations between single miRNAs and the age we also computed the Distance Correlation (DI) between age and sex. To relate the DI and the SC, we computed a smoothed spline with eight degrees of freedom and computed the minimal Euclidean distance of each data point from the spline. Points with a distance of 0.02 (the threshold of 0.02 has been computed by a histogram-based approach) were highlighted and are considered to follow a non-linear trend with aging. In the further analyses, we applied only the rank-based Spearman Correlation (SC) instead of the Pearson Correlation that assumes linear effects in data. Beyond linear and non-linear correlations between single miRNAs and the age we also performed different standard dimension reduction technologies, including principal component analysis, t-stochastic neighborhood embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). To calculate the fraction of variance attributed to the age and sex we applied principal variant component analysis (PVCA), originally developed to discover batch effects in microarray experiments.

Analysis of arm shift events. Recently, we developed the miRSwitch database and analysis tool to identify and characterize human arm shift and arm switch events[30]. To detect associations between aging and differential arm usage we considered the following criteria. First, the percentage of the 5' mature arm given the total expression of 3' and 5' arm must correlate with an absolute Spearman Correlation Coefficient > 0.2. Second, the correlation must reach a *p*-value of at least 0.05. The *p*-value is computed by the R cor.test function via the asymptotic t approximation. Third, the difference between the minimal and maximal percentage of 5' arm expression for any samples must exceed 0.2 (20%). As fourth and last condition, the 3' and 5' mature form must have a different sign, i.e., the 5' has to increase with age and the 3' to decrease or vice versa. The miRNAs that were discovered by this procedure where then checked by miRSwitch.

Cluster analysis and miRNA enrichment analysis. We split the miRNAs in 5 groups, strongly decreasing with age (SC < −0.2), decreasing with age (SC between −0.2 and −0.1), not changing with age (SC between −0.1 and 0.1), miRNAs increasing with age (SC > 0.1 and <0.2) and miRNAs increasing strongly with age (SC > 0.2). For each cluster, we computed smoothed splines for each miRNA and the cluster average allowing three degrees of freedom. Further, we computed for disjoint age windows of five years whether miRNAs are significantly higher or lower in cases versus controls at an alpha level of 0.05 and colored them, respectively, in red and green. To find categories that are significantly enriched either for miRNAs increasing or decreasing over age we performed a miRNA enrichment analysis using the miEAA tool[55], which has been recently updated[56]. Thereby, for over 14,000 categories running sum statistics are computed. The sorted list of miRNAs (increasing correlation with age) is processed from left to right. Whenever a miRNA is located in a category the running sum is increased otherwise it is decreased. The running sum is then plotted along with 100 random permutation tests. Notably, the *p*-value is not computed from the permutations but exactly by using dynamic programming. A category showing a perfect "V" like shape would contain miRNAs that are increasing over age while a category following a pyramid like shape contains miRNAs that are decreasing over age.

Sliding window analysis based on Cohen's d. Since *p*-values rely on the effect size and the cohort size different group sizes bias the results frequently. In our sliding window analysis, we observed substantial differences, i.e., cases and controls are not equally distributed across the age range. We thus performed all analyses using Cohen's d as effect size. All effects with an absolute value of above 0.5 were considered relevant. Negative effect sizes thereby characterize downregulation and positive effect sizes upregulation. We computed effect sizes for each disease in windows of 10 years, shifted by one year, starting from 30 and ending at 70 years (i.e., the last window is from 70 to 79 years). Only when at least 20 cases and control measurements were available effect sizes were computed. The calculated effect sizes were then summarized and a smoothed spline with eight degrees of freedom were computed.

Self-organizing map (SOM) for finding disease patterns. One task in high dimensional data analysis is to group features and to generate lower dimensional representation of high dimensional data. Self-organizing maps (SOMs) are one type of artificial neural networks (ANNs), relying on competitive learning. As described by Kohonen already in 1982[57], in a network of adaptive elements "receiving signals from a primary event space, the signal representations are automatically mapped onto a set of output responses in such a way that the responses acquire the same topological order as that of the primary events". From input data, a typically two-dimensional discretized representation of the input space is derived that can be visualized by heat maps. To compute self-organizing maps for patients and controls in an age dependent manner we computed the effect size for each disease group over all patients, for young patient (30–60 years) and for old patients (60–80 years) separately. Only 801 highest expressed miRNAs were included in this analysis. For the biomarker sets, a 10 × 10 hexagonal som grid was used to train a network. The data set was presented 10,000 times to the network. The learning rate was set to be between 0.05 and 0.01, meaning that the learning rate linearly decreased from 0.05 to 0.01 over the 10,000 iterations. To cluster the SOM cells, we performed hierarchical clustering. In more detail, we applied the R hclust function to carry out agglomerative complete linkage clustering. As distance measure we computed the Euclidean distance using the R dist function.

Plasma proteomics measurements. We used data from a recent study investigating the effect of aging on the human plasma proteome. In this study, 2925 proteins were measured using the SomaScan assay in 4264 subjects from the INTERVAL and LonGenity cohorts[5]. The SomaScan platform is based on modified single-stranded DNA aptamers binding to specific protein targets. Assay details were previously described. Relative Fluorescence Units (RFUs) were log10-transformed and we used a 10 years sliding window to estimate proteins trajectories throughout lifespan.

