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

Evaluation of blood-based microRNAs toward clinical use as biomarkers in common and rare diseases

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

der UNIVERSITÄT DES SAARLANDES
2022

vorgelegt von Mustafa Kahraman
geb. am 24.12.1986 in Olpe, Deutschland

Student:
"Dr. Einstein, aren't these the same questions as last year's [physics] final exam?"

Dr. Einstein:
"Yes, but this year the answers are different."

Abstract

According to the GLOBOCAN project of the International Agency for Research on Cancer, the top three common cancer diseases worldwide in the year 2020 were breast, lung and colorectal cancer. These are usually diagnosed via imaging methods (e.g. computer tomography) or invasive methods (e.g. biopsy). However, these techniques are potentially risky and expensive and thus not accessible to all patients, resulting in most cancers being detected in an advanced stage. Since the discovery of small non-coding RNAs and specifically microRNAs and their role as gene regulators, many researchers investigate their association with disease development. In particular, researchers examine body fluid based microRNAs which could present potential cost-effective and minimally- or non-invasive alternatives to the previously described established diagnosis methods.

This dissertation focuses on microRNAs and investigates their suitability as minimally-invasive blood-borne biomarkers for potential diagnostic purposes. More specifically, the goals of this work are (1) to implement a new method to predict novel microRNAs, (2) to understand stability and characteristics of these small non-coding RNAs, possibly relevant for the last goal, (3) to discover potential diagnostic biomarkers in common and rare diseases. The first goal was addressed by developing miRMaster, a web service to predict new microRNAs. The tool uses machine learning and high-throughput sequencing data to find microRNA candidates that follow the known biogenesis pathways. The second goal was pursued in four publications. First, we performed a large scale evaluation of miRMaster by generating a high-resolution map of the human small non-coding RNA transcriptome for which we analyzed and validated potential microRNA candidates. Next, we examined the influence of seasonal effects on microRNA expression profiles and observed the largest difference between spring and the other seasons. Additionally, we evaluated the evolutionary conservation of small non-coding RNAs in zoo animals and showed that the distribution of sncRNA classes varies across species, while common microRNA families are present in more diverse organisms than assumed so far. Furthermore, we analyzed if microRNAs are technically stable, and whether biological variation is preserved when using capillary dried blood spots as an alternative sample collection device to venous blood specimens. Finally, we investigated the suitability of microRNAs as biomarkers for two diseases: lung cancer and Marfan disease. We identified blood-borne biomarker candidates for lung cancer detection in a large-scale multi-center study via machine learning. For the rare Marfan disease we analyzed the paired messenger RNA and microRNA expression levels in whole-blood samples. This highlighted several significantly deregulated microRNAs and messenger RNAs, which we subsequently validated in an independent cohort.

In summary, this thesis provides valuable results toward potential clinical use of microRNAs, and the herein described projects represent comprehensive analyses of them from different perspectives: starting with microRNA discovery, addressing various technical and biological questions and ending with the potential use as biomarkers.

Zusammenfassung

Nach Angaben des GLOBOCAN-Projekts der International Agency for Research on Cancer sind die drei häufigsten Krebserkrankungen weltweit im Jahr 2020 Brust-, Lungen- und Darmkrebs. Diese werden in der Regel durch bildgebende Verfahren (z.B. Computertomographie) oder invasive Methoden (z.B. Biopsie) diagnostiziert. Diese Verfahren sind jedoch potenziell risikoreich und teuer und daher nicht für alle Patienten zugänglich. Dies führt dazu, dass die meisten Krebsarten erst in einem fortgeschrittenen Stadium entdeckt werden. Seit der Entdeckung der kurzen nichtkodierenden RNAs und insbesondere der microRNAs und ihrer Rolle als Genregulatoren untersuchen viele Forscher ihren Zusammenhang mit der Krankheitsentwicklung. Insbesondere untersuchen die Forscher die in Körperflüssigkeiten vorkommenden microRNAs, die potenziell kosteneffiziente und minimal- oder nicht-invasive Alternativen zu den bisher beschriebenen etablierten Diagnosemethoden darstellen könnten.

Diese Dissertation konzentriert sich auf microRNAs und untersucht deren Eignung als minimal-invasive blutbasierte Biomarker für potenzielle diagnostische Zwecke. Genauer gesagt sind die Ziele dieser Arbeit (1) die Implementierung einer neuen Methode zur Vorhersage neuartiger microRNAs, (2) das Verständnis über die Stabilität und Charakteristika dieser kurzen nicht-kodierenden RNAs, die möglicherweise für das nächste Ziel relevant sind, (3) die Entdeckung potenzieller diagnostischer Biomarker für verschiedene Anwendungen. Das erste Ziel wurde durch die Entwicklung von miRMaster verfolgt, einem Webdienst zur Vorhersage neuer microRNAs. Das Tool nutzt maschinelles Lernen und Hochdurchsatz-Sequenzierungsdaten, um microRNA-Kandidaten zu finden, die den bekannten Wege der Biogenese folgen. Das zweite Ziel wurde in vier Veröffentlichungen verfolgt. Zunächst führten wir eine groß angelegte Evaluierung von miRMaster durch, indem wir eine High-Resolution Map des menschlichen Transkriptoms kurzer nichtkodierender RNAs erstellten, für die wir potenzielle microRNA-Kandidaten analysierten und validierten. Anschließend untersuchten wir den Einfluss saisonaler Effekte auf die microRNA-Expressionsprofile und beobachteten den größten Unterschied zwischen dem Frühling und den anderen Jahreszeiten. Darüber hinaus untersuchten wir die evolutionäre Erhaltung kurzer nichtkodierender RNAs in Zoo-Tieren und zeigten, dass die Verteilung der kurzer nichtkodierenden RNA-Klassen zwischen den Arten variiert, während gemeinsame microRNA-Familien

in verschiedeneren Organismen vorkommen als bisher angenommen. Darüber hinaus analysierten wir, ob microRNAs technisch stabil sind und ob die biologische Variation erhalten bleibt, wenn kapillares Trockenblut als alternatives Probenentnahmeverfahren zu venösen Blutproben verwendet werden. Schließlich untersuchten wir die Eignung von microRNAs als Biomarker für zwei Krankheiten: Lungenkrebs und Marfan-Krankheit. In einer groß angelegten multizentrischen Studie identifizierten wir mit Hilfe von maschinellem Lernen Biomarker-Kandidaten aus dem Blut für die Erkennung von Lungenkrebs. Für die seltene Marfan-Krankheit analysierten wir die gepaarten Expressionsniveaus von messengerRNA und microRNA in Vollblutproben. Dabei wurden mehrere signifikant deregulierte microRNAs und messengerRNAs festgestellt, die wir anschließend in einer unabhängigen Kohorte validierten.

Zusammenfassend lässt sich sagen, dass diese Arbeit wertvolle Ergebnisse im Hinblick auf die potenzielle klinische Verwendung von microRNAs liefert. Die hier beschriebenen Projekte stellen umfassende Analysen aus verschiedenen Blickwinkeln dar: angefangen bei der Entdeckung von microRNAs, über verschiedene technische und biologische Fragen bis hin zur potenziellen Verwendung als Biomarker.

Scientific papers

This is a cumulative thesis based on the following published papers. The publications included herein are identical to the published versions.

- **M. Kahraman**, T. Laufer, C. Backes, H. Schrörs, T. Fehlmann, N. Ludwig, J. Kohlhaas, E. Meese, T. Wehler, R. Bals, and A. Keller. Technical Stability and Biological Variability in MicroRNAs from Dried Blood Spots: A Lung Cancer Therapy-Monitoring Showcase. *Clinical Chemistry*, 63(9):1476–1488, Sep 2017. [1]
- T. Fehlmann†, **M. Kahraman**†, N. Ludwig, C. Backes, V. Galata, V. Keller, L. Geffers, N. Mercaldo, D. Hornung, T. Weis, E. Kayvanpour, M. Abu-Halima, C. Deuschle, C. Schulte, U. Suenkel, A.K. von Thaler, W. Maetzler, C. Herr, S. Fähndrich, C. Vogelmeier, P. Guimaraes, A. Hecksteden, T. Meyer, F. Metzger, C. Diener, S. Deutscher, H. Abdul-Khaliq, I. Stehle, S. Haeusler, A. Meiser, H.V. Groesdonk, T. Volk, H.P. Lenhof, H. Katus, R. Balling, B. Meder, R. Kruger, H. Huwer, R. Bals, E. Meese, and A. Keller. Evaluating the Use of Circulating MicroRNA Profiles for Lung Cancer Detection in Symptomatic Patients. *JAMA Oncology*, 6(5):714–723, May 2020. [2]
- T. Fehlmann, C. Backes, **M. Kahraman**, J. Haas, N. Ludwig, A.E. Posch, M.L. Würstle, M. Hübenthal, A. Franke, B. Meder, E. Meese, and A. Keller. Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic acids research*, 45(15):8731–8744, Sep 2017. [3]
- T. Fehlmann, C. Backes, J. Alles, U. Fischer, M. Hart, F. Kern, H. Langseth, T. Rounge, S.U. Umu, **M. Kahraman**, T. Laufer, J. Haas, C. Staehler, N. Ludwig, M. Hübenthal, B. Meder, A. Franke, H.P. Lenhof, E. Meese, and A. Keller. A high-resolution map of the human small non-coding transcriptome. *Bioinformatics*, 34(10):1621-1628, May 2018. [4]
- T. Fehlmann, C. Backes, M. Pirritano, T. Laufer, V. Galata, F. Kern, **M. Kahraman**, G. Gasparoni, N. Ludwig, H.P. Lenhof, H.A. Gregersen, R. Francke, E. Meese, M. Simon, and A. Keller. The sncRNA Zoo: a repository for circulating small noncoding RNAs in animals. *Nucleic Acids Research*, 47(9):4431–4441, May 2019. [5]
- N. Ludwig, A. Hecksteden, **M. Kahraman**, T. Fehlmann, T. Laufer, F. Kern, T. Meyer, E. Meese, A. Keller, and C. Backes. Spring is in

the air: seasonal profiles indicate vernal change of miRNA activity. *RNA Biology* 16:8, 1034-1043, Aug 2019. [6]

- M. Abu-Halima, **M. Kahraman**, D. Henn, T. Rädle-Hurst, A. Keller, H. Abdul-Khaliq, and E. Meese. Deregulated microRNA and mRNA expression profiles in the peripheral blood of patients with Marfan syndrome. *Journal of translational medicine*, 16(1):60, Mar 2018. [7]

Contents

Abbreviations 15

1 Introduction 17

1.1 Diseases 17

1.1.1 Lung cancer 17

1.1.2 Marfan Syndrome 20

1.2 MicroRNA 21

1.2.1 MicroRNA basics 21

1.2.2 Biogenesis in human 22

1.2.3 Biological functionality and mechanism of action 23

1.2.4 MicroRNA databases 24

1.2.5 Tissue of origin 26

1.3 Computational methods 27

1.3.1 MicroRNA analyses 27
  1.3.1.1 Overview of microRNA analyses 27
  1.3.1.2 MicroRNA prediction 28

1.3.2 Web services - broad usage of user-friendly applications 32

1.4 Technical basics 32

1.4.1 Platforms for miRNA profiling 32
  1.4.1.1 Next Generation Sequencing (NGS) 33
  1.4.1.2 Nanostring 34

1.4.1.3 Microarray 35

1.4.1.4 Real-Time qPCR 36

1.4.2 Blood collection devices 37

1.4.2.1 PAXgene - Sampling in clinical routine 38

1.4.2.2 Dried blot spots (DBS) & Mitra - Options for self- and home-sampling 39

2 Goals of the PhD thesis 41

3 Results 45

3.1 Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs 46

3.2 A high-resolution map of the human small non-coding transcriptome 60

3.3 Technical Stability and Biological Variability in MicroRNAs from Dried Blood Spots: A Lung Cancer Therapy-Monitoring Showcase 68

3.4 The sncRNA Zoo: a repository for circulating small noncoding RNAs in animals 81

3.5 Spring is in the air: seasonal profiles indicate vernal change of miRNA activity 92

3.6 Evaluating the use of circulating microRNA profiles for lung cancer detection in symptomatic patients 102

3.7 Deregulated microRNA and mRNA expression profiles in the peripheral blood of patients with Marfan syndrome 112

4 Discussion, outlook and conclusion 131

4.1 Discussion 131

4.2 Outlook 133

4.3 Conclusion 136

Bibliography 137

Acknowledgement 157

Curriculum Vitae 159

List of Figures

1.1 Identified proportion of lung cancer patients and life expectancy 18

1.2 Circulating tumor DNA in the blood 19

1.3 Biogenesis of miRNAs in human 23

1.4 Number of miRNAs released in different miRBase versions 25

1.5 MicroRNAs determined by different platforms 33

1.6 Sequencing by synthesis 34

1.7 Hybridization-based measurement 35

1.8 Principle of qPCR 37

1.9 Lysis and RNA stabilization 38

2.1 Overview of goals and studies of this thesis 41

List of Tables

1.1 Overview of the listed miRNA prediction tools 31

Abbreviations

A

AGO Argonaute

B

bp basepair

C

cfDNA cell-free DNA
COPD chronic obstructive pulmonarydisease
CT computed tomography
CTC circulating tumor cells
ctDNA circulating tumor DNA

D

DBS dried blood spot

M

MFE minimum free energy
MFS Marfan syndrome
miRISC precursor miRNA
miRNA microRNA

N

NGS next-generation sequencing
nt nucleotide

P

PBMCs peripheral blood mononuclear cells
PCA principal component analysis
PCR polymerase chain reaction
piRNA Piwi-interacting RNA
pre-miRNA precursor miRNA
pri-miRNA primary miRNA

R

RBC red blood cell

S

SEER Surveillance, Epidemiology, and End Results
siRNA small interfering RNA
sncRNA small non-coding RNA
snoRNA small nucleolar RNA
snRNA small nuclear RNA

T

t-SNE t-distributed stochastic neighbor embedding
tRNA transfer RNA

U

UMAP uniform manifold approximation and projection

W

WBC white blood cell

1

Introduction

This chapter introduces the motivation and fundamentals for using microRNAs (miRNAs) as blood-based markers. Research over the last 10 years has shown that this particular small non-coding RNA (sncRNA) class has been implicated in the development of cancer [8], with lung cancer often being of research interest as one of the most common and deadly cancer diseases [9]. In this regard, I present here that miRNAs may be an interesting minimally-invasive diagnostic alternative to conventional diagnostic methods such as invasive biopsy methods or non-invasive possibly radiation-invasive imaging methods for early detection of this disease [10]. In addition, we raised the question of whether the use of miRNAs is also transferable to analysis of genetic-based Marfan syndrome, a fundamentally different disease to lung cancer. Furthermore, this chapter will highlight the structural, biogenesis and occurrence properties that are valuable for the discovery of new miRNAs by prediction methods and for the understanding of blood-based markers. Lastly, I will present sampling methods and the profiling platforms. I will describe the classical whole-blood PAXgene tube as well as dried sampling methods that seems to be suitable for home-sampling. Profiling of miRNAs on multiple platforms with different applications will also be addressed in this chapter.

1.1 Diseases

In this section, two different diseases will be presented. First, there is lung cancer, a common disease, which is often diagnosed too late for most patients [11]. Second, there is Marfan syndrome, a rare genetic disease, the diagnosis of which is currently done in a multiple-step approach by a group of medical doctors of different fields [12]. These challenges led to the research of biomarker-based diagnostic tools to provide quick and simplified diagnoses.

1.1.1 Lung cancer

Brief history and causes of lung cancer

While in the beginning of the 20th century documented cases of lung cancer suggested it was a rare disease, today, more than 100 years later, it is the most common cancer disease alongside breast cancer and one of the leading causes of death worldwide with around 1.8 million deaths in 2020 [13, 14]. This vast number can be explained with better identification through improvement of the medical infrastructure, wider interest in lung cancer research and broader access to medical facilities. Although in the first half of the 20th century physicians in Germany and the US were able to associate smoking with lung cancer development, the general public and the majority of medical doctors ignored this link until the Surgeon General's report in 1964. This report, based on more than a thousand of articles relating to smoking and disease, recognized the use of tobacco as a main cause of lung cancer [15, 16]. While for a long time the squamous cell carcinoma with central localization was the predominant lung cancer form among smokers, today the focus is on peripherally located adenocarcinoma, likely connected to the changes made in the cigarette productions [17]. However, there are several non-smoking risk factors which can cause lung cancer as well. Genetic predispositions alongside personal exposure to passive smoking, radon, or occupational exposures (asbestos, metals, silica, etc.) and even air pollution all lead to a risk of developing the disease [18]. It is worth mentioning that chronic obstructive pulmonary disease (COPD), a common lung disease, often precedes by tumor development [19].

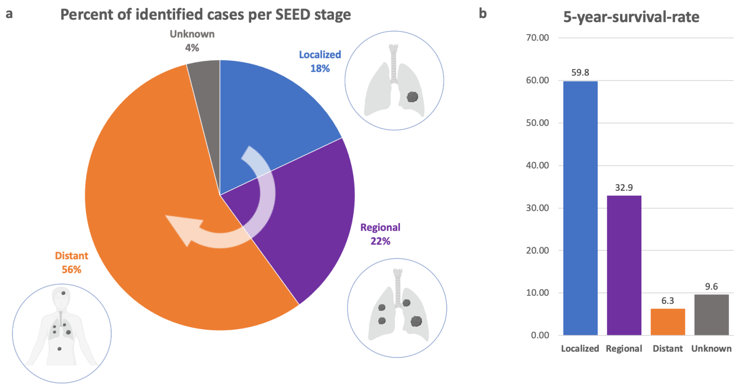


Figure 1.1: **Identified proportion of lung cancer patients and life expectancy.** (a) SEED stages. Localized: Cancer is circumscribed to its primary location. Regional: Cancer expands to regional lymph nodes. Distant: Metastases have developed. Unknown: Cancer is unstaged. (b) The 5-year-survival rate decreases with increasing stage. Created with BioRender.com

Life expectancy and current methods in diagnosis and treatment

According to the annual report of the National Cancer Institute in the United States, the 5-year-survival-rate by SEER stages (accessible at http://www.seer.cancer.gov/) was 6.3% among patients diagnosed in the distant stage when metastasis has already begun. The highest survival chance with 59.8% is when the patient's diagnosis happens in the localized stage. More than half of the identified lung cancer patients were recognized in distant stage with low survival perspectives, while a minority of 18% with potential longer life is identified in the localized stage [20] (see Figure 1.1). The main current method for disease detection is the non-invasive x-ray-based low-dose computed tomography

(CT) scan. It was applied on for the first time in 1979 and was further developed over time. The method captures images of cross-sections of parts of the body and thus enables the observation of abnormalities in different layers or levels of the examined body or tissue region [21]. The exposure to radiation from low-dose CT seems to be less dangerous for a patient in a high-risk population than the annual chest radiography or conventional CT scan [22, 23]. Due to its high rate of false positive results, the elevated radiation exposure in subsequent diagnostic CT scans and an increased amount of overdiagnosis, the worldwide and general implementation of low-dose CT screening is still subject to debate [23, 24]. A helpful tool to prioritize patients for screening or to help the prevention of unnecessary invasive diagnostic examinations in finally benign findings may be the additional use of molecular biomarkers. The blood-based analysis of tumor-associated proteins as well as circulating tumor cells (CTC) or circulating tumor DNA (ctDNA) is known as liquid biopsy. Alternative non-invasive sampling methods are saliva or sputum collection or even exhaled breath condensate sampling [25–28]. However, the predominant markers are CTCs and ctDNA. CTCs are essentially metastases *in transit*. However, their ultra-low concentration in blood (~1-10 cell/ml blood [29]) leaves them very difficult to detect in early-stage disease settings [30]. ctDNA is part of cell-free DNA (cfDNA) that typically gets shed from dying tumor cells into peripheral blood (see Figure 1.2) [31].

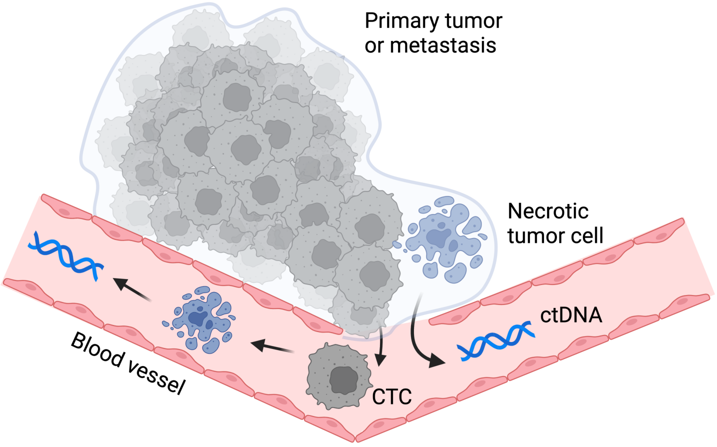


Figure 1.2: **Circulating tumor DNA in the blood.** Tumor DNA in the blood is released from necrotic tumor cells located either in a tumor (primary tumor or metastasis) or in the bloodstream. Created with BioRender.com

Due to its short half-life of about 35 min [32], ctDNA is a dynamic marker and has been used successfully in late-stage tumor monitoring and to detect minimal residual disease after surgery. One challenge of ctDNA is that its abundance in peripheral blood is relatively low – only a small fraction of cfDNA is actually ctDNA – requiring very deep sequencing or sensitive polymerase chain reaction (PCR) assays. Interestingly, different tumor types exhibit different ctDNA shedding rates, making it challenging to develop a ctDNA based assays for early detection for multiple cancer types. For example, the CancerSEEK assay (61 ctDNA amplicons + 8 proteins) applied on early-stage cancers (stage I) shows the lowest detection rate for esophageal cancer (20%) and the highest for liver cancer (100%) [33], whilst the GRAIL's Galleri methylation based cfDNA assay detects only 23% of stage I lung cancers [34]. For this reason, it is necessary to also assess alternative, or complementary biomarkers, such as small RNAs. A promising group of biomarkers besides proteins or ctDNA are miRNAs, small non-coding RNAs which take part in the post-transcriptional orchestration of metabolic and regulatory pathways. Their alteration in several pathophysiological conditions such as manifold cancers, neurodegenerative or cardiovascular diseases has already been demonstrated in our research as well as others' [28, 35–38]. For cancer therapy, there are a number of treatments depending on the type and how advanced the tumor is. To cure the cancer, the usage of surgery, radiotherapy or chemotherapy are recommended, while to improve the situation for advanced cancer patients supportive and palliative care is used [39].

1.1.2 *Marfan Syndrome*

Brief history and causes of Marfan syndrome

The Marfan syndrome (MFS), which was first reported by the pediatrician Antoine Marfan in 1896, is a rare and multi-system disorder with a prevalence of 1 in 5,000 individuals in the general population [40, 41]. In 1931 Henriculus Weve first discovered that the cause is located at chromosome 15, more precisely 15q-21.1, and that it is inherited in an autosomal-dominant manner [42]. Later in 1991 Harry C. Dietz identified mutations on the fibrillin-1 (FBN1) gene [43]. Since that first mutation discovery, over 600 different mutations connected to this protein have been reported for MFS and related disorders [44, 45]. The disease comes along with skin, cardiovascular, skeletal and ocular symptoms [46].