Target analysis and target network analysis. The main biological function of miRNAs is to bind the 3' UTR of genes and to degrade the target mRNAs. In reality, miRNAs and genes thereby follow a n:m relation, i.e., one miRNA can regulate many genes and one gene is regulated by many miRNAs. Further, there exist different confidence levels to assume a pair-wise regulation of a miRNA to a target gene. Most relations are only predicted by one or several computational analyses. Another set is composed of miRNA gene pairs with weak evidence, e.g., from microarray experiments. The most reliable category consists of miRNA gene pairs with strong evidence, e.g., validated by reporter assays. We only considered this most reliable set of miRNA gene interactions and extracted the set from the miRTarBase database[34,58]. Our analysis highlighted that around 20% of miRNAs are increasing with age, 20% are decreasing and 60% are not age dependent. We assumed the same distribution for human plasma proteins changing with age and asked how many miRNAs going down with age regulate genes/proteins going up and down with age, respectively. Similarly, we asked how many miRNAs going up with age regulate genes/proteins going up and down with the age.

To construct a reliable core network, we combined five stringed filtering approaches and only considered those connections between miRNAs and genes that fulfill all filtering criteria. In the least stringent version the filters include (a) a strong experimental evidence of a target interaction from the literature; (b) one of the most decreasing miRNAs (5%) regulates (c) one of the most upregulated proteins (5%) over aging. To avoid a bias towards genes/proteins that are targeted only by one or few miRNAs, potentially also fragmenting the network, we (d) only considered proteins that are regulated by more than eight miRNAs. Next, we analyzed the correlation between miRNAs and genes/proteins in the network over 40 discrete age ranges from 30 to 70 years. Each age range thereby spans 10 years. For the 40 data points corresponding to 40 age windows we computed the Spearman correlation between miRNA expression in this age window and protein expression. As last criterion we added (e) only edges that have an absolute Spearman correlation of at least 0.6. This network has been visualized with the igraph library. Nodes were colored with respect to changes in age and edges weights relative to the absolute Spearman correlation.

Single cell analysis. We used data that have been made available by 10× genomics (https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_v3). The profiles were subsequently processed with scater[59] and scran[60] with default parameters, cell type annotations with singleR[61].

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw microarray measurements are freely available for any scientific purpose upon request as Excel Table and Tab Delimited Text file (110 MB) to data@ccb.uni-saarland.de. The use of the data for commercial purposes is prohibited.

Code availability

The data analysis has been performed using the R software for statistical computation (R 3.3.2 GUI 1.68 Mavericks build (7288)) using freely available packages. The following packages were used: ROC, RColorBrewer, preprocessCore, tsne, effsize, UpSetR, kohonen, fmsb, igraph. All packages are available from Bioconductor or CRAN.

Received: 18 February 2020; Accepted: 16 October 2020;
Published online: 24 November 2020

References

1. Harman, D. The aging process: major risk factor for disease and death. *Proc. Natl Acad. Sci. USA* **88**, 5360–5363 (1991).
2. Valdes, A. M., Glass, D. & Spector, T. D. Omics technologies and the study of human ageing. *Nat. Rev. Genet.* **14**, 601–607 (2013).
3. Deelen, J. et al. A meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nat. Commun.* **10**, 3669 (2019).
4. Aramillo Irizar, P. et al. Transcriptomic alterations during ageing reflect the shift from cancer to degenerative diseases in the elderly. *Nat. Commun.* **9**, 327 (2018).
5. Lehallier, B. et al. Undulating changes in human plasma proteome profiles across the lifespan. *Nat. Med.* **25**, 1843–1850 (2019).
6. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).
7. Schaum, N. et al. Ageing hallmarks exhibit organ-specific temporal signatures. *Nature* **583**, 596–602 (2020).
8. Tabula Muris, C. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590–595 (2020).
9. Hahn, O. et al. A nutritional memory effect counteracts benefits of dietary restriction in old mice. *Nat. Metab.* **1**, 1059–1073 (2019).
10. Villeda, S. A. et al. Young blood reverses age-related impairments in cognitive function and synaptic plasticity in mice. *Nat. Med.* **20**, 659–663 (2014).
11. Middeldorp, J. et al. Preclinical assessment of young blood plasma for Alzheimer disease. *JAMA Neurol.* **73**, 1325–1333 (2016).
12. Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350–355 (2004).
13. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
14. Bushati, N. & Cohen, S. M. microRNA functions. *Annu. Rev. Cell Dev. Biol.* **23**, 175–205 (2007).
15. Gurtan, A. M. & Sharp, P. A. The role of miRNAs in regulating gene expression networks. *J. Mol. Biol.* **425**, 3582–3600 (2013).
16. Krek, A. et al. Combinatorial microRNA target predictions. *Nat. Genet.* **37**, 495–500 (2005).
17. Leidinger, P. et al. A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biol.* **14**, R78 (2013).
18. Keller, A. et al. Validating Alzheimer's disease micro RNAs using next-generation sequencing. *Alzheimers Dement.* **12**, 565–576 (2016).
19. Keller, A. et al. Comprehensive analysis of microRNA profiles in multiple sclerosis including next-generation sequencing. *Mult. Scler.* **20**, 295–303 (2014).
20. Vogel, B. et al. Multivariate miRNA signatures as biomarkers for non-ischaemic systolic heart failure. *Eur. Heart J.* **34**, 2812–2822 (2013).
21. Smith-Vikos, T. & Slack, F. J. MicroRNAs and their roles in aging. *J. Cell Sci.* **125**, 7–17 (2012).
22. Somel, M. et al. MicroRNA, mRNA, and protein expression link development and aging in human and macaque brain. *Genome Res.* **20**, 1207–1218 (2010).
23. Drummond, M. J. et al. Aging and microRNA expression in human skeletal muscle: a microarray and bioinformatics analysis. *Physiol. Genomics* **43**, 595–603 (2011).
24. Zhang, H. et al. Investigation of microRNA expression in human serum during the aging process. *J. Gerontol. A Biol. Sci. Med. Sci.* **70**, 102–109 (2015).
25. Noren Hooten, N. et al. Age-related changes in microRNA levels in serum. *Aging* **5**, 725–740 (2013).
26. Meder, B. et al. Influence of the confounding factors age and sex on microRNA profiles from peripheral blood. *Clin. Chem.* **60**, 1200–1208 (2014).
27. Huan, T. et al. Age-associated microRNA expression in human peripheral blood is associated with all-cause mortality and age-related traits. *Aging Cell* **17**, https://doi.org/10.1111/acel.12687 (2018).
28. Kehl, T. et al. miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkz1022 (2019).
29. Ludwig, N. et al. Distribution of miRNA expression across human tissues. *Nucleic Acids Res.* **44**, 3865–3877 (2016).
30. Kern, F. et al. miRSwitch: detecting microRNA arm shift and switch events. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkaa323 (2020).
31. Chen, L. et al. miRNA arm switching identifies novel tumour biomarkers. *EBioMedicine* **38**, 37–46 (2018).
32. Schwarz, E. C. et al. Deep characterization of blood cell miRNomes by NGS. *Cell Mol. Life Sci.* **73**, 3169–3181 (2016).
33. Valiathan, R., Ashman, M. & Asthana, D. Effects of ageing on the immune system: infants to elderly. *Scand. J. Immunol.* **83**, 255–266 (2016).
34. Huang, H. Y. et al. miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkz896 (2019).
35. Gil, V. & Del Rio, J. A. Functions of plexins/neuropilins and their ligands during hippocampal development and neurodegeneration. *Cells* **8**, https://doi.org/10.3390/cells8030206 (2019).
36. Zhou, Y. F. et al. Sema3E/PlexinD1 inhibition is a therapeutic strategy for improving cerebral perfusion and restoring functional loss after stroke in aged rats. *Neurobiol. Aging* **70**, 102–116 (2018).
37. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
38. Muller-McNicoll, M. et al. SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. *Genes Dev.* **30**, 553–566 (2016).
39. Fehlmann, T. et al. cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin. Epigenet.* **8**, 123 (2016).
40. Juzenas, S. et al. Depletion of erythropoietic miR-486-5p and miR-451a improves detectability of rare microRNAs in peripheral blood-derived small RNA sequencing libraries. *NAR Genom. Bioinform.* **2**, https://doi.org/10.1093/nargab/lqaa008 (2020).
41. Fehlmann, T. et al. A high-resolution map of the human small non-coding transcriptome. *Bioinformatics* **34**, 1621–1628 (2018).
42. Londin, E. et al. Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proc. Natl Acad. Sci. USA* **112**, E1106–E1115 (2015).
43. Keller, A. et al. Toward the blood-borne miRNome of human diseases. *Nat. Methods* **8**, 841–843 (2011).
44. Wang, H. et al. Transcriptome analysis of common and diverged circulating miRNAs between arterial and venous during aging. *Aging* **12**, 12987–13004 (2020).
45. Maffioletti, E. et al. miR-146a plasma levels are not altered in Alzheimer's disease but correlate with age and illness severity. *Front. Aging Neurosci.* **11**, 366 (2019).
46. Alibhai, F. J. et al. Cellular senescence contributes to age-dependent changes in circulating extracellular vesicle cargo and function. *Aging Cell* **19**, e13103 (2020).
47. Gomez, I. et al. Neutrophil microvesicles drive atherosclerosis by delivering miR-155 to atheroprone endothelium. *Nat. Commun.* **11**, 214 (2020).
48. Wei, Z. et al. Coding and noncoding landscape of extracellular RNA released by human glioma stem cells. *Nat. Commun.* **8**, 1145 (2017).
49. Cheng, M. et al. Circulating myocardial microRNAs from infarcted hearts are carried in exosomes and mobilise bone marrow progenitor cells. *Nat. Commun.* **10**, 959 (2019).
50. Juzenas, S. et al. A comprehensive, cell specific microRNA catalogue of human peripheral blood. *Nucleic Acids Res.* **45**, 9290–9301 (2017).