Current methods in diagnostics and treatment

To currently diagnose MFS, a pediatrician must identify anamnesis, and more crucially a "classic triad," consisting of ocular, cardiovascular and musculoskeletal abnormalities. Other doctors such as human geneticist, oculist, cardiologist and orthopedist can aid in the final diagnosis [12, 47, 48]. The current therapeutic strategies aim to avoid complications for the patient. This is implemented during regular medical check-ups by the above-mentioned medical specialists. One of the most important treatments is the annual monitoring of the heart and the administration of beta-blockers to prevent a sudden cardiac arrest through aortic dissection, [49]. In addition, surgeries treating scoliosis or ocular abnormalities can help to alleviate medical conditions [40, 50]. With the improvement of therapy, the mean life expectancy in 1990s has been significantly increased to 41 compared to the age of 32 in 1972 [51]. Due to the fact that the disorder is genetic-based, potential solutions of the gene therapy using catalytic nucleic acid molecules for correction of the mutation as the cause itself was already discussed in the review paper of Phylactou and Kilpatrick in 1999 [52]. With the development of the precise gene editing technology based on clustered regularly interspaced short palindromic repeats (CRISPR), a new promising approach to cure MFS has emerged [53]. In this process, the Cas9 protein transferred into the cell finds the target region with a guided single stranded sequence and cleaves the target DNA sequence. At this point this region can be edited by modifying, deleting or inserting new sequences [54, 55]. The CRISPR/Cas9 technology used as a copy-and-paste tool opens the door for healing genetic-based diseases such as MFS.

1.2 *MicroRNA*

Image-based instruments such as computer tomography are currently the gold standard for diagnostics to see intra-body pathologic-morphological tissue modifications caused by cancer [56]. However, the situation with the expensive procurement of these technologies and the late diagnosis for many diseases calls for low-cost and early diagnostic alternatives. Worldwide research efforts have started to work on different type of blood-based biomarkers [57, 58]. In particular, miRNAs, which will be introduced in this section, have come to the forefront of research as their potential was revealed.

1.2.1 *MicroRNA basics*

MicroRNAs were introduced to the science world in 1993 [59]. The discovery of cel-lin-4 by Lee, Feinbaum and Ambros in 1993 [59] and cel-let-7 by Reinhart et al. and 2000 [60] started a new field in biological and medical research. Gene regions that were considered junkDNA were later associated with regulatory RNAs that control mRNA translation [61, 62]. The miRNA genes belong to the small non-coding RNAs (sncRNAs) which are also composed of other classes such as transfer RNA (tRNA), small nucleolar RNA (snoRNA), small nuclear RNA (snRNA), small interfering (siRNA) and piwi-interacting (piRNA) [63]. So far, over 85,000 human transcripts of these sncRNAs have been found and annotated [64]. These small molecules – with a length of 17-28 nucleotides (nts) – are present in human, plants and other organisms. miRNA sequences that are similar across species share conserved seed regions, which are the main binding region for target identification. To find similar miRs in other species, one can apply two approaches using assumed conservation: homology- and sequence-based approaches that ignoring expression profiles and expression profile based approaches, e.g. between human and rat can be conserved [65, 66]. When miRNAs were found to play a central role in several species, especially in humans, they quickly gained attention, leading to the development of hundreds of bioinformatics-tools which allow the evaluation of the plethora of generated miRNA-based datasets today [67]. By determining correlations to pathologies, miRNAs became highly interesting molecules as potential predictive biomarkers for diagnostics and prognostics on one side and potential players for therapeutic approaches on the other side [68].

1.2.2 *Biogenesis in human*

Human mature miRNAs are transcribed RNA molecules from miRNA genes. The initial longer transcript, which is called primary miRNA (pri-miRNA), undergoes several processing steps, starting in the nucleus (see Figure 1.3), and then becomes the final mature miRNA in the cytoplasm. miRNAs are transcribed mainly by RNA Polymerase (Pol II), but Pol III can transcribe miRNAs as well [69, 70]. The primary transcribed molecule can usually be over 100 basepairs (bps) long and contains a stem sequence of around 35 bps where the mature miRNA is located [71]. Some primary miRNAs can produce multiple mature miRNAs. These long transcripts are called clusters [72]. In the nucleus, Drosha, a ribonuclease III enzyme, and its co-factor DGCR8 process the pri-miRNA to a shorter hairpin structure called precursor miRNA (pre-miRNA) with 2 nt 3′ overhang [73]. Afterwards, the new molecule is transferred to the cytoplasm, where its loop is removed by the RNA III endonuclease Dicer [74]. The resulting mature miRNA duplex is loaded into the Argonaute (AGO) which forms the miRNA-induced-silencing-complex (miRISC) [75]. To prepare the final version of this complex for RNA targeting, one of the two strands, the passenger strand (*) is ejected and degraded, while the guide strand is maintained for targeting. The seed region of a mature miRNA, starting from base position 2 with a length between 6 and 8 nts, is the main region for target recognition [70]. In the following, the sequential and structural features of the general precursor miRNA are described. These are important for biogenesis and ultimately *de novo* miRNA prediction. Usually, a precursor is ~60 bps long and has a 2 nt overhang on the 3′ end so when the stem structure is formed the 5p- and 3p-arm are shifted [73]. It includes bulges and a terminal loop, which can influence preprocessing by Drosha and Dicer [76]. Flexible terminal loop regions allow a facilitated processing for both enzymes, while mutations inside that region conduce base pairing and shorter loops, leading to possible inhibition of miRNAs biogenesis [76]. Finally, the minimum free energy (MFE) is worth to be mentioned as a significant feature indicating the stability of a structure. It describes how much energy is needed to keep the specific form of secondary structure. The lower the MFE for a secondary structure is, the more stable the form [77]. The base composition has an effect on the MFE and, in the end, on stability [78]. The MFE is a good indicator if a precursor would fold into a stable hairpin structure or an alternative form which is less likely to match the miRNA biogenesis and to be processed by Dicer [79]. Hence, this feature can be used for miRNA prediction.

For a long time, there was the general acceptance that nearly each

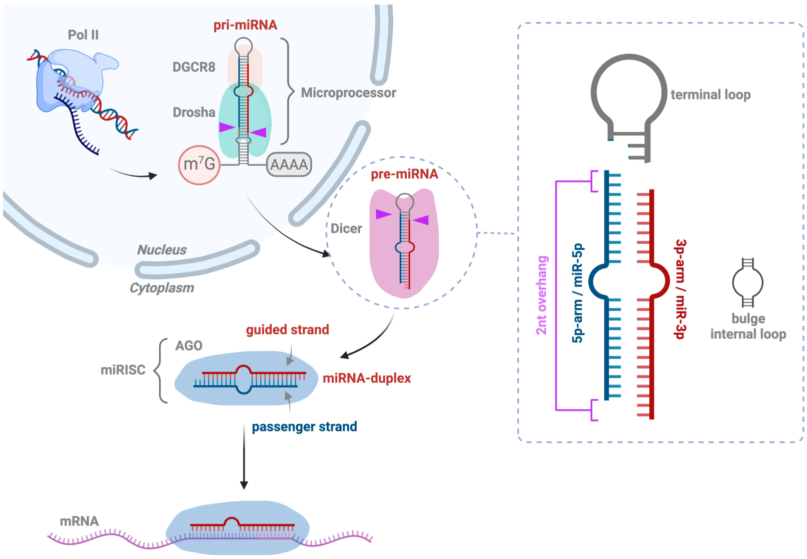


Figure 1.3: **Biogenesis of miRNAs in human.** Pol II transcribes a miRNA gene to a pri-miRNA which is processed to a pre-miRNA by Drosha and its co-factor DGCR8. This precursor is cleaved by Dicer to a miRNA-duplex which is loaded into AGO. The resulting miRISC binds to the corresponding target to destabilize the RNA or inhibit its translation. Created with BioRender.com

precursor generates one mature miRNA for each arm. However, recent sequencing studies have shown that there can be transcribed alternative forms (isomiRs) besides the canonical sequence documented in miRBase [80]. Based on the type of modifications, isomiRs can be grouped into three major groups: polymorphic, 5′- or 3′-isomiRs. While the polymorphic variants have internal substitutions, additions or deletions, the 5′- and 3′-isomiRs have modifications in the respective tail regions. These can be divided into template and non-template-based variations [81]. While the template-based modification represents bases from the precursor, the non-template variation can be a result of polyadenylation [82]. Wu et al. showed that the two-thirds of the found miRNAs in their dataset of their 81 colorectal tissue samples are non-canonical miRNAs. While the majority of these are 3′ isomiRs with additions [81], isomiR expression usually correlates highly with canonical miRNA expression [83]. Nevertheless, modifications in the seed region can lead to different targets and functions [83]. Several studies have shown that the variations can be dependent on population, gender and age [84–86], but the isomiR generation can also be cell-type specific and disease-dependent [86, 87].

1.2.3 *Biological functionality and mechanism of action*

Although miRNAs are short sequences, they play a major part in post-transcriptional gene regulation. They are involved in physiological processes of the cell cycle such as proliferation, differentiation and apoptosis [88]. Taking the first discovered miRNAs as an example, cel-lin-4 and cel-let-7 play a part in the developmental transitions in larval stages of *C. elegans* [88, 89]. In 2007, Shcherbata et al. found out that dme-miR-8 and dme-bantam are required for adult germline stem cell maintenance in *Drosophila* [90]. In addition to that, hsa-miR-449 could play a cross-functional role in cell fate determination regarding cell death, cell differentiation or cell cycle arrest in humans [91]. The gene regulation occurs at the post-transcriptional RNA level targeting mRNAs by Watson-Crick basepairing. Here, one miRNA can have multiple different targets on the one hand, and one target can be targeted by many different miRNAs on the other hand [92]. There are three mechanisms of miRNA targeting. While in plants the commonly seen perfect binding to the target leads to mRNA cleavage, the partial binding which is predominant in animals results in translational repression [65, 93]. The last mechanism is the destabilization of the target RNA by removing the protecting ends (m7G on the 5′-end and poly-A-tail on the 3′-end) causing degradation. Thereby, the GW182 protein bound to the miRISC complex recruits CCR4-NOT deadenylase for digesting the poly-A tail and DCP2 with associated proteins for decapping the RNA. Finally, the decapped mRNA can be degraded by the exoribonuclease XRN1 [72]. The miRNA-mediated mRNA regulation in animals occurs by one of the mechanisms (repression or destabilization) or sometimes by the combination of both shown in *Drosophila melanogaster* [94]. Overall, all scenarios typically yield anti-correlated miRNA-mRNA-expression and protein translation of the target gene [95]. However, in contrast to their traditional role in gene regulation, recently it was observed that an overexpressed miRNA can also up-regulate its target. Gu et al. showed that miR-191 increased the mRNA expression of the antiangiogenic factors p21 and tissue inhibitor of metalloproteinase-1 [96]. Tan et al. found that the positively-correlating miRNA-gene-pairs are consistent across various human cancer types of the The Cancer Genome Atlas [97, 98].

1.2.4 *MicroRNA databases*

Since miRBase was created in 2002, it is the most well-known public online database for miRNA sequences and their annotation. In the beginning, the database contained 218 hairpin precursors and 218 mature miRNAs of five different species (miRBase v1.0 in December 2002). Two decades later, 38,589 precursors and 48,860 mature miRNAs from 271 organisms exist in the current release 22.1 (October 2018). The immense increase can be explained by the shift of traditional expression and experimental validation methods – northern blot and reverse-transcription quantitative polymerase chain reaction (RT-qPCR) towards novel high-throughput methods by next generation sequencing (NGS) combined with computational *ab-initio* miRNA prediction approaches [99, 100]. The NGS approaches have created novel sequence data where several computational miRNA prediction methods were implemented. These generated a high number of miRNA candidates alongside the challenge to validate them experimentally. With the increasing usage of NGS, the number of predicted miRNAs grew rapidly. When considering the miRBase v1 - v10, 82 miRNAs were introduced in total in the repository by NGS validation. However, for v11 to v20 there were already 1505 NGS-validated new miRNAs [101]. Without experimental validation of these large number of new entries, there is a risk of overestimation of total numbers of miRNAs in the miRBase content because of potential false positives. [100]. The latter observation describes one limitation of miRBase. Another important observation is the absence of candidate molecules that other research groups have already found but not updated or updated much later in miRBase [101]. Focusing on the human organism, we can find 1,917 precursors and 2,656 mature miRNAs in the current miRBase version of 2018 (see Figure 1.4), while Londin et al. have already published 3494 novel precursors and their 3707 potential mature miRNAs derived from a comprehensive analysis of 13 cell types in 2015 [102].

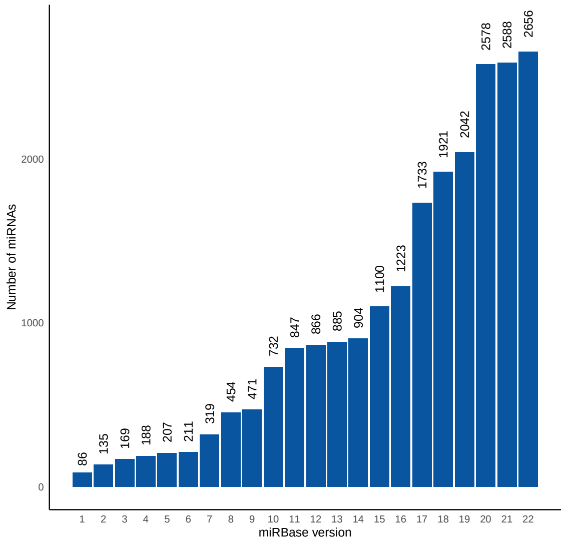


Figure 1.4: **Number of miRNAs released in different miRBase versions.** The number of new miRNA entries is increasing constantly with each miRBase release. Large increases are observed with the releases from version 17 to 20.

In addition, there is a chance that the same not yet reported novel miRNA can be found by several researchers without them noticing, since the biggest central repository is not always up-to-date. To avoid this issue, miRCarta was developed with the focus on human, and its content was derived from more than 18,000 small RNA sequencing data sets (Sequence Read Archive), The Cancer Genome Atlas (TCGA) and in-house datasets using miRMaster as a prediction tool. The current release contains over 15,000 human precursors and around 25,000 human miRNAs [101].

To understand the biological relevance of annotated miRNAs, databases documenting their functional relevance have been developed. miRecords (last update: April 2013) has 2,705 experimentally validated records of miRNA-target interactions based on over 600 miRNAs and 1,901 target genes of 9 species [103]. Other currently maintained databases are miRTarBase 8.0 (September 2019) and DIANA-TarBase v8 (2017) which both have manually curated entries and around half a million reported target interactions [103, 104]. Another well-known repository, which is the largest one, is miRwalk (launched in December 2014; current version is 3.0) with a collection of around 949 million predicted and experimentally verified miRNA-target interactions [105]. Focused on the human species and mice, miRPathDB 2.0 (released in 2019) provides target-interactions as well, but also associations with particular biological processes or pathways combined with enrichment analysis functionality [106]. Furthermore, it incorporates over 15,000 novel miRNAs listed in miRCarta [101].

1.2.5 *Tissue of origin*

Although miRNAs are universally present in the human body, they show different expression pattern depending on their tissue specificity and temporal local biological conditions [107]. They can reflect physiological and disordered physiological processes such as carcinogenesis, metastasis and drug receptivity in their cells of origin [108, 109]. The tissue specific expression of miRNAs can be useful to determine the primary origin of metastatic cancer with unknown primary tumor [110] or can be supportive to understand blood- or serum-based potential biomarkers for a disease in the respective tissue. To look up calculated abundances of miRNAs in different organs, scientist can use online repositories such as smiRNAdb (Landgraf et al., 2007) [111] or Human miRNA Tissue Atlas (Ludwig et al., 2016) [107]. However, since the tissue specific miRNAs are only obtained via biopsy, medical researchers are highly interested in biomarkers of body fluids (whole blood, serum, plasma, urine, etc.) which are easily collected in a non-invasive to minimally-invasive manner and can be collected at larger scale. The source of these circulating miRNAs are erythrocytes (red blood cells, RBCs), immune cells (white blood cells, WBCs), other circulating cells, exosomes, and general secretion from cells [112]. As such circulating miRNAs can reflect a variety of solid tissue specific observations, some of these can reflect a disease-related signal. There are miRNAs responsible for cell-to-cell communication. For example, tumor-secreted hsa-miR-21-5p and hsa-miR-29a-3p work as ligands to receptors of the Toll-like receptor family with the result of an inflammatory response and potential inducing tumor growth and metastasis [113]. Furthermore, there are miRNAs related to the immune response as part of the innate and adaptive immune system [114]. For example, while miRNAs including hsa-miR-155-5p and hsa-miR-223-3p regulate granulocytes and macrophages of the innate immune system, hsa-miR-17-5p, hsa-miR-31-5p and others are involved in the development of lymphoid immune cells (B- and T-cells) [115]. Despite these miRNAs being part of peripheral blood mononuclear cells (PBMCs), red blood cells form the largest proportion with 99% of the whole blood components. This imbalance results in two problems. First, the high-level blood-borne miRNAs can make it difficult to catch a disease-specific miRNA expression pattern [112, 116]. The miRNA detection bandwidth can turn out smaller using e.g. NGS technologies. Here, the top three dominant miRNAs can comprise over 90% of the total signal and can consequently lead to inaccurate quantification of potential important low abundant miRNAs [117]. The second problem is that a fold change is barely noticeable if the disease-related deregulation of a miRNA is not taking place in RBCs, the dominant cell type in the whole blood.

1.3 Computational methods

In this section, I present different kinds of analyses that can be performed with miRNA-related data. These cover approaches for sample similarity and biomarker discovery, prediction of miRNAs and the identification of miRNA-gene interactions which are the components of miRNA regulatory networks. The last topic introduces web services as a solution for broader usage in the scientific community.

1.3.1 MicroRNA analyses

1.3.1.1 Overview of microRNA analyses

Cluster analysis An important aspect to understand complex and high-dimensional data (each miRNA is a feature/dimension) is data visualization via unsupervised dimensional reduction while preserving the relevant information. Usually, the dimensionality is reduced to two or three dimension where samples with similar patterns (expression profiles) group together as clusters. A common method for dimensionality reduction is principal component analysis (PCA) [118]. It generates orthogonal linear combinations (principal components) of the original variables. In doing so, the first components conserve most of the largest pairwise distances by maximizing the variance explained by each component. Since PCA is based only on linear transformations, the non-linear relationships can be captured by other approaches such as t-distributed stochastic neighbor embedding (t-SNE) [119] or uniform manifold approximation and projection (UMAP) [120]. The idea of t-SNE is to preserve local similarities in non-linear manifold structures. The algorithm computes similarities between pairs of samples in the high-dimensional space (original data) and in the low-dimensional space (starting with a random distributed representation). Then it tries to minimize the two similarity derived probability distributions using a loss function with the result that the clustering in the high-dimensional space is preserved in the low-dimensional one. An alternative dimension reduction method is UMAP which preserves both local and most of the global structure and has better runtime performance according to the authors [121]. Finally a popular method, there is agglomerative hierarchical clustering. Its algorithm builds a binary tree representation of the data by starting with each sample as clusters and merging iteratively similar groups together until ending up with one final cluster (the whole dataset) [122].

Single and multi biomarker discovery Differential expression analysis is applied to quantify changes in expression levels of single miRNAs between experimental groups. By doing comparisons with statistical tests, significant deregulated markers can be identified. If the data follows a normal distribution, which can be checked with the Shapiro-Wilk test [123], then Student's t-test can be performed; otherwise, the choice is a non-parametric test, e.g. Wilcoxon-Mann-Whitney test [124]. In case of multi-group comparisons, techniques like analysis of variance (ANOVA) for normal distributed data and Friedman test for non-normal distributed data can be applied [125]. When using multiple hypothesis testing (multiple miRNAs), the false discovery rate (FDR) must be controlled by adjusting the calculated raw p-values with correction methods such as the Benjamini-Hochberg method [126]. Importantly, the effect sizes of single markers alone are often not strong enough to sufficiently differentiate between experimental conditions. Therefore, multi-marker panels are calculated as best combinations of miRNAs to predict the true health state of a patient. This is done by classification procedures of machine learning, such as support vector machines (SVM), random forest (RF) and others [127].

MicroRNA target prediction and interactions in regulatory network The identification of miRNA targets is essential to understand the resulting biological functions of miRNAs and the regulation of their targets in physiological and pathophysiological processes. Since miRNAs can have multiple target sites because of imperfect binding in mammals [128], the field of miRNA target prediction is growing steadily and offers alternatives to expensive and time-consuming experimental target validations [129]. The following properties form the general concept for most prediction tools: i) complementary target site to the seed region, ii) conservation of seed and target site region across species, iii) stable secondary structure and iv) accessibility of the target sites [130]. Around 100 tools were implemented for this category in the last 20 years [131]. Popular ones are, for example, miRanda, TargetScan and RNAhybrid [132–134]. When piecing the puzzle together, the recognition and verification of interactions between miRNAs and their target genes are the foundation for defining complex regulatory miRNA networks. Concepts of graph theory and gene set enrichment analyses can be helpful for network profiling [135].

MicroRNA prediction Since miRNAs play crucial roles in gene regulation and are involved in disease development, the identification of these sequences are important for downstream analysis. Referring to the next paragraph (1.3.1.2) of this subsection, I will present five selected miRNA prediction tools in detail categorized as homology-based or NGS-based approaches.

1.3.1.2 MicroRNA prediction

MicroRNA identification is an important and challenging step which combines biological and computer-based ideas. The resulting bioinformatic tools for miRNA prediction, which some of them are listed in the following, can be categorized in homology- and NGS-based approaches. While early tools such as MiRScan and MiRSeeker work with sequence similarity, next generation sequencing have enabled the discovery of novel miRNAs and opened the door for the development of NGS-based prediction methods using sequence and structure features [67]. From this category, the following three tools MIReNA, miRanalyzer and miRDeep are presented in addition to the comparative methods MiRScan and MiRSeeker (see Table 1.1).

MiRScan MiRScan, released in 2003, is a web server that identifies miRNAs that are conserved in two nematode genomes, *Caenorhabditis elegans* and *Caenorhabditis briggsae*, and shows feature characteristics of known nematode miRNAs [136]. The candidate sequence is evaluated by similarity and conservation checks on the training set. In detail, the criteria are based on the base pairing of the miRNA portion and rest of the stem-loop, stringent sequence conservation in the seed region, preference to form symmetric internal loops and bulges in the miRNA region, and having a consensus base pairing segment lying between the terminal loop and miRNA regions. For the development of MiRScan, a training set was built on 50 of 53 conserved miRNAs that were reported previously [137, 138]. These were identified by scanning all hairpin structures with conserved sequences in the above-mentioned genomes. Using MFE, MiRScan predicted about 36,000 hairpins that fulfilled the minimal requirements for hairpin structure and sequence conservation. 35 of them were top candidates, whereby 16 sequences could be successfully validated, and the remaining 19 candidates seemed to be false positives.

MiRSeeker Another early published web service is MiRSeeker (released in 2003) which follows a similar concept as MiRScan but to predict *Drosophila* miRNAs [139]. In the first step, conserved miRNA regions are identified via alignments between the genomes *Drosophila melanogaster* and *Drosophila pseudoobscura*. Thereby, alignments related to exons and sncRNAs that are not miRNAs/pre-miRNAas are excluded. The applied conservation criteria are derived from a reference set of 24 Drosophila precursor sequences. There should be not more than 13% gaps or 15% mismatches in a respective 100 nt alignment segment. By window-shifting, the respective genomic neighborhoods are considered, too. In the second step, stem-loops are identified and ranked. When the secondary structure is predicted by mfold [140], the longest helical arm (minimum of 23bp) is used for evaluation. The isolated arm should have a maximum MFE of -23 kcal/mol. In addition, the predicted stem-loop structure is scored by comparing it to the canonical hairpin. While continuous helical pairing is rewarded, deviations as increased size of internal loops, asymmetric loops und bulged nucleotides are penalized. In the third and last step, a candidate is evaluated for its pattern of divergence. It is considered perfectly conserved, when the query sequence lays perfectly inside the helical arm, with a size at least of 23 nts and maximal 10 nt distance to the end of the terminal loop. In this study, the authors validated 24 novel miRNAs of 38 candidates with northern blotting [139].