51. Fehlmann, T., Ludwig, N., Backes, C., Meese, E. & Keller, A. Distribution of microRNA biomarker candidates in solid tissues and body fluids. *RNA Biol.* **13**, 1084–1088 (2016).
52. Fehlmann, T. et al. Evaluating the use of circulating MicroRNA profiles for lung cancer detection in symptomatic patients. *JAMA Oncol.* https://doi.org/10.1001/jamaoncol.2020.0001 (2020).
53. Keller, A. et al. Genome-wide MicroRNA expression profiles in COPD: early predictors for cancer development. *Genomics Proteom. Bioinform.* **16**, 162–171 (2018).
54. Ludwig, N. et al. Spring is in the air: seasonal profiles indicate vernal change of miRNA activity. *RNA Biol.* **16**, 1034–1043 (2019).
55. Backes, C., Khaleeq, Q. T., Meese, E. & Keller, A. miEAA: microRNA enrichment analysis and annotation. *Nucleic Acids Res.* **44**, W110–W116 (2016).
56. Kern, F. et al. miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkaa309 (2020).
57. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59–69 (1982).
58. Hsu, S. D. et al. miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.* **39**, D163–D169 (2011).
59. McCarthy, D. J., Campbell, K. R., Lun, A. T. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).
60. Lun, A. T., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* **5**, 2122 (2016).
61. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).

Acknowledgements

This work was supported by grants from the Michael J Fox Foundation (to Andreas Keller) and the Luxembourg National Research Fund (FNR) within the PEARL programme (FNR; FNR/P13/6682797 to Rejko Krüger).

Author contributions

T.F.: Data analysis, conception of the study and analyses; B.L.: Data analysis, manuscript drafting; N.S.: Data interpretation, manuscript drafting; O.H.: Data interpretation, manuscript drafting, data representation; M.K.: Data analysis; Y.L.: Data interpretation; N.G.: Data interpretation, data representation; L.G.: Data interpretation; C.B.: Data analysis; R.B.: Data interpretation, conception of the study and analyses; F.K.: Data analysis, data representation; R.K.: Data interpretation, conception of the study and analyses, providing clinical data and patient specimens; F.L.: Data interpretation, providing clinical data and patient specimens; N.L.: Performing analyses and contributing experimental data; B.M.: Data interpretation, conception of the study and analyses, providing clinical data and patient specimens; B.F.: Data interpretation, manuscript drafting; W.M.: Data interpretation; D.B.: Data interpretation; K.B.: Data interpretation; C.D.: Data interpretation; A.K.v.T.: Data interpretation, providing clinical data and patient specimens; G.W.E.: Data interpretation, providing clinical data and patient specimens; S.M.: Data interpretation, Performing analyses and contributing experimental data; N.B.: Data interpretation, Performing analyses and contributing experimental data; M.R.: Data interpretation, providing clinical data and patient specimens; T.W.C.: Data interpretation, manuscript drafting; E.M.: Data interpretation, conception of the study and analyses, manuscript drafting; A.K.: Data analysis, Data interpretation, conception of the study and analyses, manuscript drafting.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

M.K. is also employed by Hummingbird Diagnostic GmbH. The remaining authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-19665-1.

Correspondence and requests for materials should be addressed to A.K.

Peer review information *Nature Communications* thanks Lifang Hou and the other, anonymous, reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

5. Conclusion

The dynamics of miRNA actions and the complexity of miRNA-mediated gene regulations in physiological and pathological conditions requires high throughput technologies to study their structures, pathways, expression profiles etc. In addition, high spatial and temporal resolutions will help to understand their intercellular and mode of actions in more comprehensive way. In the PhD studies, we demonstrated new technologies such as CoolMPS can be potentially used in a more accurate, cheaper way to enable large scale miRNA associated omics study. And miRNATargetlink also paves good way for miRNA targets and pathway studies, to enable biomarkers findings.

In following studies, large scale miRNA sequencing with lower costs, and combine with other omics data cohort will be possible with CoolMPS. It is continuous effort to also explore single cell miRNA studies with more powerful math models to gain even more clearer complex miRNA regulatory networks.

Further, miRNAs as good biomarkers can be potentially applied from dry blood spot, it is exciting to develop a sample to answer solution, to enable home sampling, non-invasive, cost effective and most importantly get high sensitivity and specificity for Alzheimer early diagnosis or at least started monitoring the miRNA profiles changes from earlier ages to enable better interventions and aging care.

This will be potential translational value from this PhD studies, which brings new projects of interests for miRNA at single cell level and the tissue level. At the same time, we must recognize the limitation of the work. The sncRNA data can be measured at scale only from complex mixtures but not at the single cell level. The next step of research that is mandatory to understand complex pathological pathways on a molecular level includes the parallel measurement of sncRNAs, gene expression, chromatin state, DNA methylation and other features from the same cells. Once available, respective protocols and data will pose however tremendous challenges to data scientist.

6. Bibliography

References

[1] Fehlmann, T. *et al.* cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin Epigenetics* **8**, 123, doi:10.1186/s13148-016-0287-1 (2016).

[2] Li, Y. *et al.* CoolMPS: evaluation of antibody labeling based massively parallel non-coding RNA sequencing. *Nucleic Acids Res* **49**, e10, doi:10.1093/nar/gkaa1122 (2021).

[3] Drmanac S, Callow M, Chen L, et al. CoolMPS™: Advanced massively parallel sequencing using antibodies specific to each natural nucleobase. bioRxiv; 2020. DOI: 10.1101/2020.02.19.953307.

[4] Pirritano, M. *et al.* Next Generation Sequencing Analysis of Total Small Noncoding RNAs from Low Input RNA from Dried Blood Sampling. *Anal Chem* **90**, 11791-11796, doi:10.1021/acs.analchem.8b03557 (2018).

[5] Fehlmann, T. *et al.* Common diseases alter the physiological age-related blood microRNA profile. *Nat Commun* **11**, 5958, doi:10.1038/s41467-020-19665-1 (2020).

[6] Fabian Kern, et al. miRTargetLink 2.0—interactive miRNA target gene and target pathway networks, *Nucleic Acids Research*, 2021;, gkab297, https://doi.org/10.1093/nar/gkab297

[7] Felekkis K, Touvana E, Stefanou Ch, Deltas C. microRNAs: a newly described class of encoded molecules that play a role in health and disease. Hippokratia. 2010;14(4):236-240.

[8] Ambros V. microRNAs: tiny regulators with great potential. Cell. 2001;107(7):823-826. doi:10.1016/s0092-8674(01)00616-x

[9] Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. Identification of novel genes coding for small expressed RNAs. Science. 2001;294(5543):853-858. doi:10.1126/science.1064921

[10] Tuschl T. RNA interference and small interfering RNAs. Chembiochem. 2001;2(4):239-245. doi:10.1002/1439-7633(20010401)2:4<239::AID-CBIC239>3.0.CO;2-R

[11] Ambros V. The functions of animal microRNAs. Nature. 2004;431(7006):350-355. doi:10.1038/nature02871

[12] Sempere LF, Freemantle S, Pitha-Rowe I, Moss E, Dmitrovsky E, Ambros V. Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. Genome Biol. 2004;5(3):R13. doi:10.1186/gb-2004-5-3-r13

[13] Lukiw WJ. Micro-RNA speciation in fetal, adult and Alzheimer's disease hippocampus. Neuroreport. 2007;18(3):297-300. doi:10.1097/WNR.0b013e3280148e8b

[14] Mehler MF, Mattick JS. Noncoding RNAs and RNA editing in brain development, functional diversification, and neurological disease. Physiol Rev. 2007;87(3):799-823. doi:10.1152/physrev.00036.2006

[15] Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature. 2010;466(7308):835-840. doi:10.1038/nature09267

[16] Witkos TM, Koscianska E, Krzyzosiak WJ. Practical Aspects of microRNA Target Prediction. Curr Mol Med. 2011;11(2):93-109. doi:10.2174/156652411794859250

[17] 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. doi:10.1016/0092-8674(93)90529-y

[18] O'Brien J, Hayder H, Zayed Y, Peng C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Front Endocrinol (Lausanne). 2018;9:402. Published 2018 Aug 3. doi:10.3389/fendo.2018.00402

[19] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116(2):281-297. doi:10.1016/s0092-8674(04)00045-5

[20] Hammond SM. An overview of microRNAs. Adv Drug Deliv Rev. 2015;87:3-14. doi:10.1016/j.addr.2015.05.001

[21] Ambros V. The functions of animal microRNAs. Nature. 2004;431(7006):350-355. doi:10.1038/nature02871

[22] Arteaga-Vázquez M, Caballero-Pérez J, Vielle-Calzada JP. A family of microRNAs present in plants and animals. Plant Cell. 2006;18(12):3355-3369. doi:10.1105/tpc.106.044420

[23] Lukiw WJ. Variability in microRNA (miRNA) abundance, speciation and complexity amongst different human populations and potential relevance to Alzheimer's disease (AD). Front Cell Neurosci. 2013;7:133. Published 2013 Aug 27. doi:10.3389/fncel.2013.00133

[24] Lukiw WJ. Evolution and complexity of micro RNA in the human brain. Front Genet. 2012;3:166. Published 2012 Sep 3. doi:10.3389/fgene.2012.00166

[25] Alexandrov PN, Dua P, Hill JM, Bhattacharjee S, Zhao Y, Lukiw WJ. microRNA (miRNA) speciation in Alzheimer's disease (AD) cerebrospinal fluid (CSF) and extracellular fluid (ECF). Int J Biochem Mol Biol. 2012;3(4):365-373.