MIReNA A further non-machine-learning based command line tool is the 2010 released MIReNA, [141]. It searches for genomic locations of the query sequences and checks their surrounding potential pre-miRNA structure. For the evaluation, MIReNA takes five implemented parameters (physical-chemical and combinatorial) into account. The two physicochemical features are "adjusted MFE" (AMFE) and "MFE index" (MFEI) which are based on MFE, sequence length and GC content. While AMFE is the MFE depending on sequence length, MFEI is described by AMFE and the percentage of the GC composition. The remaining three combinatorial structure features are the following: i) a potential miRNA sequence cannot form with itself base pairings within the secondary structure of a precursor; ii) similar length between the 5p and 3p arm of the candidate pre-miRNA; iii) a maximum of 26% of a candidate's miRNA length can be unmatched within the respective precursor secondary structure. The MIReNA algorithm starts with searching for similar input sequences to annotated miRNAs by reducing the number of base variations (insertions, deletions and substitutions of nucleotides). The resulting known miRNA gene location is extended on each by 200 nts. For every resulting sequence extension, secondary structures are computed via RNAfold [142]. Finally, these secondary structures are evaluated by applying the above-mentioned five-parameter-filter to get potential pre-miRNAs. The performance in the confirmation of pre-miRNAs were compared against other predictive tools (miR-ablea, miPred and microPred [143–145]) using three datasets. MiReNA yielded similar good or better results regarding sensitivity, specificity, accuracy or Mathew's correlation coefficient (MCC). The testing dataset consisted of 263 human miRNAs as positive sequences and 265 coding sequences as negative set.

miRanalyzer The miRanalyzer tool published in 2009 is a web server based on a machine learning approach for next generation sequencing data and is the predecessor of the comprehensive analysis srnRNA toolbox [146, 147]. The workflow is implemented in three steps. First, known miRNAs annotated in miRBase are detected by mapping against mature miRs. In the second step, the remaining sequences without any mapping hit are mapped against other small non-coding RNA libraries. The queries with mapping hits will be excluded to reduce potential false positives for miRNA prediction. The last step is the prediction of new miRNAs based on the remaining input sequences. The following features for the classification with random forest were used: read count and structural features such as stem length, loop length, loop GC content, MFE and others. Applied on the datasets of human (hsa), rat (rno) and *C. elegans* (cel), the resulting cross-validated sensitivities in single-species datasets are in the range of 0.74-0.77. By training on one species and using another species as unseen test data, the performances are poor (0.48-0.66). Interestingly, the performance to predict a single-species test set increases with the remaining merged cross-species training set (0.71-0.75). Potential reasons for the performance boost could be the increased size of positive test set members and the presence of conserved miRs among all three species. In conclusion, miRanalyzer and also the previous described tool MIReNA (both NGS-based) focuses on sequence-based features without the consideration of the miRNA biogenesis which is the central idea of the next tool miRDeep2.

miRDeep2 One of the most frequently used prediction tools is the command-line tool miRDeep2 (published in 2011 and its predecessor miRDeep in 2008) [148]. Its basic principle is to reconstruct the biogenesis of candidate miRs from NGS data and assess the resulting putative pre-miRNAs. In the beginning, potential precursors of the input sequences are excised. Thereby, sequences with a maximum of five perfect mappings to the genome are considered. In a 70 nt window search, the candidate miRNA with the highest read stack is chosen and flanked by 70 nts upstream and 20 nts downstream and vice versa. By doing so, two potential pre-miRs are derived, where the candidate miRNA lays either on the 5p or the 3p arm. Afterwards, the secondary structure for each excised pre-miRNA is predicted via RNAfold, and the mapping signature – how reads are mapping to a potential precursor – is derived to evaluate if the simulated biogenesis is matching with products of the Dicer processing. This information is important for the miRDeep2 core algorithm, which is the probabilistic scoring of the structure and signature features of a candidate precursor. This step evaluates the structure for its energetic stability, the presence of sufficient base pairing in the mature part and the divisibility of the mapped read stacks to the typical hairpin structure in form of 5p-, terminal loop and 3p-region. In addition, hundred permutation runs are executed with different signatures on the same hairpin. The idea is that random signatures would decrease the final score because the actual sequence and structure features fit best to each other. The evaluation was done by using sequence data from seven animal clades. While the specificity was around 0.99 for all datasets, the range for sensitivity (predicting known miRBase miRNAs in the respective species) was in the range from 0.71 (sea squirt) to 0.90 (anemone). The true positive rates for human liver and cell lines were 0.79 and 0.81. An explanation could be that the potential underestimated sensitivity could be caused by falsely annotated miRNAs in miRBase which would not follow the miRNA biogenesis.

Category	**Tool**	**Published year**	**Usage type**	**Machine learning**	**Advantages**	**Disadvantages**
Homology based	MiRScan	2003	Web server	-	Find conserved intragenic miRs	Less novel miRs possible
Homology based	MiRSeeker	2003	Web server	-	Find conserved intragenic miRs	Less novel miRs possible
NGS-based	MIReNA	2010	Command-line tool	-	Novel miRs	Increased number of false positives
NGS-based	miRanalyzer	2009	Web server	Random forrest	Novel miRs	Increased number of false positives
NGS-based	miRDeep2	2011	Command-line tool	-	Novel miRs	Increased number of false positives

Table 1.1: **Overview of the listed miRNA prediction tools.** Detailed information about category, published year, usage type, machine learning method (if existing), advantages and disadvantages.

1.3.2 *Web services - broad usage of user-friendly applications*

The immense quantity of data generated in the lab, especially as sequencing is becoming improved and affordable, is challenging with regards to the question how this complex data shall be analyzed. Since software such as BLAST [149] or packages like limma [150] can only be used for common tasks, new tools, packages or self-written workflows are required for complex and large data analysis with new upcoming specific biological questions. However, the implementation or even the use is limited only to scientists with programming knowledge and it also requires systems with sufficient computational power and storage. The lack of both can lead to the issue that a lot of data cannot be analyzed on time, especially data of scientists without programming knowledge. To solve this issue and to reach the largest possible number of users, web services can provide user-friendly environments for data analysis and databases. While a user on the client side can focus on the analysis by selecting parameters and uploading data, the developers on the server side are responsible for the administration and maintenance of hardware and software. Ultimately, whether an application is used depends not only on the underlying bioinformatics analysis program and administration (backend), but above all on a good graphical user interface (frontend) with which a user directly interacts. A good usability of a web application can be achieved via interaction design [151]. This design focuses on the dialogue between the user and the digital product which includes flow and orientation (the user knows what to do next), immediate feedback (the user receives a response or progress indicator for each interaction) and ways of interaction (how a user should interact, e.g. using input fields or sliders for numeric parameters). In addition, frontend styling frameworks, such as Bootstrap [152], are used to generate visually appealing interfaces which can increase the usability regarding navigation, feedback and interaction. In conclusion, the trend regarding bioinformatics web services shows that the development and usage of online tools has become increasingly popular in the last 10 years [153]. The progress of this topic is also supported by journals such as Nucleic Acids Research, which publishes an annual issue on the subject.

1.4 *Technical basics*

1.4.1 *Platforms for miRNA profiling*

In this section, three major laboratory platforms for expression profiling and discovery of DNA or RNA molecules are presented. While next generation sequencing has the advantage to discover novel biomarkers [154, 155], Nanostring, microarray- and PCR-based approaches are used for expression profiling of known miRNAs [156, 157] (see Figure 1.5). In addition, the PCR technology is the tool of choice for validation of candidate biomarkers [158].

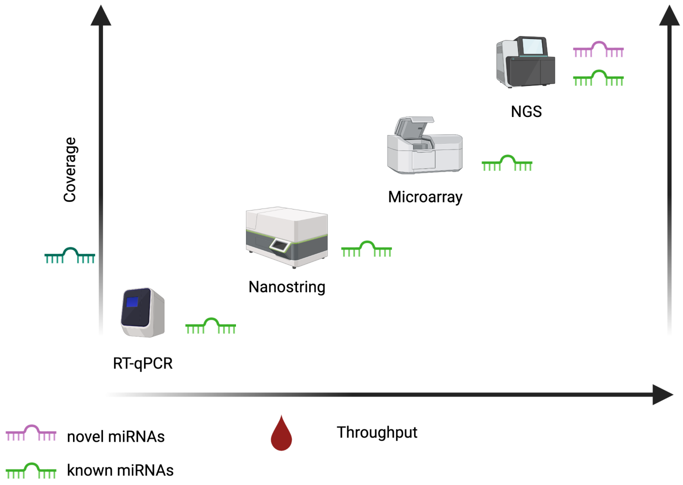


Figure 1.5: **MicroRNAs determined by different platforms.** In relation to cost-efficiency, RT-qPCR can be a suitable tool to measure a small set of known miRNAs. Microarrays and Nanostring can cover more known miRNAs than RT-qPCR. In case of discovering novel miRNAs, NGS platforms are the choice of method. Created with BioRender.com

1.4.1.1 *Next Generation Sequencing (NGS)*

Different sequencing methods

Although, Sanger sequencing was important to complete the Human Genome Project in 2003, the need for new sequencing technologies with reduced costs and increased throughput emerged to replace the known first generation method. The new methods called ”next generation sequencing” (NGS) provided massively parallel sequencing and high throughput lower costs. Therefore, they have been gradually established from the mid 2000s until today [159]. While the short-read sequencing belongs to the second generation, the third generation is characterized by long-read sequencing [160]. Of these methods, the second generation, which are widely used, can be categorized in two major groups: sequencing by synthesis (SBS) and sequencing by ligation (SBL). The SBL approaches which were used in the past are based on short labeled sequences with one or two known bases. These two bases were important for encoding the specific unknown positions of a query sequence by hybridization and ligation. This is done in so many cycles until the whole sequence is encoded. While for one sequence only every 5th position is sequenced, the procedure is repeated in total with 5 shifted primers to cover all positions [159]. The SBS approaches are based on DNA-polymerase which synthesize the original sequence regarding the complementary sequence (see Figure 1.6). Each position is identified by incorporating special labeled nucleotides [161].

Basic concept

Coming to the general concept of sequencing, the following are mostly required for all methods. In the beginning before NGS analysis (step 1), the extracted DNA or RNA material is fragmented or sized to a desired sequence length [162]. In case of small RNA sequencing, the extracted sequences are already short enough, and this step can be skipped. During library preparation (step 2), complementary DNAs (cDNAs) are synthesized from the size-filtered target sequences, and adapters which have different functions are added to the flanks of the biological inserts. These different adapters are required for amplification, definition of the sequencing start and end position, and sample indexing to enable mixing multiple samples in one pool (multiplexing) [163]. Next, the clonal amplification of the library is applied so that the signal from the sequencer is strong enough to be detected [164]. Finally, the actual sequencing step based on ligation or on synthesis is executed [163].

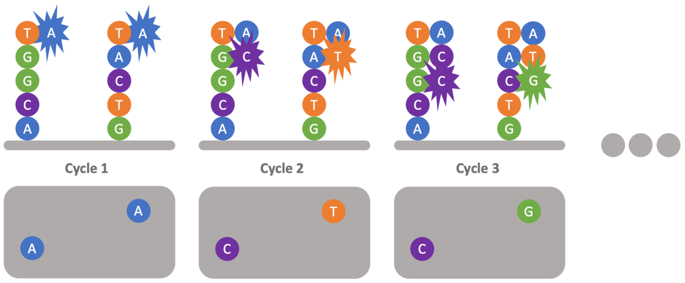


Figure 1.6: **Sequencing by synthesis.** In each cycle one special labeled complementary nucleotide is incorporated to the respective template to synthesize and reveals the sequence of the target.

Application and limitations

The high-throughput analysis by NGS allows quantification of known miRs and opens opportunities to discover novel biomarkers for diagnostics and prognostics [165]. In addition, it enables big steps towards personalized medicine, where detection of variants in genes became a topic of considerable interest in the last decade [166]. It not only allows the quantification of entirely new miRs in gene expression but also the discovery of new ones [167]. Regarding *de novo* sequencing of unknown genomes, NGS is fast and a cheaper alternative than the first generation of sequencing (Sanger method) [168]. However, one has to be aware that it has higher error rates towards the end of longer reads [169] and is still more expensive than other platforms such as microarray scanners or RT-qPCR techniques [159].

1.4.1.2 Nanostring

Basic Principle

The Nanostring nCounter is a hybridization-based system which can measures 800 miRNAs from 12 samples in one assay [170]. Referring to the manufacturer's description [171], the technology enables highly multiplexed single molecule counting in three steps. In the first step, a capture and a reporter probe containing a fluorescent barcode hybridize to the target molecule [170]. Next, the purification of samples is carried out by removing non-hybridized targets. The remaining purified target-probe complexes are bound and immobilized on the nCounter cartridge

[172]. In the last step, the barcodes of each reporter probe with a target sequence is counted digitally [173].

Application and limitations

According to the manufacturers, the Nanostring platform can be used for translational clinical studies, diagnostic fingerprinting and validation of high-throughput gene expression experiments [174]. Due to the fact that only 800 known miRNAs in 12 samples can be analysed on one cartridge [170], the technology is not ideal for biomarker discovery and can be very expensive for projects with large sample size.

1.4.1.3 *Microarray*

Different microarray methods

Since the millennium, DNA microarray devices has become a commonly used technology among of many researches with the result of over 150,000 related articles in PubMed search (keyword search in abstracts). In this time, three major type of arrays evolved for hybridization-based measurement: *in situ* synthesized arrays, spotted arrays on glass and self assembled arrays [175]. While the oligonucleotide probes of the spotted arrays are printed on glass, the probes of the *in situ* synthesized version are synthesized directly on the surface of the microarray chip [176, 177]. Microarrays produced by Affymetrix and Agilent were widely used, whereby the latter one is the main platform for analysis in this thesis [175]. However, the basic concept of this technology is the same. The probes immobilized in form of multiple copies represent known sequences, have defined positions on the array and hybridize as complementary parts to their respective target sequences which are fluorescently labeled (see Figure 1.7). With a hybridization, the fluorescent signal is increased according to the amount of the same targets [177, 178].

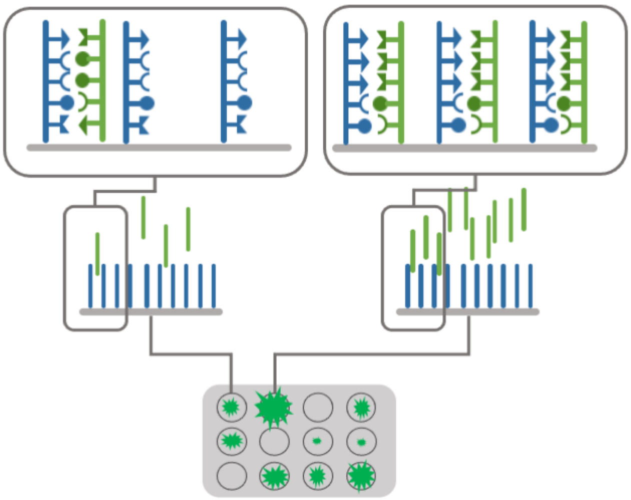


Figure 1.7: **Hybridization-based measurement.** Each position on a microarray chip defines different cluster of template copies (blue). A target sequence (green) hybridizes to its respective complementary template. The higher the amount of the target sequence is, the higher is the fluorescent signal.

Application and limitations

The main application of microarrays are general gene expression analysis [176] but they are also used for transcription factor binding analysis and single nucleotided polymorphism (SNP) genotyping [179]. In contrary to NGS, this technology can only detect known designed sequences. Microarray specific limitations are spatial effects and regional bias with artificially higher or lower intensities [180].

1.4.1.4 Real-Time qPCR

Basic principle of the commonly used PCR technology

In 1984 Kary Mullis and his co-workers invented the method known as polymerase chain reaction (PCR) which became a routinely used method in many laboratories to detect genes of interest. Until the year 2000, 617 articles were related to the keywords "real-time PCR", since then we can observe a substantial increase in the new century. We can find in the PubMed search until the year 2010 a total number of around 50,000 publications and until now a number of over 200,000. While the old PCR methods based on the question of whether a certain sequence is amplified in a prior defined number of cycles (endpoint PCR) [181], real-time quantitative PCR (qPCR) estimates the expression level of the target in real time against an arbitrary threshold (Ct) [182]. It describes how many PCR cycles are needed so that the fluorescent signal passes the threshold [182]. Thereby, it stands in an anti-proportional relation to the amount of DNA present in a sample. That means, the lower the Ct value, the higher is the amount of the target that leads to an earlier cycle threshold breakthrough [183]. On a closer view, an amplification cycle consists of three steps (see Figure 1.8). First, the DNA template with the target sequence is denatured. After that, it follows the annealing of primers to the single-stranded templates. With this, a polymerase creates a new complementary DNA strand to the template one. This extension is the last step of a cycle so that a new cycle can be started [184]. Before miRNAs can be measured by qPCR, they must first be converted into cDNA which is the first step of RT-qPCR.

Application and limitations

During the COVID-19 pandemic, the PCR technology achieved a great popularity among the world population for the Corona PCR test as the gold standard for diagnosing a coronavirus SARS-CoV-2 infection. Besides this and other pathogen diagnostics, it allows the research and identification of an accumulation of carcinogenic inherited single-nucleotide polymorphisms (SNPs) [183]. Because of the simple and solid detection of single genes, researchers use RT-qPCR to validate top candidate biomarkers derived from other platforms with large feature sets [158]. However, the gene expression analysis of cohort studies can be challenging independently of the choice of the target feature set. Due to variations caused from biological donor individuality, sample quality or experimental preparation, RT-qPCR results are normalized with a

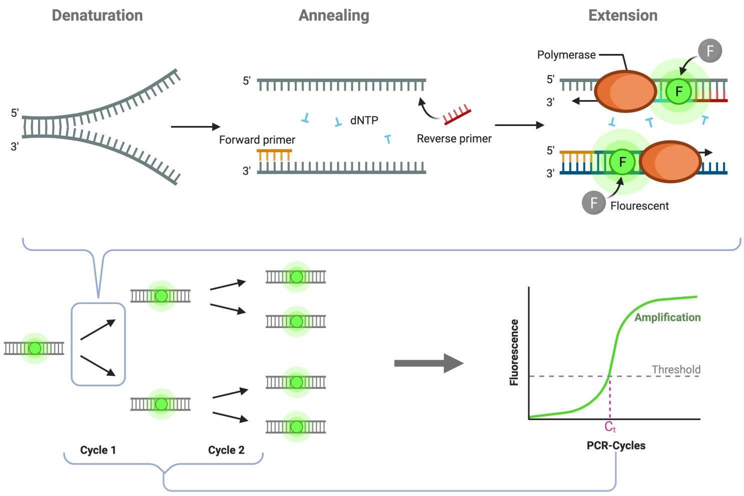


Figure 1.8: **Principle of qPCR.** With each cycle, the amount of the PCR product is growing exponentially. The integrated fluorescent molecules are monitored after each cycle to determine the amount of amplicons. When the signal intensity reaches a defined threshold, then the cycle number is noted. In each cycle, the same three steps are applied: 1. Denaturation of the double-stranded template, 2. Annealing of primers and nucleotides and 3. Extension of the respective complementary strands by polymerase, resulting in a fluorescent signal. Created with BioRender.com

selected control gene to correct for sample-to-sample variation. Ideally, this so-called housekeeping gene has a stable expression despite where the deviation is coming from and needs to be verified for its expression stability before each experiment [182, 185]. In addition to that, some researches observed that Ct levels can be determined inaccurately with known evaluation methods for exponential or sigmoidal curve fitting [186]. Nevertheless, the PCR technology remains as method of choice in the clinical use for the measurement of a few markers in single samples because of its cost-efficiency and widespread usage. Due to this logistic advantages, the PCR technology is suitable as a component of potential multi-marker diagnostic tests for complex diseases such as cancers.

1.4.2 *Blood collection devices*

Biopsy is an invasive and expensive method for sample collection and biomarker source. The costs vary highly regarding type of biopsy and setup. Average costs for a percutaneous biopsy can be around $1000 and surgical biopsies cost up to $30,000 [187]. While this complex method with high costs can limit the number of samples for a design of a research study, body fluids can be low-cost and minimally-invasive alternatives with simplified collection to achieve high numbers of samples more readily. In addition, higher sample sizes can increase the statistical reliability of analyses [188]. An important advantage of whole-blood specimens is that its composition includes PBMCs, central players of the immune responses that we want to measure. In the following, whole-blood-based collection devices such as PAXgene, dried blot spots and Mitra devices are presented.

1.4.2.1 *PAXgene - Sampling in clinical routine*

The PAXgene tube is one of the most widely used available whole blood RNA collection devices in clinical studies [189]. The blood draw is conducted through a needle pricked in a vein of an arm. After a small amount being discarded blood is collected, the stabilized stream of blood is collected by a vacuum pressure through a tubing connection into the PAXgene device. The PAXgene tube is then inverted 10 times to ensure optimal mixing of the lysis reagent with the blood. In detail, the cationic surfactant Catrimox-14 inside the tube is creating pores in cellular membranes and at the same time denaturating proteins. This cause lysis of blood cells and stabilization of RNA (RNases are unfolded) 1.9. This, together with the acidic environment of the tube (tartaric acid) prevents the degradation and ensures the prolonged stability of nucleic acids. The active component of the PAXgene tube is tetradecyltrimethylammoniumoxalate (a.k.a. Catrimox-14) which has lipid-like properties and interferes with the cell membrane [190]. The negatively charged nucleic acids are released and embedded by reverse micelles based on positively charged Catrimox-14 molecules. The longer the nucleic acid is, the more effective the formation of the micelles, thus the nuclei and ribosomal RNAs pellet most efficiently. After the storage, which can be up to a decade at -80C, the tube is centrifuged. The resulting pellets are used for RNA extraction (see PAXgene Blood RNA Kit Handbook version 2).

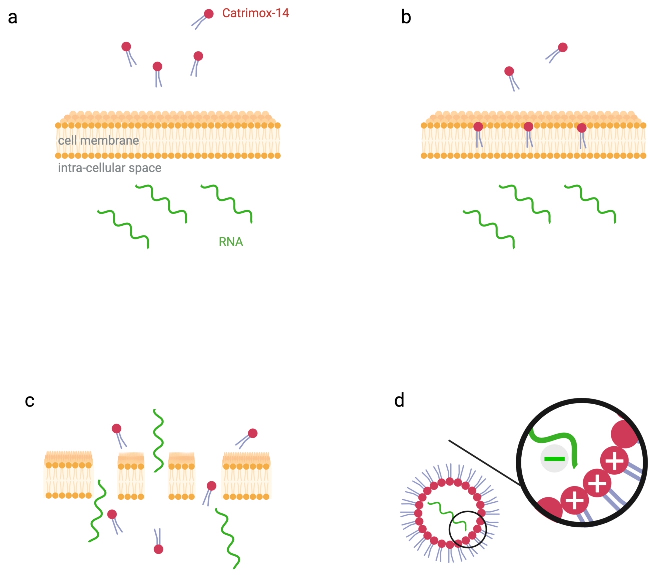


Figure 1.9: **Lysis and RNA stabilization.** (a) Catrimox-14 which is a component of PAXgene will lyse the cells. (b) The Catrimox-14 molecules are incorporated to the cell membrane to make it permeable. (c) Intracellular RNAs can pass the permeable membrane and will be mixed with RNAs of other cells. (d) In addition, Catrimox-14 molecules stabilize the release RNAs by building reverse micelles due to complementary electric charge. Created with BioRender.com

1.4.2.2 *Dried blot spots (DBS) & Mitra - Options for self- and home-sampling*

The idea of sampling blood with filter paper reaches back to the beginning of the 20th century. In 1913, Ivar Christian Bang, considered as founder of the modern clinical microanalysis, could successfully identify glucose from blood collected on dried blot spots (DBS) [191]. The advantages of DBS were already pointed out by Chapman in the 1920s. It requires less blood volume and the collection is simple, minimally invasive and cheap [192]. In the 1960s, using these benefits, Guthrie developed a screening test, which is still used today, for various diseases of newborns such as phenylketonuria (impairment of brain development) and galacatosemia (impairment of galactose metabolization) [193, 194]. The collection of DBS samples is carried out with a finger prick by a lancet, as for example in glucose measurement. The resulting blood drops fall on the specially manufactured cellulose filter paper. After the complete drying, the DBS sample is ready to be stored or shipped [195].

One further collection device which needs less blood volume is Mitra, a recent technology based on volumetric absorptive microsampling (VAMS) [196]. It seems to be a more feasible and reliable alternative novel dried sampling approach than DBS, especially with the result of a constant blood volume. Here, we prick again to obtain a blood drop, which is absorbed by a white cellulose tip attached to a plastic stick. It is fabricated in that way that it always ensures a fixed blood volume [197]. The complete drying occurs in a cartridge or clamshells specially made for this [196].

In conclusion, due to their simplified blood collection compared to tube-bases solutions, low blood volume and direct drying and thus stabilisation of the RNA, these two blood dried microsampling methods are becoming the focus of research for home sampling [198], especially in corona times with limited mobility [199]. In case of pandemic situations with social distancing and lockdowns, simple microsampling approaches could ensure the continuation of non-acute medical services such as health screening of older adults.

2
Goals of the PhD thesis

As described in the previous chapter, miRNAs, especially derived from body fluids, are the central focus of this thesis. The research goals were built on these blood-borne molecules and defined as follows: (1) the prediction of miRNA candidates, (2) validation of them and their understanding regarding technical, seasonal and cross-species-related aspects, and (3) detection of significant biomarker candidates in diseases (Figure 2.1).

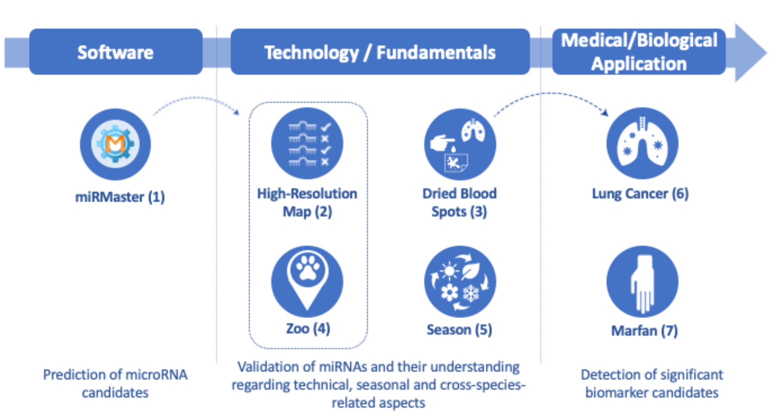

Software
Technology / Fundamentals
Medical/Biological Application
miRMaster (1)
High-Resolution Map (2)
Dried Blood Spots (3)
Lung Cancer (6)
Zoo (4)
Season (5)
Marfan (7)
Prediction of microRNA candidates
Validation of miRNAs and their understanding regarding technical, seasonal and cross-species-related aspects
Detection of significant biomarker candidates

Figure 2.1: **Overview of goals and studies of this thesis.** (1) Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs [3], (2) A high-resolution map of the human small non-coding transcriptome [4], (3) Technical Stability and Biological Variability in MicroRNAs from Dried Blood Spots [1], (4) The sncRNA Zoo: a repository for circulating small noncoding RNAs in animals [5], (5) Spring is in the air: seasonal profiles indicate vernal change of miRNA activity [6], (6) Evaluating the Use of Circulating MicroRNA Profiles for Lung Cancer Detection in Symptomatic Patients [2], (7) Deregulated microRNA and mRNA expression profiles in the peripheral blood of patients with Marfan syndrome [7].

With increased usage of small non-coding RNA-sequencing, the development of bioinformatics tools has gained traction and has produced several tools for miRNA identification and prediction with different approaches. We also developed a software, called miRMaster [3], addressing the following design criteria: user-friendliness, high computational performance, machine learning and solid derivation of miRNA candidates. Thereby, we implemented a web service with different parameter options on a central server with high computational power. To obtain reliable results, we focused on carefully selected training sets, a wide range of machine learning configurations and well-chosen biological characteristics of miRNAs. We used high confidence miRNAs based on early miRBase versions as our positive test set, and a broad combination of artificial psuedo precursors, protein coding sequences, false positive miRNAs, and ncRNA-overlapping pseudo precursors as our negative test set. Based on these datasets, we obtained the best classifier for miRMaster usage, testing 180 various combinations of classification, feature scaling and subset selection methods. On the other hand, to improve the prediction with biological characteristics, miRMaster uses sequence and structure features of miRNAs and evaluates the reconstructed miRNA biogenesis of candidates. While Fehlmann was responsible for the implementation of the backend of miRMaster, I focused on the frontend with the goal to develop a user-friendly graphical user interface. Whereas the integration of tooltips, tutorials and default parameters can be helpful and provide orientation for first-time users, advanced users can configure the expert settings. An upload wizard with drag-and-drop property can simplify the data upload. The last aspect for a user-friendly tool was the implementation of giving immediate feedbacks to the users. This is important to know in which state the requested analysis is.

The next goals were the validation of miRNA candidates and the development of fundamental understandings for miRNA stability and characteristics from different perspectives. In our work regarding a high-resolution map of human sncRNAs with miRNAs as the main class [4], we applied miRMaster on 24,554 samples from different data sources to obtain novel miRs, and performed a staged validation for a final set of over 11,000 sequences. In the first validation step, these sequences were used for a custom microarray analysis. In the final step, selected candidates of significantly deregulated markers in lung carcinoma were confirmed by northern blots. A further important sub-goal was to understand the stability of miRNAs by analysing the following three different aspects. First, we addressed the question of whether different blood-based collection devices can affect the diagnostic outcomes. For that, we tested dried blood spots from lung cancer patients for technical stability and biological variability [1]. Second, we investigated the seasonality of physiological state effects [6]. In the same project, we tested Mitra as an additional blood-based collection device. Third, we applied miRMaster on sequenced Mitra samples from animals to understand cross-species relations on small non-coding RNA level and found miRNA candidates due to their evolutionary conservation [5].

The results of our DBS study indicated promising results for blood-borne miRNA biomarkers regarding technical stability and biological significance between lung cancer and control patients. For this reason, our third goal was to identify significant biomarker candidates for diseases. The main work was here a multi-center and multi-cohort study for detection of lung cancer with over 3,000 samples [2]. This large-scale project included different analysis scenarios comparing lung cancer patients against non-tumor lung disease and other control patients. We also performed analyses where the groups were matched by age, gender and smoking behavior to control potential confounding effect. In addition, we performed analysis for early detection by choosing only lung cancer stage I and II for the case group. In a different study, we researched the potential of miRNAs as diagnostic and prognostic markers for Marfan syndrome [7]. We carried out integrated miRNA and mRNA expression analyses of samples collected from Marfan patients and healthy volunteer controls. Aside from the publications that are listed here for my thesis, I worked on a breast cancer study [200] identifying potential diagnostic biomarkers and analyzing the prognostic potential of miRNAs for predicting the pathological complete response of neoadjuvant chemotherapy.

3
Results

This cumulative thesis is based on seven peer-reviewed publications whose published versions are included in this chapter.

Published online 12 July 2017

Nucleic Acids Research, 2017, Vol. 45, No. 15 ***8731–8744***
doi: 10.1093/nar/gkx595

Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs

Tobias Fehlmann[1,*], Christina Backes[1], Mustafa Kahraman[1,2], Jan Haas[3,4,5], Nicole Ludwig[6], Andreas E. Posch[7], Maximilian L. Würstle[8], Matthias Hübenthal[9], Andre Franke[9], Benjamin Meder[3,4,5], Eckart Meese[6] and Andreas Keller[1]

[1]Chair for Clinical Bioinformatics, Saarland University, Saarbrücken, Germany, [2]Hummingbird Diagnostics GmbH, Heidelberg, Germany, [3]Department of Internal Medicine III, University Hospital Heidelberg, Heidelberg, Germany, [4]German Center for Cardiovascular Research (DZHK), Heidelberg, Germany, [5]Klaus Tschira Institute for Integrative Computational Cardiology, Heidelberg, Germany, [6]Department of Human Genetics, Saarland University, Homburg, Germany, [7]Ares Genetics GmbH, Vienna, Austria, [8]Siemens Healthcare GmbH, Strategy and Innovation, Erlangen, Germany and [9]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany

Received April 25, 2017; Revised June 23, 2017; Editorial Decision June 26, 2017; Accepted July 04, 2017

ABSTRACT

The analysis of small RNA NGS data together with the discovery of new small RNAs is among the foremost challenges in life science. For the analysis of raw high-throughput sequencing data we implemented the fast, accurate and comprehensive web-based tool miRMaster. Our toolbox provides a wide range of modules for quantification of miRNAs and other non-coding RNAs, discovering new miRNAs, isomiRs, mutations, exogenous RNAs and motifs. Use-cases comprising hundreds of samples are processed in less than 5 h with an accuracy of 99.4%. An integrative analysis of small RNAs from 1836 data sets (20 billion reads) indicated that context-specific miRNAs (e.g. miRNAs present only in one or few different tissues / cell types) still remain to be discovered while broadly expressed miRNAs appear to be largely known. In total, our analysis of known and novel miRNAs indicated nearly 22 000 candidates of precursors with one or two mature forms. Based on these, we designed a custom microarray comprising 11 872 potential mature miRNAs to assess the quality of our prediction. MiRMaster is a convenient-to-use tool for the comprehensive and fast analysis of miRNA NGS data. In addition, our predicted miRNA candidates provided as custom array will allow researchers to perform in depth validation of candidates interesting to them.

INTRODUCTION

MicroRNAs (miRNAs) play a central role in orchestrating human gene regulation and are consequently prime targets in biomedical research. Many miRNAs from *Homo sapiens* and other species are collected in the miRBase (1). Currently, the fraction of actually true positive miRNAs in this database is controversially discussed (2–10), especially later versions seem to contain many false positives (11). On the one hand, this calls for curated databases, on the other hand not all miRNAs, especially context specific ones, seem to be discovered yet.

Various experimental approaches are applied for measuring miRNA expression levels including approaches for small sets of selected miRNAs like RT-qPCR, CMOS based assays (12) or immunoassays (13). The most frequently employed genome-wide assays include microarray screening and high-throughput sequencing (HT-seq). A comparison of 12 different experimental approaches is provided by Mestdagh *et al.* (14).

HT-seq enables—beyond quantitative analysis of known miRNAs—single-base resolution of known and novel miRNAs (15) and thus is currently applied to discover the afore mentioned context-specific miRNAs. For the analysis of HT-seq data, a wide range of stand-alone and web-based bioinformatics tools have been implemented allowing the prediction of novel miRNA candidates and quantification of miRNAs (16,17), detection of miRNA isoforms (18,19), miRNA set enrichment analyses (20,21), and prediction of miRNA targets (22,23) among others. Akthar *et al.* published a comprehensive review on 129 available miRNA bioinformatics tools (24). The different data formats used in these tools and the challenges to combine web-based and stand-alone solutions, however, complicate the design of integrated pipelines.

---

[*] To whom correspondence should be addressed. Tel: +49 681 30268603; Email: tobias.fehlmann@ccb.uni-saarland.de

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Our ambition was to develop a web-based application that combines the most frequently requested analyses. An important aspect of our tool termed miRMaster (www.ccb.uni-saarland.de/mirmaster) was to facilitate HT-seq data analysis of human samples from raw sequencing files provided in the FASTQ format. Building up on the basic principle of miRDeep2 (16) as the most frequently used prediction tool for miRNAs, we implemented an own predictor with an extended feature set including our previously developed prediction score (11). Furthermore, we implemented functionality to report the presence of miRNA motifs to the user (25–27). MiRMaster allows to search for novel miRNA candidates, to quantify miRNA expression, to identify isoforms and variants of miRNAs. Another feature of miRMaster is the mapping of non-human small RNA reads against the NCBI RefSeq collection of bacterial and viral genomes (28), thereby allowing the detection of contaminations, infections or exogenous miRNAs. To allow the analysis of targets regulated by miRNAs, we implemented Application Programming Interfaces (APIs) to available web-based tools for considering the targetome (miRTargetLink (29)) and to carry out miRNA set enrichment (miEAA (20)).

Since different research groups measured various specimens using different experimental protocols and bioinformatics pipelines and not all data stored in a central repository, a redundancy between the studies exist. Besides the miRNAs in the miRBase, and specific studies mentioned before, several comprehensive analyses (e.g. Londin *et al.* (30), Backes *et al.* (11), Friedländer *et al.* (31), Jha *et al.* (32)) propose hundreds to thousands of new miRNAs. To detect as many as possible miRNA candidates we performed a comprehensive analysis of 1836 data sets containing 20 billion reads.

MATERIALS AND METHODS

Sample collection

As case study we analyzed an in-house NGS miRNA sample collection of 1097 samples from blood and blood cell components (33–39). Further we downloaded 739 samples from four series of the GEO database (40): GSE64142, GSE53080, GSE49279 and GSE45159. All samples have been sequenced using Illumina Next-Generation sequencing. Table 2 presents an overview of these samples including a description, number of samples, number of reads and file size.

Positive miRNA dataset for training miRMaster

A straightforward positive dataset would consist of the complete miRBase (1). However, others and we have observed that miRBase may contain false positives, especially in the last versions (41). Therefore, we selected all miRNA precursors from miRBase 1 to 7 and all precursors of miRNAs containing strong experimental evidence in the miRTarBase (42), leading to 487 high-confidence positive miRNAs. We defined precursors by their 5′ and 3′ mature miRNAs, i.e. they start with the first base of the 5′ miRNA and end with the last base of the 3′ miRNA. For miRBase precursors that had only one form annotated we derived the other from its hairpin, as described for our prediction algorithm. Therefore, our predictions are independent of the size of the stem loops provided in miRBase.

Negative miRNA dataset for training miRMaster

Choosing an appropriate negative dataset is a challenging task, since miRNAs can be located anywhere in the genome (43). A correct negative dataset plays an important role for the creation of a well-trained classifier. Overall, since only a small fraction of the genome and of sequences that form hairpins are actually precursors, we built five different sets to cover as many potential wrong predictions as possible. The different negative datasets were derived from separate assumptions and combined for our training procedure. The first dataset was built to cover predictions, where one actual miRNA is contained in the predicted precursor but the other miRNA is wrongly annotated. We assume that real precursors do not overlap. It was created by splitting in half all known stem–loops from miRBase that contained two annotated mature miRNAs. We adjusted the length to the original stem-loop by including the flanking regions. To determine the positions of the miRNAs in the two new pseudo precursors, we kept the original miRNAs and derived the other based on it, as in our prediction algorithm. This dataset was composed of 298 precursors. The second dataset was created to cover predictions that could stem from protein coding sequences of genes without known alternative splicing events. It was derived from the widely used pseudo precursor set built by Xue *et al.* (44). We first kept only sequences that aligned perfectly to the latest assembly of the human genome (hg38). Then we segmented these sequences to enable the computation of segment specific features. Therefore, we determined the position of one of the pseudo miRNAs by assigning it to the segment with most base pairs, having a length of 20 nucleotides and non-overlapping with the loop region. The other was derived from it, as in our predicting algorithm. The resulting set contained 3916 pseudo precursors. The third dataset was created to cover predictions that could arise from stem-loops of other ncRNAs. It was shown by others (45) that for a very small portion of all known miRNAs this could actually be the case. However, due to their low number and the false positives largely outweighing the true positives we considered this set to be useful to reduce the false positive prediction rate. The dataset was derived from Rfam (46) (release 11) and composed of 3342 negative precursors. We considered all human ncRNAs that were not miRNAs and derived pseudo precursors by retaining only those that could be partitioned into 5′, 3′ and loop parts. The fourth dataset was created to account specifically for predictions that would pass the filtering steps in our algorithm, but which would overlap with other ncRNAs. It is in fact an extension of the third dataset. We derived 4031 pseudo precursors by running our prediction on 705 in-house samples and keeping only those that passed all filtering steps but overlapped with other ncRNAs of Rfam. The fifth dataset was created to account for predictions that were not covered by the other negative datasets. It was derived from early predictions performed by our algorithm (trained on the other four datasets) on our in-house samples. This set addresses specifically predictions where the miRNAs contained many repeated bases and further, miRNA duplexes with high normalized free energy and precursors with high normalized free energy. We kept all predictions that displayed evidence for being false positives, i.e. precursors with miRNAs containing at least seven consecutive A or U or 8 C or G. Further we kept all with a normalized ensemble free energy of over –0.15 kcal/mol*nt or with a normalized duplex minimum free energy of over –0.15 kcal/mol*nt. The cutoffs were determined empirically by analyzing the distribution of the properties of known precursors. This led to 797 additional negative miRNAs. For the first four datasets we further retained only those pseudo precursors without bifurcations, with at least 50% paired bases between the 5′ and 3′ pseudo miRNAs and with a 5′-3′ miRNA length difference of at most 10. The combination of all negative datasets resulted in 12 384 pseudo precursors, which are listed in Supplementary Table S2.

Independent test sets for evaluating miRMaster

To validate the performance of our model we created two additional independent test sets. The first set was composed of human precursors of MirGeneDB (10) that were not used in our training process, resulting in 129 precursors. For the pseudo precursors we selected all sequences that were annotated as human precursors in earlier miRBase versions (1–20) and that were not duplicates or merged with known precursors. This resulted in 28 sequences, of which 6 were discarded by our algorithm when trying to determine a valid corresponding second miRNA arm. In addition, we created a second set composed of mouse precursors of MirGeneDB that had different sequences than our training precursors, resulting in 350 precursors. We selected the negative set analogously to the first negative set from early annotated mouse precursors, leading to 65 sequences. We mapped those sequences against the mouse genome (mm10) and removed all sequences which were not found or found at multiple positions. Of the remaining 56 sequences, 11 were discarded by our algorithm when trying to determine a valid second miRNA.

Features of miRMaster for predicting novel miRNAs

We created a feature set composed of 216 properties, based on 186 existing features described in (44,47–51) and 30 novel features. Novel features included our previously developed novoMiRank score (11), open/close parentheses and unpaired nucleotides in all thirds of a precursor, 5′-3′ miRNA duplex minimum free energy, the number of base pairs in the 5′ and 3′ miRNAs and in-between, and the nucleotide ratio of the 5′ and 3′ miRNAs. Supplementary Table S1 lists all features including a brief description, their runtime impact and the *P*-value resulting from a two sided Wilcoxon rank-sum test after Benjamini–Hochberg adjustment for multiple testing (52) (alpha = 0.05) on our positive and negative datasets.

Classifier selection for predicting miRNAs

To obtain the best classifier for our positive and negative dataset in terms of specificity and sensitivity we evaluated 180 different combinations of feature scaling, subset selection and classification methods using the scikit-learn Python toolkit (53), as shown in Supplementary Table S9. Since a large fraction of features can be computed in minimal time while very few features take very much computing time we built two models: one is based on all features and one based on the features with low runtime. For each combination we tuned the classifier's hyper-parameter via particle swarm optimization towards maximum ROC AUC, resulting in a total of 130,105 models. From those we then selected all models that performed at least as good as the best 25% according to ROC AUC, Precision-Recall AUC, sensitivity, specificity and Matthews correlation coefficient (MCC). The final model was chosen according to the highest $F_{0.5}$ measure. Supplementary Figure S15 sketches this process.

Input data of users to miRMaster

Since our ambition was to facilitate comprehensive miRNA analysis for all researchers, we implemented upload functionality for FASTQ files that are processed and compressed in the browser before being sent to the server. Thus, no additional software installation that compresses the files on the user's computer is needed. This feature is supported by only few tools, such as MAGI (54). Further we provide support for gzip compressed FASTQ files, since they are the typical storage format of sequencing files, thereby obviating the need to decompress files before inputting them to miRMaster.

Preprocessing

Before sending the input files to our server we perform three preprocessing steps consisting of adapter trimming, quality filtering and read collapsing. Adapter trimming is performed via fuzzy string matching and can be customized by the user. We allow one mismatch and require an overlap of at least 10 nucleotides with the read per default. Further the user has the possibility to trim leading and trailing *N*, discard reads containing any remaining *N* and remove reads shorter than a specific size. For the quality filtering step, we re-implemented the sliding window filtering approach used by Trimmomatic (55). This allows reducing the amount of data sent by up to 99.9% (depending on the sample specimens). To take advantage of multi-core processor capabilities we use JavaScript web workers to allow the preprocessing of multiple files at the same time.

Mapping to various ncRNA databases

We map the collapsed reads using Bowtie (56) and allow per default no mismatches against human rRNAs, snRNAs, snoRNAs, scaRNAs and lincRNAs of the Ensembl noncoding RNA database (release 85) (57), against piRNAs of piRBase (1.0) (58) and tRNAs of GtRNAdb (59). This allows the user to easily verify if the distribution of reads is as expected or to investigate specific RNAs. To allow the user to investigate specific ncRNAs we provide detailed expression counts for all ncRNAs we are mapping against, as well. The expression is determined by the number of reads mapping to a specific sequence using Bowtie. Further we report the mapping of reads against the human miRBase (version 21), which can be used to estimate the potential of finding novel miRNAs in the samples.

Mapping to reference

Mapping the collapsed reads to the reference genome is performed using Bowtie. Analogous to miRDeep2 (16), we require no mismatches in the first 18 nucleotides and discard reads that map to over five different locations.

Precursor excision, segment determination and filtering

The precursor excision, segment determination and filtering according to their structure and signature is performed analogous to miRDeep2. Briefly, local maximum read stacks in downstream windows of 70 nucleotides are searched and two precursors excised from each stack. The secondary structure is computed for each precursor using RNAfold (60). The maximum read stack represents one miRNA of the precursor. The other miRNA is determined by the paired sequence on the other arm with a 2-nucleotide overhang. Filtering steps are composed of a structure and signature filter. The secondary structure is required to have no bifurcations, a minimum percentage of base pairs in the highest expressed miRNA of 60% and a length difference of both miRNAs of at most five nucleotides. The signature is checked by mapping all reads with at most one mismatch against all excised precursors. At least 90% of all reads need to map to either a miRNA or in between, thereby discarding reads that do not map according to Dicer processing. All these thresholds can be customized in the web interface.

Feature computation and prediction

After the potential precursors have been excised and filtered we compute their feature values and perform the prediction using our classifier as described in previous parts of the Materials and Methods section.

Prediction merging and global signature filtering

Once the predictions for all samples have been performed we merge the resulting potential precursors in order to avoid multiple predictions shifted by only a few bases. Therefore, we group all precursors that differ by at most 10 positions and keep the one that was found in most samples. To make use of additional information provided by multiple samples we first normalize the expression of each read of each sample to reads per million (RPM) and sum up identical reads. Then we map the normalized reads of all samples against the merged predictions and score their signature. We weight each read using the following formula

$$
\begin{aligned}
&score(read)\\
&= total\_RPM(read) \cdot length(read) \cdot \sqrt{\frac{occuring\_samples(read)}{\#total\_samples}}
\end{aligned}
$$

Thereby, we penalize reads that occur in only few samples while giving more weight to longer reads. Reads mapping with mismatches are penalized per default by a dividing factor if they occur in at most 10% of all samples (but at most 10 samples). The dividing factor is the limit of occurring samples minus 1, but at least 2. We then remove all predictions that have a signature with an inconsistent dicer processing read portion representing at most 20% of the total score.

Categories of new miRNAs

We assign to each predicted precursor one of six categories. (1) *Known*: when the prediction is overlapping with a miRBase entry and both miRNAs are overlapping with known miRNAs by at least 75%. (2) *Shifted known*: when the prediction is only partially overlapping with miRBase and only one miRNA is overlapping by at least 75% with a known miRNA. (3) *One annotated*: when the prediction is overlapping with a miRBase entry, but only one miRNA is annotated for that entry and this one is overlapping by at least 75%. (4) *Dissimilar overlapping*: when the prediction is overlapping with a miRBase entry, but the miRNAs are not overlapping with the annotated ones. (5) *Half novel*: when the prediction is not overlapping with any miRBase entry, but contains at least 75% of one known miRNA. (6) *Novel*: when the prediction is not overlapping with any miRBase entry and does not contain any known miRNA.