[26] Wang M, Qin L, Tang B. MicroRNAs in Alzheimer's Disease. *Front Genet*. 2019;10:153. Published 2019 Mar 1. doi:10.3389/fgene.2019.00153

[27] Huntzinger E, Izaurralde E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat Rev Genet. 2011;12(2):99-110. doi:10.1038/nrg2936

[28] Xu W, San Lucas A, Wang Z, Liu Y. Identifying microRNA targets in different gene regions. BMC Bioinformatics. 2014;15 Suppl 7(Suppl 7):S4. doi:10.1186/1471-2105-15-S7-S4

[29] Loffreda A, Rigamonti A, Barabino SM, Lenzken SC. RNA-Binding Proteins in the Regulation of miRNA Activity: A Focus on Neuronal Functions. Biomolecules. 2015;5(4):2363-2387. Published 2015 Sep 30. doi:10.3390/biom5042363

[30] Wu M, Wang G, Tian W, Deng Y, Xu Y. MiRNA-based Therapeutics for Lung Cancer. Curr Pharm Des. 2018;23(39):5989-5996. doi:10.2174/1381612823666170714151715

[31] Bernardo BC, Ooi JY, Lin RC, McMullen JR. miRNA therapeutics: a new class of drugs with potential therapeutic applications in the heart. Future Med Chem. 2015;7(13):1771-1792. doi:10.4155/fmc.15.107

[32] Rupaimoole R, Slack FJ. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. Nat Rev Drug Discov. 2017;16(3):203-222. doi:10.1038/nrd.2016.246

[33] Bayraktar R, Van Roosbroeck K. miR-155 in cancer drug resistance and as target for miRNA-based therapeutics. Cancer Metastasis Rev. 2018;37(1):33-44. doi:10.1007/s10555-017-9724-7

[34] Saliminejad K, Khorram Khorshid HR, Soleymani Fard S, Ghaffari SH. An overview of microRNAs: Biology, functions, therapeutics, and analysis methods. J Cell Physiol. 2019;234(5):5451-5465. doi:10.1002/jcp.27486

[35] Trang P, Weidhaas JB, Slack FJ. MicroRNAs as potential cancer therapeutics. Oncogene. 2008;27 Suppl 2:S52-S57. doi:10.1038/onc.2009.353

[36] Fasanaro P, Greco S, Ivan M, Capogrossi MC, Martelli F. microRNA: emerging therapeutic targets in acute ischemic diseases. Pharmacol Ther. 2010;125(1):92-104. doi:10.1016/j.pharmthera.2009.10.003

[37] Hydbring P, Badalian-Very G. Clinical applications of microRNAs. F1000Res. 2013;2:136. Published 2013 Jun 6. doi:10.12688/f1000research.2-136.v3

[38] Chakraborty C, Sharma AR, Sharma G, Lee SS. Therapeutic advances of miRNAs: A preclinical and clinical update. J Adv Res. 2020;28:127-138. Published 2020 Aug 29. doi:10.1016/j.jare.2020.08.012

[39] Tonge DP, Gant TW. What is normal? Next generation sequencing-driven analysis of the human circulating miRNAOme. BMC Mol Biol. 2016;17:4. Published 2016 Feb 9. doi:10.1186/s12867-016-0057-9

[40] Cheng HH, Yi HS, Kim Y, et al. Plasma processing conditions substantially influence circulating microRNA biomarker levels. PLoS One. 2013;8(6):e64795. Published 2013 Jun 7. doi:10.1371/journal.pone.0064795

[41] Backes C, Leidinger P, Altmann G, et al. Influence of next-generation sequencing and storage conditions on miRNA patterns generated from PAXgene blood. Anal Chem. 2015;87(17):8910-8916. doi:10.1021/acs.analchem.5b02043

[42] Ludwig N, Fehlmann T, Galata V, et al. Small ncRNA-Seq Results of Human Tissues: Variations Depending on Sample Integrity. Clin Chem. 2018;64(7):1074-1084. doi:10.1373/clinchem.2017.285767

[43] Kahraman M, Laufer T, Backes C, 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. doi:10.1373/clinchem.2017.271619

[44] Guo L, Chen F. A challenge for miRNA: multiple isomiRs in miRNAomics. Gene. 2014;544(1):1-7. doi:10.1016/j.gene.2014.04.039

[45] Alles J, Fehlmann T, Fischer U, et al. An estimate of the total number of true human miRNAs. Nucleic Acids Res. 2019;47(7):3353-3364. doi:10.1093/nar/gkz097

[46] Keller A, Meese E. Can circulating miRNAs live up to the promise of being minimal invasive biomarkers in clinical settings?. Wiley Interdiscip Rev RNA. 2016;7(2):148-156. doi:10.1002/wrna.1320

[47] Backes C, Meese E, Keller A. Specific miRNA Disease Biomarkers in Blood, Serum and Plasma: Challenges and Prospects. Mol Diagn Ther. 2016;20(6):509-518. doi:10.1007/s40291-016-0221-4

[48] Prince M, Bryce R, Albanese E, Wimo A, Ribeiro W, Ferri CP. The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimers Dement*. 2013;9(1):63-75.e2. doi:10.1016/j.jalz.2012.11.007

[49] Villemagne VL, Burnham S, Bourgeat P, et al. Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. *Lancet Neurol*. 2013;12(4):357-367. doi:10.1016/S1474-4422(13)70044-9

[50] Reiman EM, Quiroz YT, Fleisher AS, Chen K, Velez-Pardos C, Jimenez-Del-Rio M, et al. Brain imaging and fluid biomarker analysis in young adults at genetic risk for autosomal dominant Alzheimer's disease in the presenilin 1 E280A kindred: A case-control study. *Lancet Neurol* 2012;**11**(2):1048-56.