Prediction flagging of other ncRNAs

In order to reduce the number of potential false positives, we map the predicted precursors to the Ensembl human non-coding RNA database (release 85) and to NONCODE 2016 (61) using BLAST+ (62) and flag them accordingly when matches are found. Further we map against the whole miRBase (v21) to highlight similar miRNAs in other species. Mappings are valid when over 90% of the aligned sequences overlap and at most one mismatch is present.

Quantification of known and novel miRNAs, isomiRs and mutations

The quantification of known and novel miRNAs is performed analogously to miRDeep2. Reads are mapped against the precursors using Bowtie while allowing one mismatch. The counts are reported for all reads overlapping the annotated miRNAs in a window of up to two nucleotides upstream and five nucleotides downstream. IsomiRs are detected by mapping against the precursors using Bowtie while tolerating two mismatches. We allow up to two non-template additions to the 5′ and 3′ ends and up to one mismatch in between. We also allow a variability of two nucleotides at the 5′ end and of five nucleotides at the 3′ end per default. When detecting mutations, we focus on single nucleotide substitutions. The mapping and counting is performed the same way as the quantification, however miRNAs with mutations are explicitly counted.

Exogenous read mapping

We map non-human reads (all reads that did not align to the human genome with at most one mismatch) to all 7556 bacteria and 7026 virus sequences of NCBI RefSeq (28) release 74 and report the number of perfectly mapping reads.

Reads mapping to bacteria or viruses can indicate exogenous miRNAs, but also reagent contamination or diseases such as sepsis.

Motif detection

Recently five miRNA motifs have been reported, namely the UG, UGU/GUG, CNNC (25), GHG (26) and GGAC (27) motif. We report for each prediction the present motifs, allowing matching up to two nucleotides upstream or downstream of the expected motif position.

Usability

To analyze NGS miRNA samples with miRMaster, the user needs to provide sequencing files in FASTQ format (uncompressed or gzip compressed) without barcode sequence and the 3′ adapter used in the library preparation. After clicking on the 'Launch experiment' button on the homepage or in the navigation bar, the user will be guided through three steps. During the first one, one should name the experiment and also optionally provide an e-mail address to receive a notification as soon as the analysis of the uploaded samples is done. During the second step the user needs to specify the used 3′ adapter and has the opportunity to fine-tune the parameters of the analysis. The third step consists of the upload of the sequencing files. If the samples stem from multiple cohorts, groups can be specified by either clicking on the 'Add second group' button or by uploading a tab separated sample-to-group file. Once the files are chosen and the user has clicked the 'Launch' button, the data will be preprocessed and sent to the server. The preprocessing progress is shown directly on the web page whereas the server progress can be followed in real time by clicking the 'Follow' button. This will open the experiment status page in a new tab, where the user will be able to track the progress of the analysis of all uploaded samples. Real-time web reports are provided for each sample that has been uploaded, allowing to directly inspect the data. These reports provide information on the preprocessing, mapping, quantification and prediction steps. As soon as all samples have been analyzed, the results can be downloaded and an overall web-report is created with a link to it on the top of the status page.

Validation using custom microarray

To perform a first pass iteration and to minimize the risk of false positives due to either NGS artifacts or low sample quality containing many degraded RNAs we designed a custom microarray containing all human miRNAs from the miRBase, the miRNAs from the study by Londin *et al.* (30) as well as over 5000 miRNAs from the present study. Among our predicted miRNAs we selected only those expressed in at least 50 samples which were not flagged as similar to other ncRNAs. The final microarray contained 11 866 miRNA candidates that have been measured each in 20 replicates (237 320 features per sample).

In order to measure the expression of the novel miRNAs in different human cells and tissues, we compiled a set of eight different human RNA samples: we purchased human total RNA samples from lung, brain, kidney, testis and heart tissues from Life Technologies (Cat. No. AM7968, AM7962, AM7976, AM7972 and AM7966, respectively) and the human miRNA reference kit from Agilent Technologies (Cat. No. 750700), that represents a pool of several human tissues and cell lines. Furthermore, we used a PAX blood RNA pool and a plasma RNA pool. The PAX blood RNA pool comprised of 11 blood samples collected in PAX gene tubes and purified with PAXgene Blood miRNA Kit from Qiagen according to manufacturer's instructions. Blood samples derived from four lung cancer patients, two Alzheimer's Disease patients, two patients with Wilms Tumor, and three healthy donors. The plasma RNA pool comprised of 10 plasma samples from healthy donors and was isolated using miRNeasy Serum/Plasma Kit after manufacturers recommendation with minor adaptations. To ensure sufficient RNA precipitation, we added 1 μl 20 mg/ml glycogen (Invitrogen) in the precipitation step. RNA concentration was measured using Nanodrop (ThermoFisher). RNA quality was assessed using Agilent Bioanalyzer Nano kit (for all tissue derived RNAs) or Small RNA kit (for the plasma sample).

The expression of 11 866 miRNAs and miRNA candidates was determined using the customized Agilent human miRNA microarrays. As input we used 100 ng total RNA as measured in Nanodrop for all tissue derived RNAs, and 1 ng miRNA as measured using Bioanalyzer Small RNA chip for the plasma sample. Using Agilent miRNA Complete Labeling and Hyb Kit after manufacturer's instructions, RNAs were dephosphorylated and labeled with Cy3-pCp. Labeled RNAs were hybridized to the custom microarrays for exactly 20 hours at 55°C. After hybridization, arrays were washed for 5 min in each Gene Expression Wash Buffer 1 (room temperature) and 2 (37°C). Subsequently, arrays were dried and scanned in an Agilent microarray scanner (G2505C). Expression data was extracted using Agilent feature extraction software. Downstream processing of signals has been carried out with R (version 3.2.4). Specifically, for clustering the expression intensities hierarchical clustering using the Euclidean distance has been performed as implemented in the Heatplus package.

To enable other researchers to repeat the experiments and to perform measurements on own samples, the microarrays that can be used analogously to standard Agilent microarrays using the Agilent protocols and SureScan platform, will be distributed by Hummingbird Diagnostics (Heidelberg, Germany) in three versions: human-mirna-candidate(full) containing all miRNA candidates from this study; mirna-candidate(detected) containing all miRNAs positive in any experiment of this study; mirna-candidate(blood) containing all miRNAs that have been detected in blood or serum.

RESULTS AND DISCUSSION

The aim in developing miRMaster (www.ccb.uni-saarland.de/mirmaster) was to implement a comprehensive tool for the analysis of miRNA NGS data sets. Starting from raw or compressed FASTQ files with billions of reads and gigabytes of data, miRMaster allows a wide variety of miRNA analyses. The complete workflow is described in detail in the Methods section and sketched in Figure 1. A brief de-

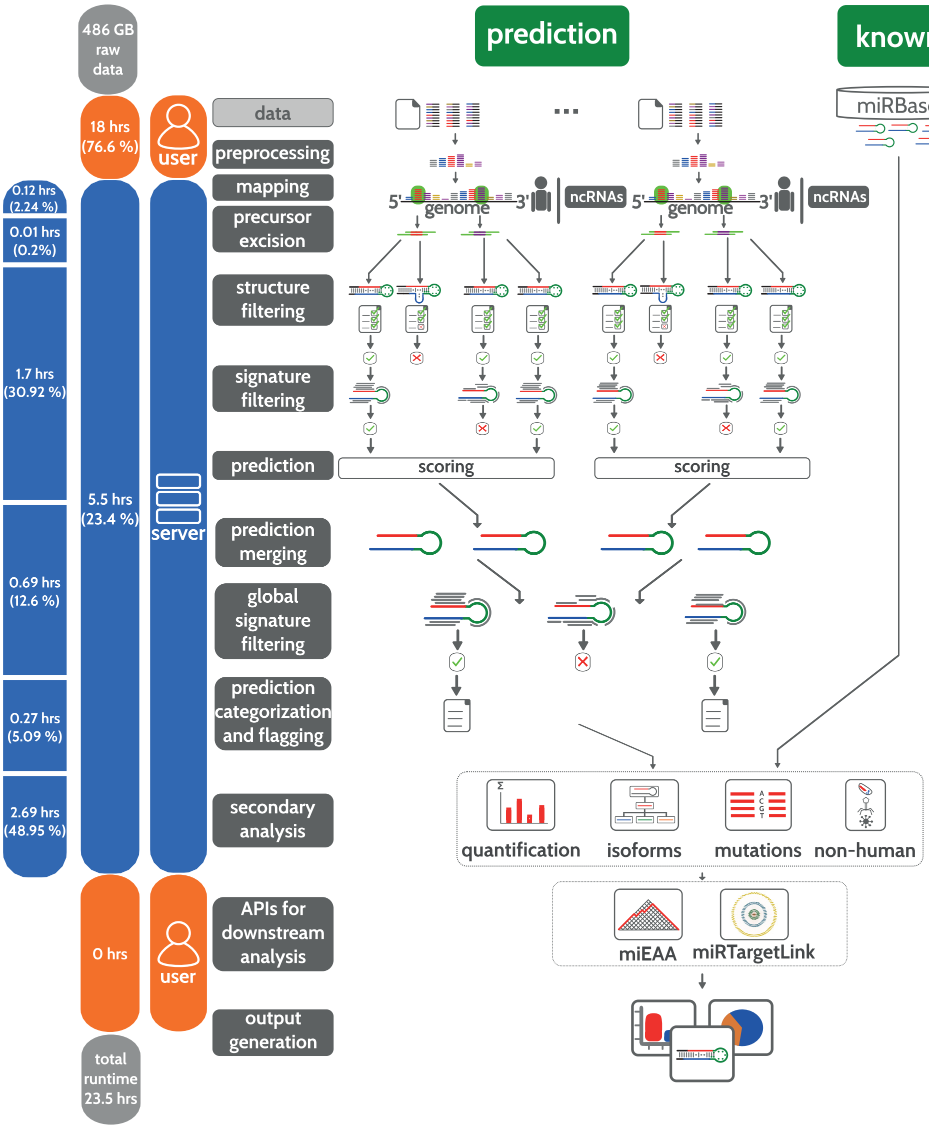


Figure 1. Schematic workflow of miRMaster. The bar at the left shows the runtime impact of each step. Steps performed by the user are shown in orange and steps performed by the server in blue.

scription on the usability of miRMaster is available in the Methods section.

In the following, we first focus on the performance of the novel algorithm for the prediction of new miRNAs. In total, we investigated 1097 miRNA NGS data sets containing 15 billion reads within a 486 GB file size and compare the miRMaster results – in terms of performance and runtime—to those of miRDeep2 using the same data sets. We next provide a detailed description of the different components of our miRNA NGS analysis framework and their application to the above-mentioned data set. Then we report a coarse description of the human miRNome by predicting small RNAs from 1836 data sets with 20 billion reads. Finally, we analyze the expression of potential miRNA candidates using custom microarrays.

Evaluation of miRNA features

In contrast to most other comparable tools, our miRNA prediction relies on a broad set of features that are derived both from precursor sequences and from their mature forms. These features are considered as weak learners as each feature has a limited impact on the overall decision to classify or declassify a new miRNA as true miRNA. The feature set consists of 216 single features including nucleotide composition, secondary structure and others (the full list is available in Supplementary Table S1). To gain first insight into the discrimination power of single features we derived a positive miRNA precursor set from early miRBase (63) versions and from targets with strong experimental evidence in miRTarBase (42) (487 precursors), as well as a negative miRNA precursor set from various sources (12 384 negative precursors). A detailed explanation on the creation of these sets can be found in the Methods section (the sequences and locations of both sets are shown in Supplementary Table S2). We calculated the significance of all features by comparing both sets via Wilcoxon rank-sum tests. The performance of the 216 features is listed in Supplementary Table S1. The smallest significance value (10^{-219}) was calculated for the minimum free energy index 1. Following adjustment for multiple testing, 158 of the 216 features remained significant ($P < 0.05$). Since our analysis pipeline is designed to support the evaluation of large data collections of up to several thousand samples, performance in runtime of feature calculation is of importance. We grouped all features in three different runtime categories with the fastest category containing features with 10,000-fold decreased runtime as compared to the slowest features. Supplementary Figure S1 shows the negative decadic logarithm of the *P*-values for features in the three categories. Since the two fast categories already contained 54 and 86 significant features, respectively, we evaluated their combined information content for predicting miRNAs. We derived classifiers not only from the complete feature set, but also from the fast features set only. Prior to classifying miRNAs based on the features we evaluated the redundancy of the features selected. As shown in the correlation heat map in Supplementary Figure S2 many of the features were redundant.

Classification of precursors

For combining the predictive power of the weak learners we applied different feature selection and classification approaches. We selected a large variety of classifier and feature selection approaches, since there is no ‘one size fits all’ approach and our goal was to build a model that performs best on our datasets. Each of the tested classifiers and feature selection approaches have their strengths and weaknesses (e.g. SVMs with different kernels are suitable for different kinds of separation spaces). Since several single features show low discriminatory power (Supplementary Figure S1) and many features are correlated to each other (Supplementary Figure S2) it is important to define feature subsets that allow to classify or declassify a new miRNA precursor as true precursor. Different scaling and feature selection methods can have substantial effects on the used classifier. Therefore, we performed an exhaustive analysis of all combinations. We evaluated 130 105 different combinations of feature selection and classifiers using repeated stratified 5-fold cross validation. Even with the cross-validation, the evaluation of so many different classification attempts may lead to overoptimistic results. To address this problem, we performed permutation tests. The evaluation of the key performance criteria in Table 1 shows that almost all classifications were highly accurate. The area under the receiver operating characteristic curve (ROC AUC) highlights median performance of 99%, with the 90% quantile of all approaches being at 99.5% and more impressively the 10% quantile being at 95.8%. In consequence, 90% of all 130 105 tested classifiers had an AUC exceeding 95.8%.

For both, the complete and the fast feature set AdaBoost outperformed the other models with an AUC of 99.6%, a specificity of 99.9% and a sensitivity of 86.9% for the complete feature set, and an AUC of 99.4%, a specificity of 99.9% and a sensitivity of 83.4% for the fast feature set. The selected AdaBoost classifier by itself selects only features known to improve the prediction and is therefore well suited for our broad set of features. This comparison demonstrates that the performance of the fast feature set is only marginally weaker than the performance of the full feature set. Nonetheless, we evaluated the performance of these two models and carried out stratified 5-fold cross-validation with 1000 repetitions each. The same approach was done with 1000 permutation tests each. As shown in Supplementary Figure S3, random test performance did not compare to the true performance in any of the cases and cross validation performance was stable and good in all cases. This further suggests that the composition of the cross-validation splits plays no major role for the model performance. In addition to the cross-validation performance we evaluated our model with the fast feature set on two independent test sets. A description of the independent test sets can be found in the Materials and Methods section. The first test set was composed of 129 human precursors and 28 human pseudo precursors. On this set our model reached a sensitivity of 82.9% and a specificity of 100%. The second test set contained 350 mouse precursors and 56 mouse pseudo precursors and resulted in a sensitivity of 81.4% and a specificity of 98.2%.

Evaluation of prediction from 1097 miRNA NGS samples

Having evaluated the performance of our classifier on the positive and negative training set we applied the models to 1097 in-house data sets (33–39). These contain 15 billion reads in a total file size of 486GB (see Table 2). Again, we first compared the fast feature set versus the complete feature set. The prediction was carried out for each sample individually. They were then merged and filtered according to their global read signature. The differences between the models regarding known miRNAs were minimal with both models discovering 900 precursors, while 55 additional were uniquely found in the fast model opposed to 34 in the full model, as shown in Supplementary Figure S4. As for the novel miRNAs both models discovered 10 651 precursors. We then compared the unique predictions of both models in regard to their mean probability, novoMiRank score and the number of samples they were predicted in. We found that their mean scores and the mean number of samples they were predicted in were very similar (score of 1.18 for the complete model, 1.19 for the fast one; predicted in 7.5 samples for the complete and 7.6 for the fast model). However, we noticed also that for both sets the majority of the differing predictions were near the decision boundary with a mean probability below 60% (in contrast to an average of 70% for the common set), meaning that these predictions were among the less likely precursor miRNA candidates. Therefore, since both models performed very similarly, ex-

Table 1. Cross validation performance

	Specificity	Sensitivity	Accuracy	NPV	Precision	ROC AUC	$F_{0.5}$
Median	99.78%	70.62%	98.61%	98.90%	91.37%	98.98%	85.10%
90% quantile	99.91%	82.35%	99.18%	99.34%	95.61%	99.50%	91.81%
10% quantile	99.44%	45.17%	97.41%	97.97%	73.60%	95.85%	64.80%
AdaBoost (all features)	99.98%	86.85%	99.51%	99.51%	99.54%	99.58%	96.71%
AdaBoost (fast features)	99.98%	83.37%	99.38%	99.38%	99.26%	99.39%	95.60%

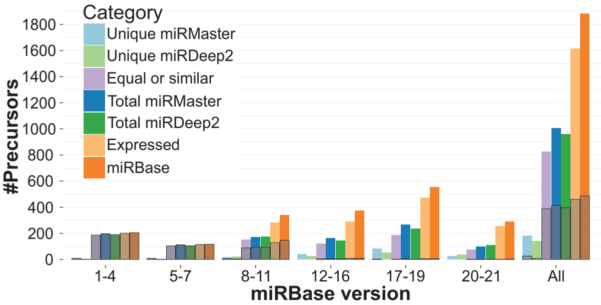


Figure 2. Distribution of recovered known miRBase precursors using miRMaster and miRDeep2. Predicted precursors are regarded as similar if they overlap by at least 90%. The black boxes show the number of precursors contained in the training set of miRMaster.

cept for the less likely candidates, we further focused on the fast model, due to its runtime advantage.

Comparison between miRMaster and miRDeep2

To further evaluate the performance of miRMaster we compared its predictions with the predictions of miRDeep2, one of the central programs for miRNA discovery. In detail, we ran miRDeep2 with default parameters on our 1097 NGS samples and merged the overlapping precursors predicted by miRDeep2 by retaining the precursors predicted in most samples. The same procedure was applied for miRMaster. A more detailed description of the different analysis steps can be found in the Methods section.

As shown in Figure 2, miRDeep2 recovered 59.5% of the known miRBase (version 21) precursors detected by quantification while miRMaster found 62.3% of them. Further, miRMaster consistently recovered more precursors from our training set than miRDeep2 (in total 414 versus 396). Specifically, 181 precursors were exclusively found by miRMaster and 138 by miRDeep2 as shown in Supplementary Table S3. Figure 2 shows that both tools perform especially well in earlier miRBase versions with both tools reporting nearly all precursors up to miRBase version 7. Precursor miRNAs exclusively detected by miRDeep2 are mainly found in later miRBase versions and contained only 7 precursors of miRNAs with strong experimental evidence for targets in miRTarBase. By contrast miRMaster detected 21 precursors in later miRBase versions with strong experimental evidence for targets in miRTarBase. These results might be biased since our models contain many more features and are trained using human high-confidence miRNAs on the one hand, and many miRNAs in later miRBase versions have already been reported by miRDeep2 on the other. Overall, the data suggest that our classifier identifies more known miRNAs and especially more of the strongly confident miRNAs.

To present a realistic comparison in runtime of miRMaster and miRDeep2, we measured execution time on the same infrastructure starting from pre-processed data. The computations were performed on a node with four AMD Opteron 6378 (4 × 16 cores totaling 64 cores) at 2.4 GHz and 512GB DDR3-RAM. MiRDeep2 required 102.5 h (4.4 days) without PDF generation (usually increases the runtime by 40% and produces reports for each known and predicted precursor). The respective steps of miRMaster required only 5.5 h which is a 19-fold decrease in runtime compared to miRDeep2. The difference is especially notable since miRMaster performed many additional analyses such as prediction of isoforms, variants in miRNAs and others. This difference in runtime is explained by the computed features and by different implementations. While miRDeep2 is implemented in Perl, miRMaster relies on a more efficient implementation in C++ for substantial parts of the program. One example is the precursor excision step, a reimplementation of the miRDeep2 Perl code in C++. This part of the program is roughly 40-fold faster in miRMaster as compared to miRDeep2.

A detailed break-down of the runtime in the different steps is presented in Supplementary Figure S5. The reads are mapped against miRBase and multiple other ncRNA databases (1.52% of the runtime) and to the human genome using Bowtie (56) (0.72% of the runtime). The afore mentioned precursor excision step requires 0.2% of the runtime. The following steps that are central for miRMaster include precursor segmentation, filtering, feature computation and prediction, altogether requiring 30.92% of the runtime. The predicted miRNA precursors from different samples are subsequently merged and filtered according to the read profiles of all samples (12.60% of the runtime). The following assignment to one of six categories ‘known’, ‘shifted known’, ‘one annotated’, ‘dissimilar overlapping’, ‘half novel’ or ‘novel’ requires 0.75% of the runtime. For the prediction flagging step, ncRNAs from Ensembl (57), lncRNAs from NONCODE (61) and known miRNAs from miRBase are mapped against the precursors (4.34% of the runtime). Finally, different secondary analyses are carried out on known and novel miRNAs, including quantification, which is again a reimplementation of miRDeep2, detection of isoforms and single base mutations. These steps, including the mapping of non-human reads to a collection of 7556 bacteria and 7026 viruses of NCBI RefSeq, permitting the detection of potential exogenous miRNAs, require in total 48.96% of the server runtime.

Table 2. Composition of all 1836 NGS samples

Source / Description	#Samples	#Reads	Compressed File Size
CNS lymphoma patients and controls (in-house)	44	884 Mn	25GB
Alzheimer patients and controls (in-house)	203	3.4 Bn	114GB
Cardiovascular disease patients and controls (in-house)	485	6.9 Bn	205GB
Multiple sclerosis patients and controls (in-house)	217	1.2 Bn	44GB
Blood cell fractions from healthy donors (in-house)	148	3.3 Mn	98GB
GSE64142 (monocyte-derived dendritic cells upon bacterial infection)	116	1.4 Bn	43GB
GSE53080 (myocardium, plasma and serum in heart failure patients)	185	925 Mn	36GB
GSE49279 (adrenocortical tumors)	78	1.2 Bn	34GB
GSE45159 (adipose tissue)	360	786 Mn	24GB
Sum	**1836**	**20 Bn**	**623GB**

Web-based analysis using miRMaster

With the development of miRMaster we aimed to provide a comprehensive web-based toolbox for an all-in-one miRNA analysis. In detail, the web-based tool has to (a) enable the analysis of HT-sequencing raw data without installing any software, even for data sets in the range of dozens of gigabytes; (b) perform the most common and further specialized analyses in an integrative manner; (c) return the results in a manner to be used for identifying interesting hits and for publication purposes by wet-lab scientists. These analyses are carried out in a fully integrated manner. From the raw data input (1097 compressed FASTQ files, 486GB) to final results for all calculations, miRMaster required 23.5 h. Data upload at client side was performed on an Intel Core i5–5200U Notebook with 12GB DDR3-RAM using Mozilla Firefox 48 and required most of the time (18 of the 23.5 h), while the analysis of pre-processed data took only 5.5 h. At client side, FASTQ files are first pre-processed (adapter trimming, quality filtering, read collapsing) and subsequently uploaded. The functionality is implemented in JavaScript such that no software has to be installed by the user. The runtime of this step may vary based on the equipment at user site and the bandwidth for data upload. Real world tests have demonstrated that studies including e.g. 50–100 samples can be evaluated in well below 5 h.

Evaluation of variations in miRNAs by miRMaster

First, we investigated the mutation frequency. For each known miRNA of each of the 1097 samples we searched the number of single base mutations. To reduce a bias depending on the coverage we considered only miRNAs and their variants covered by at least 30 reads in 100 samples. Out of 2147 detected miRNAs 333 fulfilled the criteria. Supplementary Table S4 lists the mutations found in all miRNAs. Overall the largest number of variants was discovered for hsa-miR-486-5p, which is abundantly expressed across all samples with two precursors. However, for the majority of miRNAs the number of variants is low with most miRNAs having two or less variants (67.3%). For some miRNAs, such as hsa-miR-6131 the unmutated form was almost never detected and only variants with mutations at position 8 and 14 were found. Another example is hsa-miR-1260b with the most abundant form showing an A→G mutation at position 8 (Supplementary Figure S6). However, for most miRNAs (91.6%) the wildtype was most expressed. Our results suggest that only a small set of miRNAs is frequently affected by mutations e.g. due to RNA editing. The low number of mutations is to be expected, since mutations, especially in the seed region, are likely to highly affect the miRNA regulation network.

Next, we calculated for each known miRNA the number of isoforms, analogously to the steps performed for the detection of single base mutations. After applying the abovementioned filter criteria, we found 277 miRNAs isoforms that are listed in Supplementary Table S5. As for the mutated miRNAs we found the by far largest number of isoforms for hsa-miR-486-5p, which is highly expressed in blood. In consistence with the single base mutation results, the number of variants is low for the majority of miRNAs with most miRNAs (53.8%) showing four or less variants. For most miRNAs (71.5%) we detected the canonical form as annotated in miRBase. The miRNA with most variants and without canonical form was hsa-miR-107. As shown in Supplementary Figure S7, the most expressed form of hsa-miR-107 with a median of over 60% was trimmed by four nucleotides from the 3′ end, resulting in a miRNA with 19 nucleotides. Further, we frequently observed a lack of a dominating isoform over all samples, as for example for hsa-miR-29a-3p (Figure 3). This is consistent with the idea that isoform expression varies depending on the context, such as the cell type, time or population. Since the canonical form was most expressed in only 33.6% cases, isomiRs apparently play an essential role in miRNA function.

Comprehensive version of the human miRNome

Currently, the total number of human miRNAs is controversially discussed. While miRBase currently contains 2588 human mature miRNAs (version 21), several studies propose even larger sets (e.g. Londin *et al.* (30), Backes *et al.* (11), Friedländer *et al.* (31), Jha *et al.* (32)). There exist two major challenges. First, the different miRNA sets are partially overlapping or contain miRNAs shifted only by few bases, adding a substantial redundancy. Second, the miRBase contains many false positive miRNAs, especially in later versions.

Using miRMaster we attempted to generate a coarse description of the human miRNome, i.e. we wanted to describe as many putative miRNA candidates as possible, being well aware that false positives are included (e.g. tRNA fragments, piRNAs or artifacts). This collection of potential candidates can be used to minimize further redundancy in upcoming high throughput studies.

Thus, in addition to our in-house NGS samples, we collected 739 samples from GEO (40), resulting in 1836 NGS

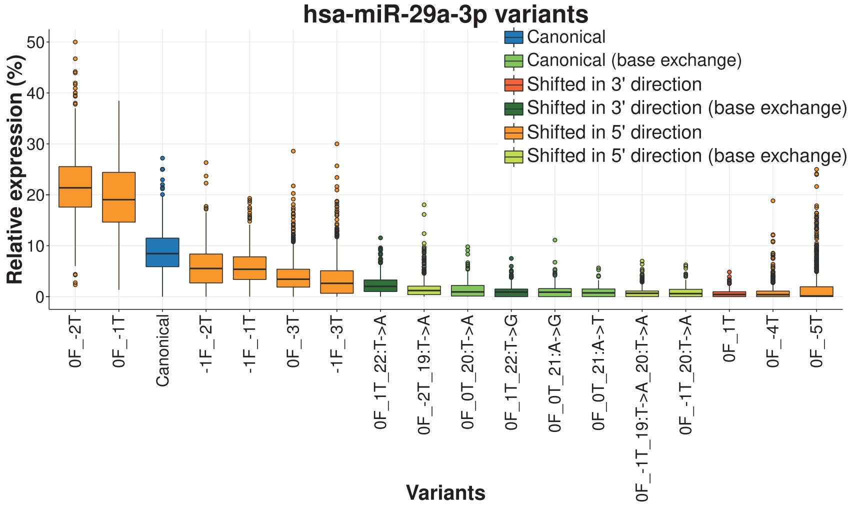


Figure 3. Isoform distribution of hsa-miR-29a-3p. Only variants appearing with an evidence of at least 30 reads in 100 samples are shown on the x-axis. Only reads occurring at least 30 times in a sample are shown for the relative expression to avoid large outlier due to low raw expression. Isoform notation: the number before F stands for the distance to the canonical 5′ end, in 5′-3′ direction (i.e. positive for trimmed, negative for extended); the number before the T stands for the distance to the canonical 3′ end (i.e. negative for trimmed, positive for extended). The canonical form is the third most frequent one and is highlighted in blue. Variants without base exchange are frequently shorter or shifted in the 5′ direction (orange), those with base exchanges match either the star/stop of the canonical miRNA (green) or are shifted slightly to the 5′ (light green) or 3′ (dark green) direction.

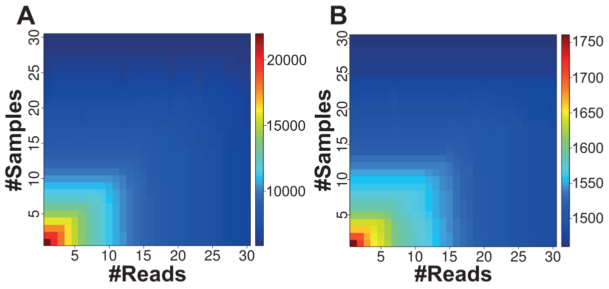


Figure 4. Distribution of the number of expressed precursors according to an evidence in a minimum number of samples and a total minimum number of reads. (**A**) The distribution of the number of expressed novel precursors. (**B**) The distribution of the number of known precursors.

samples (Table 2), and predicted novel miRNAs on those samples. The run resulted in 21 996 novel predicted miRNA precursors that are listed Supplementary Table S6. Those predictions can be inspected on the miRMaster webpage and downloaded as FASTA format. As shown in Figure 4A, most of the novel precursors were weakly expressed and in few samples. Considering only miRNAs with an expression in at least 30 samples reduced the number of predictions to 5845. As displayed in Figure 4B, the known precursors of miRBase (version 21) seem to be less affected by the augmenting number of samples or reads. Supplementary Figure S8 shows the number of expressed known and novel precursors according to their expression in multiple samples. The number of novel precursors decreases exponentially and faster than the known precursors with increasing number of required samples. This suggests that the majority of the commonly expressed miRNome is already known and that mainly tissue specific, time specific or other context specific miRNAs remain to be discovered.

Precursors of known and new miRNAs are evenly distributed on the positive and negative strands as shown in Supplementary Figure S9. The chromosomal distribution of known precursors largely matches with the distribution of the novel precursors as displayed in Supplementary Figure S10. In both cases, the least number of precursors can be found on chromosome Y. Chromosome 13, 18 and 21 harbor few known and novel precursors.

As for the number of motifs found in known and novel precursors with two annotated mature miRNAs, we found a slight enrichment of motifs in miRBase miRNAs (Supplementary Figure S11). A more fine-grained motif distribution is shown in Supplementary Figure S12.

Since miRNAs often occur in genomic clusters, we also searched genomic regions that are enriched by novel miRNAs. Supplementary Table S7 lists the positions of clusters when allowing a distance of at most 10 kb between the middle position of known or novel precursors. The largest cluster was composed of 46 known precursors and spanned 96 kb on chromosome 19. The largest cluster that contained both known and novel precursors was found on chromosome 14 and contained 42 known and 2 novel precursors and spanned 45 kb. In total 3969 clusters contained either known or novel precursors. Of these, 3423 clusters contained exclusively novel precursors. Further, 455 clusters contained both known and novel precursors and 91 exclusively known precursors. Supplementary Figure S13A and B shows the number of clusters with at least two or five precursors on each chromosome. Most clusters (394) with a minimum size of 2 could be found in chromosome 1. When focusing on clusters with at least five members, the numbers decreased to 154 clusters, 93 of which contained ex-

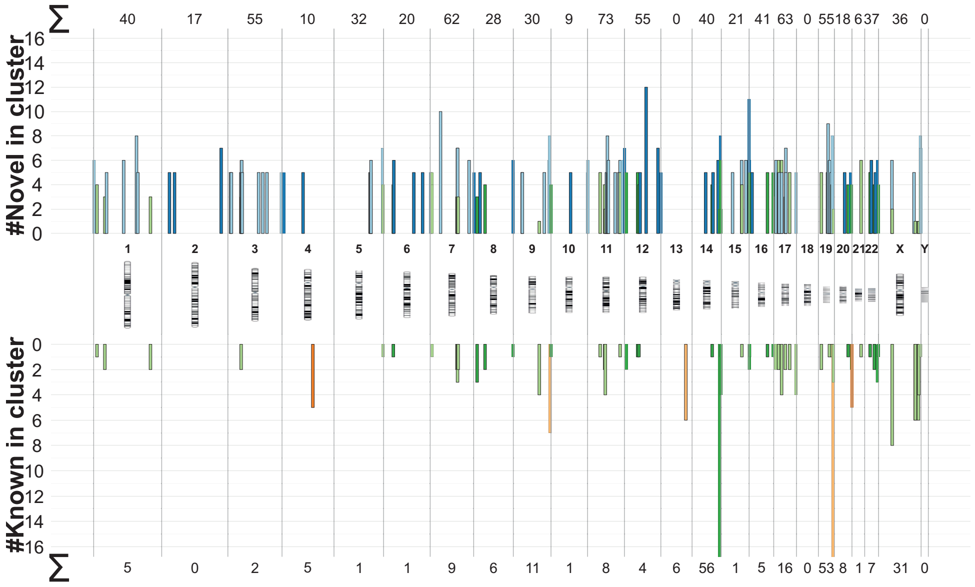


Figure 5. Distribution of the known and novel precursor clusters and their size on the human genome. Green clusters contain both novel and known precursors. Blue clusters contain only novel precursors and orange clusters contain only known precursors. The two known clusters on chromosome 14 and 19 (size 42 and 46) were trimmed for a better visualization. The sum of the number of novel or known precursors in all clusters of a chromosome with at least five members are shown on the top and bottom of the plot.

clusively novel precursors. Most clusters were observed on chromosome 11. Figure 5 shows the distribution of all clusters with five or more precursors over the human genome and demonstrates that many clusters contain both, known as well as novel precursors. The largest novel cluster with 12 precursors was found on chromosome 12.

To estimate how close our reported predictions might be to the coverage of the human miRNome, we performed predictions for different numbers of samples, each 10x randomly selected from our sample set. Supplementary Figure S14 shows the number of predictions according to the number of samples. We observe that the increase in number of predictions clearly exponentially diminishes with the number of samples. Since these predictions contain many false positives we expect the real part to be much smaller and the increase in predictions smaller as well. Therefore, we suggest that, at least for the tissues covered by our samples, we are close to the complete coverage of the human miRNome. We are aware and expect that the addition of samples of further tissue types or different conditions might add new candidates to our predicted set.

Expression analysis of miRNA candidates using custom microarrays

To provide further evidence that a relevant fraction of the aforementioned mature miRNAs is not only due to NGS bias or other artifacts such as RNA degradation, we built a custom human microarray. This array contains all miRBase v21 miRNAs, the miRNAs from the study by Londin *et al.* (30) and the top ranking miRNAs from the present study. The final microarray contained 11 866 miRNA candidates that have been measured each in 20 replicates (237 320 features per sample). For the microarray hybridization, we selected tissues from our Tissue Atlas (64) that contained the most miRNAs and added body fluids harboring likewise many miRNAs (65). The set of samples included a pool of PAXGene blood samples, a pool of plasma samples, lung tissue, brain tissue, kidney tissue, testis tissue, heart tissue and a reference pool from Agilent. Since degraded RNA is known to affect the miRNA patterns, we ensured high-quality of the used RNA samples. The RIN values of the different specimens ranged between 7.5 and 9. For the three sets of miRNAs the percentage of positive miRNAs in the hybridization experiments is presented in Figure 6A. For 56% of miRBase miRNAs, 55% of miRNAs by Londin *et al.* and 73% of miRNAs from the present study no positive signal in any sample was observed. On the other extreme, 11%, 17% and 8% were respectively positive in all experiments. The larger fraction of miRNAs not detected in any sample in the third set can be explained by the fact that many of the high abundant markers were previously already detected while we selected the candidates from the not yet discovered and likely much less abundant fraction. Still the results presented above can contain false positives (e.g. reagent contamination or positive signals induced by fragmented other RNAs) and false negatives (e.g. since other tissues or samples may harbor the miRNAs negative in the presently used samples or that are negative because of the limit of detection of microarrays). The same pattern as described can be recovered from the cluster analysis of all miRNAs from the three sets in Figure 6B. The lower part of this heat map shows that especially context sensitive miRNAs are observed among the set of miRNAs candidates only reported by miRMaster. In sum, the data

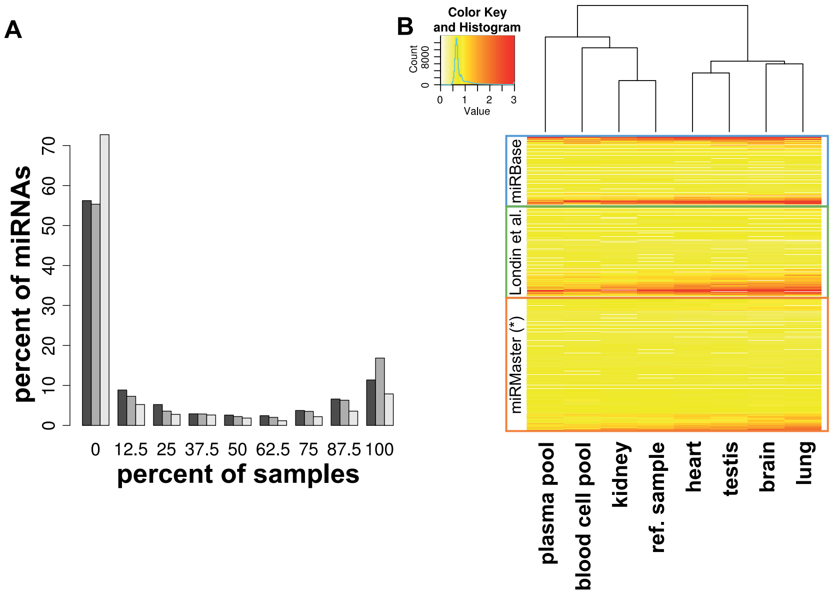


Figure 6. Expression of miRNA candidates on custom microarrays. (**A**) Distribution of the percentage of detected miRNAs in different samples. The colors correspond to the miRNAs of three studies: miRBase, dark gray; Londin *et al.*, medium grey; this study, light gray. (**B**) Heatmap of the logarithmized expression intensities of all miRNAs according to different tissues. For better visualization all expression values superior to 1000 were trimmed. The hierarchical clustering was performed with Euclidean distance.

strongly suggest that miRNAs exist which are currently not annotated in the miRBase. These miRNAs deserve further validation. All miRNAs from this analysis are contained in Supplementary Table S8.

CONCLUSIONS

The use of multiple web-based and standalone tools combined with different data formats makes the analysis of HT-seq miRNA data difficult, especially for wet-lab scientists. Therefore, we propose a web service that performs the most frequently requested applications directly from the raw FASTQ files. At the same time, experimental methods are advanced such that large-scale studies are feasible. Studies with many hundred or thousand samples are hard to be evaluated by current tools. Besides accuracy and specificity, runtime is among the most important criteria. Although miRMaster carries out a far greater number of analyses than other tools like miRDeep2, the running time of the miRMaster analysis was up to 20-fold faster. Of course, the precursor candidates predicted by miRMaster should in subsequent steps undergo a manual inspection and the selected ones be experimentally validated before calling them real miRNAs. A first validation step could be performed with our custom microarray followed by a more in depth validation of the detected interesting candidates using e.g. northern blotting. Applications such as target prediction, functional analysis and differential expression of known and novel miRNAs will in the future complete the portfolio of miRMaster.

ACCESSION NUMBERS

NGS samples are available on GEO under the following accession numbers: GSE64142, GSE53080, GSE49279, GSE45159 and GSE46579.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Best Ageing [306031] European Commission (in part); Saarland University; Siemens. Funding for open access charge: Internal funds of Saarland University.
Conflict of interest statement. None declared.

REFERENCES

1. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.
2. Castellano,L. and Stebbing,J. (2013) Deep sequencing of small RNAs identifies canonical and non-canonical miRNA and endogenous siRNAs in mammalian somatic tissues. *Nucleic Acids Res.*, **41**, 3339–3351.
3. Chiang,H.R., Schoenfeld,L.W., Ruby,J.G., Auyeung,V.C., Spies,N., Baek,D., Johnston,W.K., Russ,C., Luo,S., Babiarz,J.E. *et al.* (2010) Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev.*, **24**, 992–1009.
4. Jones-Rhoades,M.W. (2012) Conservation and divergence in plant microRNAs. *Plant Mol. Biol.*, **80**, 3–16.
5. Langenberger,D., Bartschat,S., Hertel,J., Hoffmann,S., Tafer,H. and Stadler,P.F. (2011) *Brazilian Symposium on Bioinformatics*. Springer, Vol. **6832**, pp. 1–9.

6. Meng,Y., Shao,C., Wang,H. and Chen,M. (2012) Are all the miRBase-registered microRNAs true? A structure- and expression-based re-examination in plants. *RNA Biol.*, **9**, 249–253.
7. Tarver,J.E., Donoghue,P.C. and Peterson,K.J. (2012) Do miRNAs have a deep evolutionary history? *Bioessays*, **34**, 857–866.
8. Taylor,R.S., Tarver,J.E., Hiscock,S.J. and Donoghue,P.C. (2014) Evolutionary history of plant microRNAs. *Trends Plant Sci*, **19**, 175–182.
9. Wang,X. and Liu,X.S. (2011) Systematic curation of miRBase annotation using integrated small RNA high-throughput sequencing data for C. elegans and Drosophila. *Front. Genet.*, **2**, 25.
10. Fromm,B., Billipp,T., Peck,L.E., Johansen,M., Tarver,J.E., King,B.L., Newcomb,J.M., Sempere,L.F., Flatmark,K., Hovig,E. *et al.* (2015) A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu. Rev. Genet.*, **49**, 213–242.
11. 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.
12. Hofmann,S., Huang,Y., Paulicka,P., Kappel,A., Katus,H.A., Keller,A., Meder,B., Stahler,C.F. and Gumbrecht,W. (2015) Double-stranded ligation assay for the rapid multiplex quantification of microRNAs. *Anal. Chem.*, **87**, 12104–12111.
13. Kappel,A., Backes,C., Huang,Y., Zafari,S., Leidinger,P., Meder,B., Schwarz,H., Gumbrecht,W., Meese,E., Staehler,C.F. *et al.* (2015) MicroRNA in vitro diagnostics using immunoassay analyzers. *Clin. Chem.*, **61**, 600–607.
14. Mestdagh,P., Hartmann,N., Baeriswyl,L., Andreasen,D., Bernard,N., Chen,C., Cheo,D., D'Andrade,P., DeMayo,M., Dennis,L. *et al.* (2014) Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nat. Methods*, **11**, 809–815.
15. Backes,C., Sedaghat-Hamedani,F., Frese,K., Hart,M., Ludwig,N., Meder,B., Meese,E. and Keller,A. (2016) Bias in High-Throughput Analysis of miRNAs and Implications for Biomarker Studies. *Anal. Chem.*, **88**, 2088–2095.
16. Friedländer,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2011) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
17. Hackenberg,M., Rodriguez-Ezpeleta,N. and Aransay,A.M. (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res.*, **39**, W132–W138.
18. Rueda,A., Barturen,G., Lebron,R., Gomez-Martin,C., Alganza,A., Oliver,J.L. and Hackenberg,M. (2015) sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res.*, **43**, W467–W473.
19. Guo,L., Yu,J., Liang,T. and Zou,Q. (2016) miR-isomiRExp: a web-server for the analysis of expression of miRNA at the miRNA/isomiR levels. *Sci. Rep.*, **6**, 23700.
20. Backes,C., Khaleeq,Q.T., Meese,E. and Keller,A. (2016) miEAA: microRNA enrichment analysis and annotation. *Nucleic Acids Res.*, **44**, W110–W116.
21. 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.
22. Enright,A.J., John,B., Gaul,U., Tuschl,T., Sander,C. and Marks,D.S. (2004) MicroRNA targets in Drosophila. *Genome Biol.*, **5**, R1.
23. Agarwal,V., Bell,G.W., Nam,J.W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, doi:10.7554/eLife.05005.
24. Akhtar,M.M., Micolucci,L., Islam,M.S., Olivieri,F. and Procopio,A.D. (2016) Bioinformatic tools for microRNA dissection. *Nucleic Acids Res.*, **44**, 24–44.
25. Auyeung,V.C., Ulitsky,I., McGeary,S.E. and Bartel,D.P. (2013) Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell*, **152**, 844–858.
26. Fang,W. and Bartel,D.P. (2015) The menu of features that define primary microRNAs and enable de novo design of microRNA genes. *Mol. Cell*, **60**, 131–145.
27. Alarcon,C.R., Lee,H., Goodarzi,H., Halberg,N. and Tavazoie,S.F. (2015) N6-methyladenosine marks primary microRNAs for processing. *Nature*, **519**, 482–485.
28. Tatusova,T., Ciufo,S., Fedorov,B., O'Neill,K. and Tolstoy,I. (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.*, **42**, D553–D559.
29. 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.
30. Londin,E., Loher,P., Telonis,A.G., Quann,K., Clark,P., Jing,Y., Hatzimichael,E., Kirino,Y., Honda,S., Lally,M. *et al.* (2015) 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.*, **112**, E1106–E1115.
31. Friedlander,M.R., Lizano,E., Houben,A.J., Bezdan,D., Banez-Coronel,M., Kudla,G., Mateu-Huertas,E., Kagerbauer,B., Gonzalez,J., Chen,K.C. *et al.* (2014) Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biol.*, **15**, R57.
32. Jha,A., Panzade,G., Pandey,R. and Shankar,R. (2015) A legion of potential regulatory sRNAs exists beyond the typical microRNAs microcosm. *Nucleic Acids Res.*, **43**, 8713–8724.
33. Leidinger,P., Backes,C., Deutscher,S., Schmitt,K., Mueller,S.C., Frese,K., Haas,J., Ruprecht,K., Paul,F., Stahler,C. *et al.* (2013) A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biol.*, **14**, R78.
34. Keller,A., Leidinger,P., Vogel,B., Backes,C., ElSharawy,A., Galata,V., Müller,S., Marquart,S., Schrauder,M., Strick,R. *et al.* (2014) miRNAs can be generally associated with human pathologies as exemplified for miR-144*. *BMC Med.*, **12**, 224.
35. Backes,C., Leidinger,P., Altmann,G., Wuerstle,M., Meder,B., Galata,V., Mueller,S.C., Sickert,D., Stahler,C., Meese,E. *et al.* (2015) Influence of next-generation sequencing and storage conditions on miRNA patterns generated from PAXgene blood. *Anal. Chem.*, **87**, 8910–8916.
36. Roth,P., Keller,A., Hoheisel,J.D., Codo,P., Bauer,A.S., Backes,C., Leidinger,P., Meese,E., Thiel,E., Korfel,A. *et al.* (2015) Differentially regulated miRNAs as prognostic biomarkers in the blood of primary CNS lymphoma patients. *Eur. J. Cancer*, **51**, 382–390.
37. Keller,A., Leidinger,P., Meese,E., Haas,J., Backes,C., Rasche,L., Behrens,J.R., Pfuhl,C., Wakonig,K., Giess,R.M. *et al.* (2015) Next-generation sequencing identifies altered whole blood microRNAs in neuromyelitis optica spectrum disorder which may permit discrimination from multiple sclerosis. *J. Neuroinflammation*, **12**, 196.
38. Schwarz,E.C., Backes,C., Knorck,A., Ludwig,N., Leidinger,P., Hoxha,C., Schwar,G., Grossmann,T., Muller,S.C., Hart,M. *et al.* (2016) Deep characterization of blood cell miRNomes by NGS. *Cell. Mol. Life Sci.*, **73**, 3169–3181.
39. 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.
40. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
41. Sacar,M.D., Hamzeiy,H. and Allmer,J. (2013) Can MiRBase provide positive data for machine learning for the detection of MiRNA hairpins? *J. Integr. Bioinform.*, **10**, 215.
42. Chou,C.H., Chang,N.W., Shrestha,S., Hsu,S.D., Lin,Y.L., Lee,W.H., Yang,C.D., Hong,H.C., Wei,T.Y., Tu,S.J. *et al.* (2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.*, **44**, D239–D247.
43. Kim,V.N., Han,J. and Siomi,M.C. (2009) Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.*, **10**, 126–139.
44. Xue,C., Li,F., He,T., Liu,G.P., Li,Y. and Zhang,X. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, **6**, 310.
45. Babiarz,J.E., Ruby,J.G., Wang,Y., Bartel,D.P. and Blelloch,R. (2008) Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes Dev.*, **22**, 2773–2785.

46. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.
47. Ng,K.L. and Mishra,S.K. (2007) De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, **23**, 1321–1330.
48. Batuwita,R. and Palade,V. (2009) microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, **25**, 989–995.
49. Lertampaiporn,S., Thammarongtham,C., Nukoolkit,C., Kaewkamnerdpong,B. and Ruengjitchatchawalya,M. (2013) Heterogeneous ensemble approach with discriminative features and modified-SMOTEbagging for pre-miRNA classification. *Nucleic Acids Res.*, **41**, e21.
50. Lee,M.T. and Kim,J. (2008) Self containment, a property of modular RNA structures, distinguishes microRNAs. *PLoS Comput. Biol.*, **4**, e1000150.
51. Zhang,B.H., Pan,X.P., Cox,S.B., Cobb,G.P. and Anderson,T.A. (2006) Evidence that miRNAs are different from other RNAs. *Cell. Mol. Life Sci.*, **63**, 246–254.
52. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B. Methodol.*, **57**, 289–300.
53. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
54. Kim,J., Levy,E., Ferbrache,A., Stepanowsky,P., Farcas,C., Wang,S., Brunner,S., Bath,T., Wu,Y. and Ohno-Machado,L. (2014) MAGI: a Node.js web service for fast microRNA-Seq analysis in a GPU infrastructure. *Bioinformatics*, **30**, 2826–2827.
55. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
56. 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.
57. Yates,A., Akanni,W., Amode,M.R., Barrell,D., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P., Fitzgerald,S., Gil,L. *et al.* (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.
58. Zhang,P., Si,X., Skogerbo,G., Wang,J., Cui,D., Li,Y., Sun,X., Liu,L., Sun,B., Chen,R. *et al.* (2014) piRBase: a web resource assisting piRNA functional study. *Database (Oxford)*, **2014**, bau110.
59. 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.
60. Lorenz,R., Bernhart,S.H., Honer Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
61. Zhao,Y., Li,H., Fang,S., Kang,Y., Wu,W., Hao,Y., Li,Z., Bu,D., Sun,N., Zhang,M.Q. *et al.* (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.*, **44**, D203–D208.
62. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
63. 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.
64. 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.
65. Fehlmann,T., Ludwig,N., Backes,C., Meese,E. and Keller,A. (2016) Distribution of microRNA biomarker candidates in solid tissues and body fluids. *RNA Biol.*, **13**, 1084–1088.

3.2 *A high-resolution map of the human small non-coding transcriptome*

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

3.3 *Technical Stability and Biological Variability in MicroRNAs from Dried Blood Spots: A Lung Cancer Therapy-Monitoring Showcase*

This article is available under: https://doi.org/10.1373/clinchem.2017.271619

This article is available under: https://doi.org/10.1373/clinchem.2017.271619

This article is available under: https://doi.org/10.1373/clinchem.2017.271619

This article is available under: https://doi.org/10.1373/clinchem.2017.271619

This article is available under: https://doi.org/10.1373/clinchem.2017.271619

This article is available under: https://doi.org/10.1373/clinchem.2017.271619

This article is available under: https://doi.org/10.1373/clinchem.2017.271619

This article is available under: https://doi.org/10.1373/clinchem.2017.271619

This article is available under: https://doi.org/10.1373/clinchem.2017.271619

This article is available under: https://doi.org/10.1373/clinchem.2017.271619

This article is available under: https://doi.org/10.1373/clinchem.2017.271619

This article is available under: https://doi.org/10.1373/clinchem.2017.271619

This article is available under: https://doi.org/10.1373/clinchem.2017.271619

3.4 *The sncRNA Zoo: a repository for circulating small noncoding RNAs in animals*

Published online 2 April 2019

Nucleic Acids Research, 2019, Vol. 47, No. 9 ***4431–4441***
doi: 10.1093/nar/gkz227

The sncRNA Zoo: a repository for circulating small noncoding RNAs in animals

Tobias Fehlmann[1], Christina Backes[1], Marcello Pirritano[2,3], Thomas Laufer[1,4], Valentina Galata[1], Fabian Kern[1], Mustafa Kahraman[1,4], Gilles Gasparoni[5], Nicole Ludwig[6], Hans-Peter Lenhof[7,8], Henrike A. Gregersen[9], Richard Francke[10], Eckart Meese[6], Martin Simon[2,3] and Andreas Keller[1,8,*]

[1]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [2]Molecular Cell Dynamics, Center for Human and Molecular Biology, Saarland University, 66123 Saarbrücken, Germany, [3]Molecular Cell Biology and Microbiology, University of Wuppertal, 42097 Wuppertal, Germany, [4]Hummingbird Diagnostics GmbH, 69120 Heidelberg, Germany, [5]Department of Genetics, Center for Human and Molecular Biology, Saarland University, 66123 Saarbrücken, Germany, [6]Department of Human Genetics, Saarland University Hospital, 66421 Homburg, Germany, [7]Chair for Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [8]Center for Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [9]Zoological Garden Neunkirchen, 66538 Neunkirchen, Germany and [10]Zoological Garden Saarbrücken, 66121 Saarbrücken, Germany

Received March 08, 2019; Editorial Decision March 20, 2019; Accepted March 29, 2019

ABSTRACT

The repertoire of small noncoding RNAs (sncRNAs), particularly miRNAs, in animals is considered to be evolutionarily conserved. Studies on sncRNAs are often largely based on homology-based information, relying on genomic sequence similarity and excluding actual expression data. To obtain information on sncRNA expression (including miRNAs, snoRNAs, YRNAs and tRNAs), we performed low-input-volume next-generation sequencing of 500 pg of RNA from 21 animals at two German zoological gardens. Notably, none of the species under investigation were previously annotated in any miRNA reference database. Sequencing was performed on blood cells as they are amongst the most accessible, stable and abundant sources of the different sncRNA classes. We evaluated and compared the composition and nature of sncRNAs across the different species by computational approaches. While the distribution of sncRNAs in the different RNA classes varied significantly, general evolutionary patterns were maintained. In particular, miRNA sequences and expression were found to be even more conserved than previously assumed. To make the results available for other researchers, all data, including expression profiles at the species and family levels, and different tools for viewing, filtering and searching the data are freely available in the online resource ASRA (Animal sncRNA Atlas) at https://www.ccb.uni-saarland.de/asra/.

INTRODUCTION

Since the establishment of the central dogma of molecular biology by Crick (1), for decades the main role of RNAs was believed to be either in the transfer of information between DNA and proteins (mRNAs) or in housekeeping functions (tRNAs, rRNAs). With the discovery of microRNAs in the early 1990s (2), research on small noncoding RNAs (sncRNAs) and later on long noncoding transcripts (3) gained traction. Moreover, advances in high-throughput sequencing technology that allowed the sequencing of millions to billions of small RNA fragments with reasonable effort and cost (4) led to a further growth in the field. Via sequencing-based approaches, the number of identified sncRNAs, especially of miRNAs, increased markedly in just a few years. While the reference repository miRBase (5) was established in the year 2000 with only 222 miRNAs in five species, the most recent version stores 48 885 miRNAs in 271 species. miRCarta (6), a database that collects mature miRNAs independently of the organism, suggests up to 44 347 miRNA candidates; however, only a fraction of these can be assumed to actually be true miRNAs. Because miRNAs have been described in a variety of organisms, their assumed conservation is frequently used to identify additional miRNAs in related species by homology- and sequence-based approaches (7–11), which often exclude expression profiling. Interestingly, the expression patterns of homologous miRNAs also appear to be comparable between organs in dif-

---

[*]To whom correspondence should be addressed. Tel: +49 681 302 68611; Fax: +49 681 302 58094; Email: andreas.keller@ccb.uni-saarland.de

© The Author(s) 2019. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

ferent species, as we successfully showed for human and rat (12).

One of the most commonly performed types of study on sncRNAs is biomarker discovery analysis (13–15). Here, human serum, plasma or blood cells are sequenced, or expression profiling using microarrays or real-time quantitative reverse transcription PCR (RT-qPCR) is performed. Blood cells are especially suitable for this as they contain many hundred to over 1000 human miRNAs (12,16). It has already been demonstrated that the use of standardized protocols for collecting and analysing blood-borne miRNA profiles has huge potential for comparing biomarker profiles across different human pathologies (17,18). Because blood can be obtained in a standardized manner and miRNA expression patterns are technically very stable, it is easy to accurately compare expression between different animal species. In particular, dried blood spots (19) (DBS) or microsampling devices (20) appear to be well suited as containers for miRNAs. While such decentralized collection kits are perfectly suited to collecting samples from different sites, the small amount of RNA that can be purified presents a challenge for further investigations. Previously, analyses based on DBS were mostly limited to microarrays and RT-qPCR, but excluded next-generation sequencing (NGS). However, the application of NGS was mandatory for our study to be able to compare the total sncRNA repertoires amongst different species. Thus, we developed a novel low-input-volume NGS protocol to facilitate sequencing from capillary microsampling devices starting with only 50 pg of RNA (20).

In the present study, we sequenced blood samples of 21 animals collected at two regional German zoos: in Saarbrücken and Neunkirchen. A phylogenetic tree of the animals is presented in Figure 1. The primary data analysis was performed with our tool miRMaster (21). We analysed and compared the read profiles as well as the distribution and composition of small RNAs across species. In addition, an online resource for the collected data was implemented and is freely available at: https://www.ccb.uni-saarland.de/asra/. This resource provides access to all detected sncRNAs, their families and their expression patterns across all species in this study. In summary, the compiled dataset and associated online web server constitute a valuable resource for sncRNA research, either for finding and validating miRNA candidates because of their conservation, or for general research on evolutionary aspects of sncRNAs.

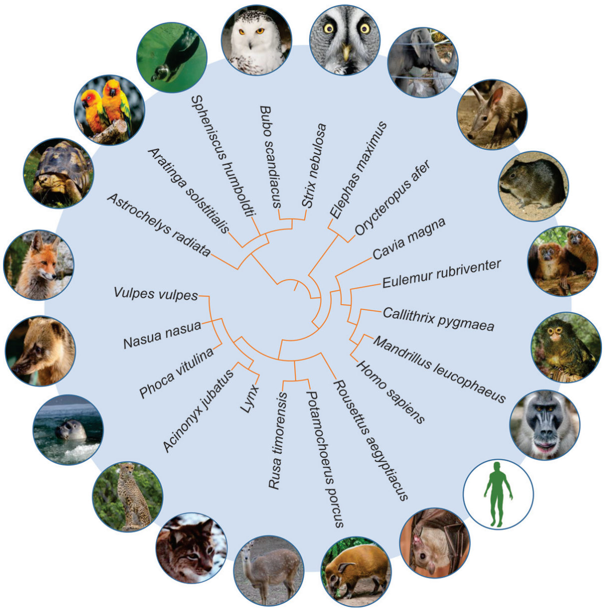


Figure 1. Circular taxonomy tree based on the species that were sequenced in our study.

MATERIALS AND METHODS

Sample collection

We collected 21 animal samples from regional zoos in Saarbrücken and Neunkirchen (Germany) comprising 19 different species. In addition, we collected four human samples as a reference. All blood samples were collected with the Mitra™ microsampler device (Neoteryx, CA). The samples were collected from remaining blood samples in the context of veterinary examinations. No additional examinations were performed with the animals. The study was permitted by the regional authority, the State Office for Consumer Protection (Landesamt für Verbraucherschutz). Human blood samples were collected from volunteers with informed consent. An overview of the samples in this study with their corresponding taxonomic classification is given in Table 1. Metadata containing the age, gender, as well as the health condition for each specimen are available in Supplementary Table S1.

RNA extraction and sequencing

Animal blood was collected onto Mitra™ collection devices (Neoteryx, CA) and dried at least for 2 h. Small RNAs were extracted by a modified version of the manufacturer's procedure using the miRNeasy Serum/Plasma Kit (Qiagen, Hilden, Germany). Size distribution and concentration were analysed using Agilent Bioanalyzer small RNA chips (Agilent Technologies, Santa Clara, CA). A total of 500 pg of sRNA with a size range of ∼15–150 nt was subjected to library preparation using a ligation-free procedure involving 3'-polyadenylation and template switch-based cDNA synthesis using the CATS sRNA-seq Kit (Diagenode, Liege, Belgium), omitting any dephosphorylation to enrich 3'-OH. Library size enrichment was carried out using 1.8 vol AMPure XP beads (Beckman Coulter, Krefeld, Germany) to achieve the enrichment of libraries containing RNAs larger than 15–20 nt (library size >160 bp). Libraries were multiplex-sequenced in an Illumina HiSeq 2500 platform in high-output mode with 50 cycles, except for common seal (1), human (3), pygmy marmoset, radiated tortoise and red-bellied lemur that were (re)sequenced with 40 cycles. Lynx (2) was sequenced with 47 cycles.

Table 1. Overview of the sequenced species ordered by phylogeny, their taxonomic classification, their total generated reads and remaining valid reads after filtering and trimming, as well as the availability of a genome assembly

Taxid	Species	Superorder	Order	Total reads (Mio)	Valid reads (Mio)	Genome
9568	*Mandrillus leucophaeus*	*Euarchontoglires*	*Primates*	72.65	52.19	✓
9606	*Homo sapiens*	*Euarchontoglires*	*Primates*	25.45	12.14	✓
9606	*Homo sapiens*	*Euarchontoglires*	*Primates*	15.46	10.02	✓
9606	*Homo sapiens*	*Euarchontoglires*	*Primates*	16.98	9.87	✓
9606	*Homo sapiens*	*Euarchontoglires*	*Primates*	24.50	19.26	✓
9493	*Callithrix pygmaea*	*Euarchontoglires*	*Primates*	38.80	27.76	✗
34829	*Eulemur rubriventer*	*Euarchontoglires*	*Primates*	35.50	21.09	✗
297387	*Cavia magna*	*Euarchontoglires*	*Rodentia*	36.54	25.38	✗
273791	*Potamochoerus porcus*	*Laurasiatheria*	*Artiodactyla*	32.85	24.68	✗
1088130	*Rusa timorensis*	*Laurasiatheria*	*Artiodactyla*	37.23	25.41	✗
9720	*Phoca vitulina*	*Laurasiatheria*	*Carnivora*	24.57	16.15	✗
9720	*Phoca vitulina*	*Laurasiatheria*	*Carnivora*	23.73	14.95	✗
9651	*Nasua nasua*	*Laurasiatheria*	*Carnivora*	46.87	34.31	✗
9627	*Vulpes vulpes*	*Laurasiatheria*	*Carnivora*	29.23	20.26	✓
13124	*Lynx*	*Laurasiatheria*	*Carnivora*	47.84	22.31	✗
13124	*Lynx*	*Laurasiatheria*	*Carnivora*	28.72	17.23	✗
32536	*Acinonyx jubatus*	*Laurasiatheria*	*Carnivora*	30.62	20.65	✓
9407	*Rousettus aegyptiacus*	*Laurasiatheria*	*Chiroptera*	33.75	24.99	✓
9783	*Elephas maximus*	*Afrotheria*	*Proboscidea*	97.67	63.16	✗
9818	*Orycteropus afer*	*Afrotheria*	*Tubulidentata*	36.68	26.45	✓
371907	*Bubo scandiacus*	*Neognathae*	*Strigiformes*	58.79	38.58	✗
126836	*Strix nebulosa*	*Neognathae*	*Strigiformes*	37.81	27.92	✗
176015	*Aratinga solstitialis*	*Neognathae*	*Psittaciformes*	43.77	28.29	✗
9240	*Spheniscus humboldti*	*Neognathae*	*Sphenisciformes*	72.75	53.41	✗
66190	*Astrochelys radiata*	*Chelonia*	*Testudines*	25.24	17.76	✗

Bioinformatics

Sample preprocessing. All samples were trimmed and cleaned using miRMaster (21). In detail, we first removed the template switch motif, i.e. the first three bases of the reads. Then, we removed the bases resulting from the polyadenylation process. Therefore, we first checked the reads for adenine homopolymers with at least 13 bases and at most one mismatch and, if no match was found, we relaxed the requirement for an adenine homopolymer with at least five bases and no mismatch starting at position 15 of the read. Finally, we removed sequencing adapter contamination. The quality filtering was performed using default parameters together with a sliding window of four bases and a quality threshold of 20. The resulting reads that were shorter than 17 nt were discarded.

Statistics and visualizations. All statistical tests were computed using the free statistical programming language R (22) (version 3.4.4). If not specified otherwise, reported *P*-values were adjusted for multiple testing using the Benjamini-Hochberg procedure (23). Cramer's V was computed using the R package rcompanion (24). Wilcoxon-rank sum test was applied when the data did not follow normal distribution according to Shapiro–Wilk test. Plots were generated using the R packages ggplot2 3.1.0 (25) and pheatmap 1.0.12.

Sample distance estimation and similarity to NCBI phylogenetic tree. We computed Mash sketches for all samples (using Mash 2.0 (26)) with a k-mer size of 17 and a signature size of 1000 and used them to estimate the pairwise sample distances. Reads were subsampled using Seqtk 1.2. We constructed a phylogenetic tree using the neighbour-joining approach (27) implemented in the R-package phangorn (28) and visualized it using the Interactive Tree of Life (29). The similarity to the phylogenetic tree provided by NCBI was computed using the normalized Robinson-Founds distance. To be able to compare both trees, we collapsed the nodes of the same species. We determined the significance of the similarity of both trees by creating 100 000 random trees with 20 leaves, labeled by the analysed species and comparing them with the NCBI tree. We then tested if the resulting distances were smaller than the computed distance and derived from this the *P*-value.

Rfam. We downloaded all Rfam family sequences from the Rfam FTP server (ftp://ftp.ebi.ac.uk/pub/databases/Rfam, version 13, accessed on 27/3/2018). Then, we determined that sequences were related to Metazoa by performing an SQL query against the Rfam database, and selected them accordingly. To this end, we used the following SQL query:

```
SELECT fr.rfam_acc, fr.rfamseq_acc,
   fr.seq_start, fr.seq_end, f.type
FROM full_region fr, rfamseq rf,
   taxonomy tx, family f
WHERE rf.ncbi_id = tx.ncbi_id
AND f.rfam_acc = fr.rfam_acc
AND fr.rfamseq_acc = rf.rfamseq_acc
AND tx.tax_string LIKE '
AND is_significant = 1
```

Next, we mapped all samples against the Metazoa Rfam sequences using RazerS 3 (30), while requiring at least 95% identity and allowing only forward mappings. We determined the RNA composition based on the RNA class annotations of each family. If a read mapped to multiple classes, it was counted in full for each.

miRNA homology determination. We collected the miRNA sequences of miRBase v22, miRCarta v1.0 and MirGeneDB 2.0 via their respective websites (accessed on 18 July 2018). To determine the expression of each miRNA, we mapped the samples against the databases with Bowtie (31) (version 1.1.2), while allowing no mismatches and disabling mapping against the reverse complement, using the following command:

```
bowtie -f -v 0 -a --fullref --norc
  -S <reference_mirnas_idx> <sample.fa>
```

To ensure that each read corresponds to a real miRNA, we discarded all reads with lengths different from those of their mapped miRNA. A miRNA was considered to be expressed in a species if it was present in at least one of its samples.

miRNA expression and potential precursor determination. MiRNAs found in any of the three considered databases were first clustered according to 90% sequence similarity using vsearch 2.7.1 (32), thereby merging potential isoforms into one cluster. The RPM normalized counts for each cluster were determined by summing up the expression of each miRNA contained in it. MiRNA arms were determined according to their annotation in the databases. Potential precursors were determined for the miRNAs by considering all combinations of 5′ and 3′ miRNAs of precursors of the same precursor family for MirGeneDB, with the same base name for miRBase and according to the exact annotations in miRCarta. MiRNAs that could not be assigned unambiguously to one arm were discarded. Using the thereby obtained potential precursors, we could then compute arm ratio differences to investigate arm switches.

MiRNA candidate prediction. MiRNA candidates were predicted using mirnovo (33) (downloaded on 20 July 2018) with the default parameters, except for the brown-nosed coati, for which we had to increase the required minimum number of isoform variants from 1 to 3 because the program was not terminating with lower numbers. Predicted miRNAs were filtered in a first step by only keeping those that did not map with at least 90% identity to any known miRNA. The mapping was performed with RazerS 3 (version 3.5.8). Subsequently, we built a scoring scheme similar to our tool novoMiRank (34). In a first step, we determined the values of the features used by mirnovo for known miRNAs in our dataset. To this end, we restricted the known miRNAs to those contained in the high-confidence set of miRBase v22, as we recently showed that this subset contains by far the largest fraction of true miRNAs (35). The features of mirnovo do depend not only on the miRNAs but also on the samples. It is thus possible that some miRNAs that are more expressed than others bias the feature distribution. To avoid this bias, we took the mean feature values for every miRNA. We then normalized all features to a mean of zero and a variance of once, since they were all on different scales and computed z-scores for all known miRNAs. To avoid too large influences of single features, we restricted the absolute values to 3. We then computed for every predicted miRNA its distance to the distribution of known miRNAs, for every feature, and reported the mean z-score. As filtering threshold we chose the 80th percentile of the z-scores of known miRNAs, corresponding to 0.8 standard deviations above or below the mean of the known miRNAs.

ASRA. In the web resource, we provide a species specificity index (SSI) for miRNAs and for Rfam families that describe the variability of their expression patterns. It is computed analogously to the tissue specificity index used in our miRNA tissue atlas (12). It allows measurement of the specificity of expression of an miRNA/Rfam family over different species. The SSI ranges from 0 to 1, where values closer to 1 represent molecules expressed in a few or only one species (species-specific molecules) and values closer to 0 represent molecules similarly expressed in many species (well-conserved molecules). To this end, the SSI for an miRNA/Rfam family j is calculated as follows:

$$ssi_j = \frac{\sum_{i=1}^{N}(1 - x_{j,i})}{N - 1}$$

where N corresponds to the total number of species and $x_{j,\,i}$ is the RPM expression of the miRNA/Rfam family j in species i normalized by the maximal expression in any species of miRNA/Rfam family j.

RESULTS

Using the Mitra™ system, we collected a total of 21 specimens from two regional zoos, including 19 animal species, as well as four human samples. The species in this study belong to five different superorders and 11 different orders. The samples were sequenced on an Illumina HiSeq 2500, yielding a total of 973 994 362 reads. After quality filtering and adapter trimming 654 217 441 reads remained and were used for downstream analysis. An overview of the collected samples, their taxonomy and read counts is presented in Table 1. Due to the fact that for only five of the sequenced animal species a genome assembly is available to date, of which all are on scaffold level, no genome mappings were computed. Also, no miRNAs were annotated in any of the considered reference databases. All downstream analyses were performed only with the valid reads.