[51] Jack CR, Lowe VJ, Weigand SD, Wiste HJ, Senjem ML, Knopman DS, et al. Serial PiB and MRI in normal, mild cognitive impairment and Alzheimer's disease: Implications for sequence of pathological events in Alzheimer's disease. *Brain* 2009;**132**:1355-65.

[52] Bateman RJ, Xiong C, Benzinger TL, Fagan AM, Goate A, Fox NC, et al. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *N Engl J Med* 2012;**367**(9):795-804.

[53] Gordon BA, Blazey TM, Su Y, Hari-Raj A, Dincer A, Flores S, et al. Spatial patterns of neuroimaging biomarker change in individuals from families with autosomal

dominant Alzheimer's disease: A longitudinal study. *Lancet Neurol* 2018;**17**(3):241-50.

[54] Braak H, Thal DR, Ghebremedhin E, Del Tredici K. Stages of the pathologic process in Alzheimer disease: age categories from 1 to 100 years. *J Neuropathol Exp Neurol* 2011;**70**(11):960-9

[55] El-Hayek YH, Wiley RE, Khoury CP, et al. Tip of the Iceberg: Assessing the Global Socioeconomic Costs of Alzheimer's Disease and Related Dementias and Strategic Implications for Stakeholders. *J Alzheimers Dis*. 2019;70(2):323-341. doi:10.3233/JAD-190426

[56] Hampel H, Toschi N, Babiloni C, et al. Revolution of Alzheimer Precision Neurology. Passageway of Systems Biology and Neurophysiology. *J Alzheimers Dis*. 2018;64(s1):S47-S105. doi:10.3233/JAD-179932

[57] Cacabelos R, Cacabelos P, Torrellas C. Personalized Medicine of Alzheimer’s Disease. *Handbook of Pharmacogenomics and Stratified Medicine*. 2014;563-615. doi:10.1016/B978-0-12-386882-4.00027-X

[58] Games D, Adams D, Alessandrini R, et al. Alzheimer-type neuropathology in transgenic mice overexpressing V717F beta-amyloid precursor protein. *Nature*. 1995;373(6514):523-527. doi:10.1038/373523a0

[59] Frank RA, Galasko D, Hampel H, et al. Biological markers for therapeutic trials in Alzheimer's disease. Proceedings of the biological markers working group; NIA initiative on neuroimaging in Alzheimer's disease. *Neurobiol Aging*. 2003;24(4):521-536. doi:10.1016/s0197-4580(03)00002-2

[60] Sonnen JA, Montine KS, Quinn JF, Kaye JA, Breitner JC, Montine TJ. Biomarkers for cognitive impairment and dementia in elderly people. *Lancet Neurol*. 2008;7(8):704-714. doi:10.1016/S1474-4422(08)70162-5

[61] Petersen RC, Jack CR Jr. Imaging and biomarkers in early Alzheimer's disease and mild cognitive impairment. *Clin Pharmacol Ther*. 2009;86(4):438-441. doi:10.1038/clpt.2009.166

[62] Trojanowski JQ, Vandeerstichele H, Korecka M, et al. Update on the biomarker core of the Alzheimer's Disease Neuroimaging Initiative subjects. *Alzheimers Dement*. 2010;6(3):230-238. doi:10.1016/j.jalz.2010.03.008

[63] Hardy J, Selkoe DJ. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics [published correction appears in Science 2002

Sep 27;297(5590):2209]. *Science*. 2002;297(5580):353-356. doi:10.1126/science.1072994

[64] Shaw LM. PENN biomarker core of the Alzheimer's disease Neuroimaging Initiative. *Neurosignals*. 2008;16(1):19-23. doi:10.1159/000109755

[65] Jack CR Jr, Knopman DS, Jagust WJ, et al. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol*. 2010;9(1):119-128. doi:10.1016/S1474-4422(09)70299-6

[66] Leidinger P, Backes C, Deutscher S, et al. A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biol*. 2013;14(7):R78. Published 2013 Jul 29. doi:10.1186/gb-2013-14-7-r78

[67] Doecke JD, Laws SM, Faux NG, et al. Blood-based protein biomarkers for diagnosis of Alzheimer disease. *Arch Neurol*. 2012;69(10):1318-1325. doi:10.1001/archneurol.2012.1282

[68] Keller A, Leidinger P, Bauer A, et al. Toward the blood-borne miRNome of human diseases. *Nat Methods*. 2011;8(10):841-843. Published 2011 Sep 4. doi:10.1038/nmeth.1682

[69] Diener D, Galata V, Keller A et al. MicroRNA profiling from dried blood samples. Critical Reviews in Clinical Laboratory Science, 2019; 56:2, 111-117, DOI: 10.1080/10408363.2018.1561641

[70] Hébert SS, Horré K, Nicolaï L, et al. Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/beta-secretase expression. *Proc Natl Acad Sci U S A*. 2008;105(17):6415-6420. doi:10.1073/pnas.0710263105

[71] Wang WX, Rajeev BW, Stromberg AJ, et al. 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(5):1213-1223. doi:10.1523/JNEUROSCI.5065-07.2008

[72] Geekiyanage H, Chan C. MicroRNA-137/181c regulates serine palmitoyltransferase and in turn amyloid β, novel targets in sporadic Alzheimer's disease. *J Neurosci*. 2011;31(41):14820-14830. doi:10.1523/JNEUROSCI.3883-11.2011

[73] Long JM, Lahiri DK. MicroRNA-101 downregulates Alzheimer's amyloid-β precursor protein levels in human cell cultures and is differentially expressed. *Biochem Biophys Res Commun*. 2011;404(4):889-895. doi:10.1016/j.bbrc.2010.12.053

[74] Geekiyanage H, Jicha GA, Nelson PT, Chan C. Blood serum miRNA: non-invasive biomarkers for Alzheimer's disease. *Exp Neurol*. 2012;235(2):491-496. doi:10.1016/j.expneurol.2011.11.026

[75] Wang H, Zhou Y, Yin Z, et al. Transcriptome analysis of common and diverged circulating miRNAs between arterial and venous during aging. *Aging (Albany NY)*. 2020;12(13):12987-13004. doi:10.18632/aging.103385

[76] Maffioletti E, Milanesi E, Ansari A, et al. miR-146a Plasma Levels Are Not Altered in Alzheimer's Disease but Correlate With Age and Illness Severity. *Front Aging Neurosci*. 2020;11:366. Published 2020 Jan 17. doi:10.3389/fnagi.2019.00366

[77] Schaum N, Lehallier B, Hahn O, et al. Ageing hallmarks exhibit organ-specific temporal signatures. *Nature*. 2020;583(7817):596-602. doi:10.1038/s41586-020-2499-y

[78] Fehlmann, T., Lehallier, B., Schaum, N. *et al.* Common diseases alter the physiological age-related blood microRNA profile. *Nat Commun* **11,** 5958 (2020). https://doi.org/10.1038/s41467-020-19665-1

[79] Miska, E.A., Alvarez-Saavedra, E., Townsend, M. *et al.* Microarray analysis of microRNA expression in the developing mammalian brain. *Genome Biol* **5,** R68 (2004). https://doi.org/10.1186/gb-2004-5-9-r68

[80] Fehlmann T, Backes C, Alles J, et al. A high-resolution map of the human small non-coding transcriptome. *Bioinformatics*. 2018;34(10):1621-1628. doi:10.1093/bioinformatics/btx814

[81] Londin E, Loher P, Telonis AG, 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-E1115. doi:10.1073/pnas.1420955112

[82] Schmartz GP, Kern F, Fehlmann T, Wagner V, Fromm B, Keller A. Encyclopedia of tools for the analysis of miRNA isoforms [published online ahead of print, 2020 Dec 14]. Brief Bioinform. 2020;bbaa346. doi:10.1093/bib/bbaa346

[83] Kolanowska M, Kubiak A, Jażdżewski K, Wójcicka A. MicroRNA Analysis Using Next-Generation Sequencing. Methods Mol Biol. 2018;1823:87-101. doi:10.1007/978-1-4939-8624-8_8

[84] Jirak, P., Wernly, B., Lichtenauer, M. et al. Next-generation sequencing analysis of circulating micro-RNA expression in response to parabolic flight as a spaceflight analogue. npj Microgravity 6, 31 (2020). https://doi.org/10.1038/s41526-020-00121-9

[85] Drmanac R, Sparks AB, Callow MJ, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010;327(5961):78-81. doi:10.1126/science.1181498

[86] Fehlmann T, Backes C, Kahraman M, 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. doi:10.1093/nar/gkx595

[87] Kern F, Backes C, Hirsch P, et al. What's the target: understanding two decades of in silico microRNA-target prediction. *Brief Bioinform*. 2020;21(6):1999-2010. doi:10.1093/bib/bbz111

Acknowledgement

Firstly, I would like to express my sincerest gratitude to my supervisor Prof. Dr. Andreas Keller for giving me the opportunity to join his group, Chair for Clinical Bioinformatics (CCB) at the Saarland University; and for his continued support of my Ph.D. studies and the related research, especially his motivation to understand aging process has been constantly encouraging me throughout my other healthcare devotions.

Furthermore, I would like to thank the rest of my thesis committee for their insightful comments and evaluation of my thesis.

I am grateful to my fellow Ph.D. students Mr. Tobias Fehlmann, Mr. Fabian Kern from University Saarland and special thanks go to the PIs Dr. Nicole Ludwig and Dr. Eckart Messe, Prof. Rolf Müller, Dr. Cord Staehler, Prof. Tony Wyss-Coray, and so on for their helpful advice and collaborations on different topics of research.

I would also like to show my gratitude to the employees of MGI Tech Co.,Ltd and BGI Genomics Co.,Ltd for working together with me on the coolMPS project which became the main part of my study.

Last but not least, I would like to thank my family for supporting me throughout my studies and life in general. Special thanks for my roommates, husband Xin Lu, and my Daughters Lulia Lu and Klara Lu, for their laugh and background noise during my writings.