Read profiles resemble phylogenetic descriptors

One of the core hypotheses in this study is that the differences in read profiles between the species also mirror their known taxonomic classification. To test this hypothesis, we conducted a minHash analysis using Mash (26). The top panel of Figure 2 shows the resulting 2D embedding based on the computed sample Mash distances for superorders (2 A) and orders (2 B). For the superorders, we observe a cluster pattern matching what one would expect from their taxonomy, with the exception of *Afrotheria*. In the more detailed 2D embedding for orders, we see that samples belonging to *Primates*, *Carnivora* and *Strigiformes* cluster together well. Since the amount of reads for our samples varied greatly we wanted to estimate this influence. Therefore, we generated embeddings based on 15 times subsampling of the depth of the smallest sample, for each sample. This way, we ensure that all samples have the same size, while still keeping a realistic sequencing depth. The resulting plots

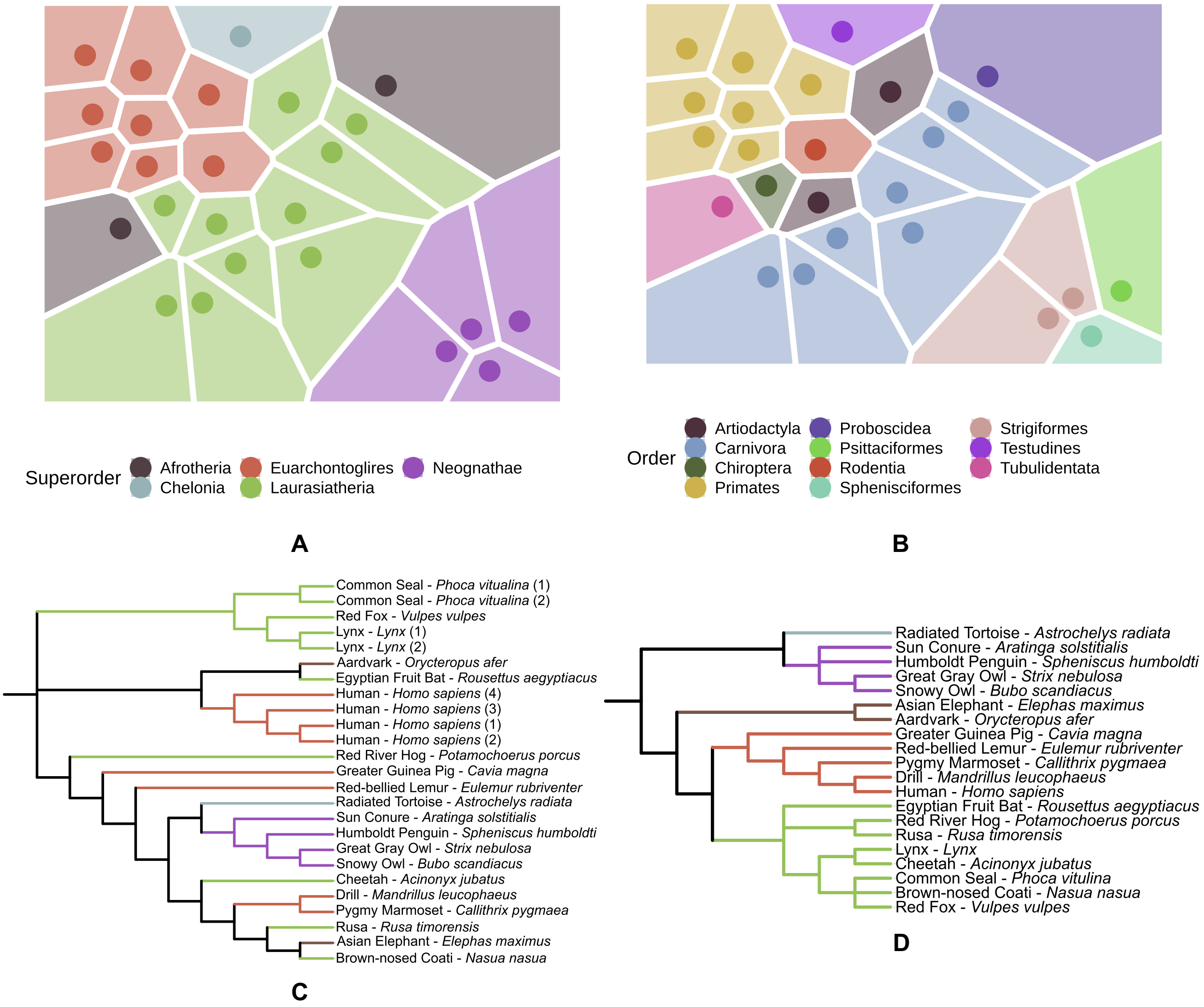


Figure 2. 2D embedding including a Voronoi diagram of the pairwise sample Mash distances for superorders (**A**) and orders (**B**). Each point in the plot represents a sample. Taxonomic tree built using the computed Mash distances of the read profiles at the species level (**C**) in comparison to the taxonomic tree derived from NCBI (**D**). The branches are colored according to the superorder of the corresponding species.

(Supplementary Figure S1) show that the sample depth has only a minor influence on the clustering. To increase the resolution to the species level, we visualized the computed Mash distances as a phylogenetic tree, as shown in the lower panel of Figure 2, in comparison to the phylogenetic tree from NCBI. The biological replicates for human, common seal and lynx cluster together, confirming the reproducibility of the sample collection and sequencing process. For some species, the clustering in the Mash tree matches very well with the partitioning in the NCBI taxonomy tree; for example, the two owls cluster with the Humboldt penguin and the sun conure, which form a larger cluster with the radiated tortoise. Drill and pygmy marmoset also cluster together in both trees; however, the human samples do not cluster with these species as we would expect from the NCBI phylogenetic tree, which is partly related to the heuristic nature of the neighbour-joining algorithm used to create the tree. To quantify the resemblance of both trees, we computed the normalized Robinson-Foulds distance between both trees ($D = 0.8$) and found that it was significantly lower than expected by chance ($P = 4 \times 10^{-5}$). While some of the remaining sample clusters do not fit the known taxonomy perfectly, we still see that, based on the distance of read profiles alone, we can derive evolutionary relationships to a certain extent.

Distribution of sncRNAs varies across species

To obtain an overview of the distribution and composition of sncRNAs across species, we mapped their reads to the sequences from the Rfam database (36) with a threshold of 95% identity. We then evaluated the quality of the mappings

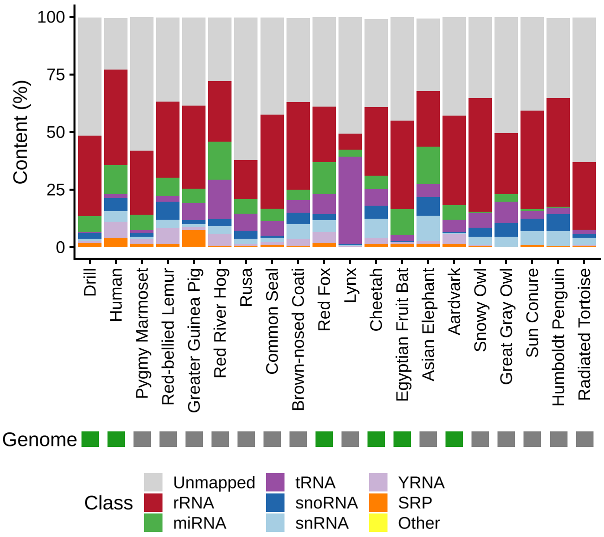


Figure 3. Overview of reads mapped to the different Rfam classes for all species in this study. The colors are ordered according to the median mapping ratio of each class. Classes with mapped reads <0.05% are summarized in the category 'Other'.

by inspecting the distribution of their read lengths after trimming (Supplementary Figure S2) and comparing them with the distribution of the mappings against every RNA class of Rfam (Supplementary Figures S3–10). We observe in all sample peaks at the length of the sequenced reads (minus 3 nt of the template switch motif), i.e. at 47 nt and for some that ran with less cycles at 37 nt. In general, we would expect that for RNA classes that are longer than the read lengths, and which have no known functional fragments, mostly untrimmed reads map. This is the case for rRNAs where we observe mainly untrimmed reads. It holds also for snRNAs, where only in few species over 15% of the reads shorter than 30 nt map. Reads mapping to SRP RNAs are mainly untrimmed reads as well; however, in some species the length of the mapped reads is nearly evenly distributed. YRNAs, as well as tRNAs, are either mostly covered by untrimmed reads or reads in the length of YRNA and tRNA fragments (around 26 nt and around 32 nt). For reads mapping to miRNAs, we observe clear mapping patterns that show peaks at 21–22 nt, with mostly no mapping read exceed a length of 24 nt. Considering snoRNAs, we observe mostly mappings of untrimmed reads, except for some species with peaks around 26 nt. Finally, all other mapping reads are composed mostly of untrimmed reads or short reads around 20 nt. The overall results of the mapping distribution are presented per sample in Supplementary Figure S11 and summarized per species by taking the average mapping fraction in Figure 3. As expected, in almost all species, the most dominant read fraction is represented by rRNAs. However, the percentages vary substantially across species: from 7% in lynx to 49.3% in snowy owl, with a median of 35.2%. In particular, the composition of the RNA classes in both lynx samples diverge the most from those in the other species. Here, not only is the rRNA fraction very small, but also the tRNA fraction (which is in median the third most abundant class) represents 38.1% of the sncRNA reads. In most other species, the fraction of tRNAs is under 10% (median 5.5%). The distribution of miRNAs, which are the second most abundant RNA class, also varies amongst the different species, ranging from 0.2% in radiated tortoise to 16.4% in red river hog. Similar patterns could be observed for all other RNA classes. Interestingly, the fraction of miRNAs, but also of YRNAs, was highly underrepresented in all species of the *Neognathae* and *Chelonia* superorder (miRNA mean: 1.1% versus 8.7%, Wilcoxon rank-sum test $P = 5 \times 10^{-6}$; YRNA mean: 0.27% versus 2.9%, Wilcoxon rank-sum test $P = 4 \times 10^{-4}$). The differences in the compositions of RNA classes might also be influenced by the number of unmapped reads. Human reads are much better recovered in Rfam than reads of rusa and radiated tortoise, for example (unmapped: ~23% versus ~62%, respectively). We investigated if the mapping rates were associated with the presence of a genome assembly; however, no significant association was found (Wilcoxon rank-sum test (two-sided) $P = 0.968$). A chi-square test of homogeneity showed that all pairwise sample comparisons differ significantly ($P = 0$). Since the *P*-values are strongly affected by large read counts, we also computed the effect sizes using Cramer's V, see Supplementary Table S2. Thereby, we found that the values for samples of the same species (median: 0.16) were significantly smaller (i.e. the class distributions were more similar to each other) than for samples between different species (median: 0.31, Wilcoxon rank-sum test (one-sided) $P = 9 \times 10^{-6}$), highlighting that even though all RNA class distributions were significantly different, the heterogeneity between samples of different species was higher than between samples of the same. To assess if the observed class distributions of some RNA classes are related to each other, we computed all pairwise Spearman correlation coefficients (Supplementary Figure S12) on the number of reads mapped to each class. This showed that miRNA and YRNA levels, as well as snoRNAs and snRNAs, are significantly and positively correlated to each other ($\rho = 0.72$, $P = 6 \times 10^{-4}$ for miRNAs and YRNAs, and $\rho = 0.89$, $P = 3 \times 10^{-5}$ for snoRNAs and snRNAs).

Zoo animals express common miRNA families that are more conserved than previously assumed

We also evaluated the coverage of known miRNA sequences and miRNA families in the different species. To obtain a comprehensive overview, we made use of three different miRNA databases with different scope: miRBase v22 (5), miRCarta v1.0 (6) and MirGeneDB 2.0 (37). miRBase is the gold standard resource for miRNAs; miRCarta also collects many miRNA candidates, of which only a fraction might be true miRNAs; and MirGeneDB collects miRNA genes that are manually curated and validated. We mapped the reads of the different species against the mature miRNA sequences of the three different databases, allowing only exact matches, which means that we count only reads that have exactly the same sequence and length as the sequence deposited in the corresponding database. Figure 4 summarizes the findings for the three databases separately, as well as the results overlapping amongst them. As a me-

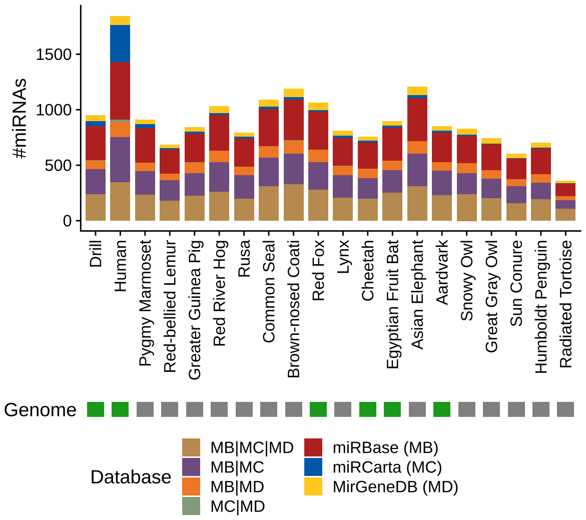


Figure 4. Comparison of mapping the reads of the different species against the three miRNA databases: miRBase, miRCarta and MirGeneDB. The mapping was performed with perfect matches, allowing no mismatches or differences in lengths between read and database sequence. The stacked barplot shows the number of miRNAs found uniquely in the corresponding databases, as well as the different overlaps amongst the databases.

dian, we recovered 847 miRNAs per sample. Because human is the organism with the most annotated miRNAs, we recovered the most miRNA sequences in human (n = 1846), followed by Asian elephant (n = 1210) and brown-nosed coati (n = 1187). At the lower end, the reads of the radiated tortoise sample recovered only 358 miRNAs. We could expect the number of recovered miRNAs to be significantly higher in species with known genome; however, this was not the case (Wilcoxon rank-sum test (one-sided) $P = 0.1037$). Although a large proportion of the miRNA sequences overlap with references in the three databases or in any combination thereof, we still found many unique hits of the reads, especially for miRNAs from miRBase. While this is surprising at first glance, it can be explained by the difference in set-up between miRCarta and miRBase. In these databases, similar miRNAs are merged into one representative, but miRBase might contain variants of the same miRNA sequence with different lengths. Nonetheless, for assessing which miRNAs actually exist, these sequences uniquely recovered in the different databases might provide new insights, because they appear to be expressed in different species in our study. To this end, we analysed the uniquely recovered sequences in miRBase in more detail. In total, we discovered 862 unique miRBase sequences, of which 44 were found in all 20 species in our deep sequencing approach. Interestingly, most of these have been described in only three different organisms in miRBase, on average. Amongst those 44 recovered sequences, there are many representatives of well-known families, such as let-7, mir-17, mir-103, mir-24, mir-181 and mir-92. Our findings indicate that these miRNAs are expressed in substantially more species than previously assumed and provide new insights into their conservation. If we look at the unique miRBase sequences recovered that have the most miRBase organisms' annotations, but are found in only a few of the species in our analysis, we might conclude that these are either not as evolutionarily conserved or predominantly expressed as isoforms with different sequence lengths, or might even represent artefacts that have been derived by sequence-based homology but not by expression analysis. One such example is the sequence 5′-CUGCCCUGGCCCGAGGGACCGA-3′, which is only found in one species amongst our samples, but is annotated in 10 miRBase organisms. However, if we remove one base at the 3′ end from this, we also find this sequence in seven further organisms in our study and in two from miRBase. Essentially, this shows that this sequence might be a conserved miRNA, but occurs in at least two isoforms of different lengths. The uniquely recovered miRBase sequences, the number of species they cover in our study and in how many miRBase organisms the sequences are annotated are shown in Supplementary Table S3.

Some sncRNAs are processed depending on the superorder of their species

Small noncoding RNAs and especially miRNAs are known to be expressed differently in organisms depending on various factors such as diseases, developmental stages or tissues. Therefore, we asked if we could find such relationships between our species as well, and in particular if this would be related to phylogeny. In a first step, to avoid biases related to isoforms, we clustered all detected miRNAs with an identity of at least 90% together and summed their expression values. Next, we clustered the miRNAs that represented at least 0.1% of the total miRNA expression in the corresponding species and that were present in at least 5 species (see Supplementary Figure S13). There, we observed that the strongest split between the species happened between those of the superorders of Neognathae and Chelonia in comparison to the other three. This is in concordance with our observations made in the previous analyses, as well as with the phylogenetic tree provided by NCBI. One example of miRNA expressed nearly exclusively in Neognathae and Chelonia is miR-2188-5p. This miRNA is expressed with a median of over 30 000 reads in those species, whereas in others we found it in at most 328 reads. In opposition, for example miR-423-3p is mostly expressed in Afrotheria, Euarchontoglires and Laurasiatheria (median of over 25 000 reads) but nearly not in Chelonia and Neognathae (at most 467 reads). We also evaluated if either 5′or 3′ miRNAs were over-represented amongst the evaluated miRNAs; however, their numbers were very similar (66 5′ miRNAs, 63 3′ miRNAs and 48 either undetermined or miRNAs that have been annotated on 5′ and 3′ positions). The observed differences led us to the question if there were potential miRNA precursors that indicated arm switches between species of different superorders. Supplementary Figure S14 shows the fraction of 5′ minus 3′ miRNA reads (1 being thus precursors exclusively expressing their 5′ miRNA and -1 their 3′ miRNA) of potential precursors, derived from the known annotations. We see that most precursors express mainly one form across all species. However, there are some for which there is no clear form. We decided to investigate those further, in particular regarding differences at the superorder level and found nine potential precursors with large differences between the Neognathae and Chelonia superorders in contrast to the Afrotheria, Euarchontoglires and Laurasiatheria superorders (see Supplementary Figure S15). However, differential processing seems to be not only limited to miRNAs, since we found for example different processing profiles for fragments of SNORD14 enriched in most species at the 5′ end, but showing clear preferences for fragments at the 3′ end in great gray owl, red fox and sun conure, as shown in Supplementary Figure S16.

Gender and health condition have limited impact in cross-species RNA expression

Others and we have shown that expression levels of certain sncRNAs, in particular miRNAs, are driven by gender or disease conditions (38–40). Therefore, we evaluated if we could observe different expression levels of Rfam families or miRNAs according to the gender or health conditions (unaffected versus affected) of our sequenced species. We did not perform a more fine grained comparison by disease, since the group sizes would have been too small and some miRNAs, such as miR-144-5p, have been shown to be deregulated independent of the disease in human (39). While significantly differing miRNA and Rfam family levels were found according to a two-sided Wilcoxon rank-sum test (gender specific: RF01412 ($P = 0.013$), miR-224 ($P = 0.026$); health condition specific: RF00009 ($P = 0.0025$), miR-238|miR-548c|miR-1842 ($P = 0.009$)), none remained significant after adjustment for multiple testing. Therefore, we conclude that the impact of these variables in a cross-species setup is too small and that differences between the species dominate the expression levels.

Many miRNA candidates are not covered by known databases

In addition to known miRNAs from the databases above, it is likely that there are other small noncoding RNAs that have not yet been annotated. A mapping-based analysis using a reference genome usually supports the discovery of these candidates. Because, for the majority of the animals included in this study, no reference genome is available, we applied mirnovo for genome-free miRNA prediction (33). First, we assessed how many known miRNAs can be recovered by a run of this tool. Figure 5A shows a stacked barplot for the number of recovered miRNAs deposited in the databases miRBase, miRCarta and MirGeneDB. In this case, we defined a positive hit if the reads mapped with at least 90% identity to the miRNA sequence in a database taking into account mismatches and differences in length. The prediction algorithm recovers most known miRNAs for human, followed by lynx, Egyptian fruit bat and common seal. In contrast to the comparison of the perfect matches above, we see that the largest fraction of recovered miRNAs is shared by all three databases for each organism and that miRCarta entries contribute the largest proportion. Still, the number of recovered miRNAs is moderate overall; even for human, we recover only 360 miRNAs. As a median, we recover only 40.5 miRNAs across all samples. Second, we analysed the results of the mirnovo algorithm by excluding known miRNAs and illustrate the numbers of novel predictions in Supplementary Figure S17. Here, as a median, approximately 575 miRNAs per species remain. The organism yielding the most candidates is sun conure, with more than 2000 predicted miRNAs, followed by Asian elephant with 1298. Because the gap between known recov-

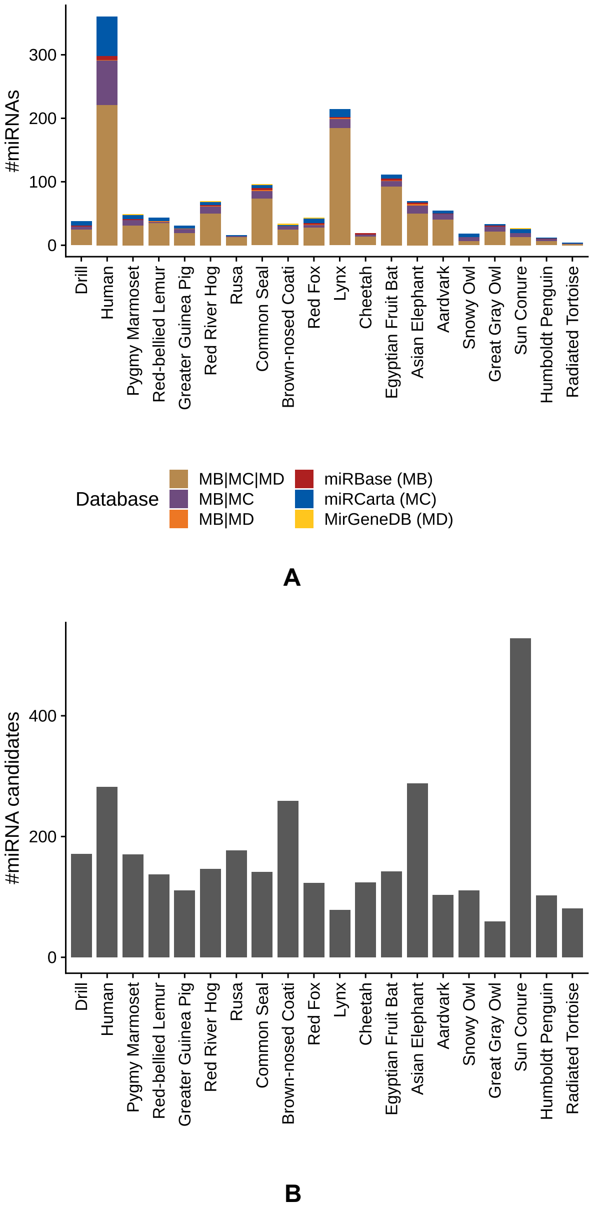


Figure 5. Prediction of novel miRNAs with the tool mirnovo. (**A**) Comparison of recovered known miRNAs deposited in the three databases: miRBase, miRCarta and MirGeneDB. For the mapping, we required at least 90% identity between read and database sequence. The stacked barplot shows the number of miRNAs found uniquely in the corresponding databases, as well as the different overlaps amongst the databases. (**B**) Number of novel miRNAs predicted by mirnovo and filtered by us for the samples in this study.

ered miRNAs and novel miRNAs is quite large, it is questionable how many of the predicted candidates represent true positive findings. To increase the likelihood of predicting true miRNAs, we applied a score filtering similar to novoMiRank (34), based on the features of mirnovo. The obtained scores (see Supplementary Figure S18) highlight that many predicted miRNAs are very different from the miRNAs of the high confidence set of miRBase. By filtering the predictions according to their scores, we reduced the number of predictions by 4-fold in median, as show in Figure 5B, while the number of recovered miRNAs dropped in median only by 2-fold (see Supplementary Figure S19). The results of the filtered mirnovo analysis are available in our online repository.

ASRA: the online resource

In the previous sections, we provide only a snapshot of the potential analyses that are possible using the NGS dataset, excluding many further considerations, such as animal-specific miRNA arm expression preferences, isoforms and others. To make our findings and data easily accessible to others and to promote secondary analyses, we implemented the online resource ASRA (Animal sncRNA Atlas), available at https://www.ccb.uni-saarland.de/asra/. ASRA consists of five major modules. First, we provide an overview of all studied samples and display their read profile similarity in comparison to their phylogenetic annotations, represented as a 2D embedding plot and a phylogenetic tree. Second, users can search specific miRNAs or Rfam families in the databases considered here (miRBase, miRCarta, MirGeneDB and Rfam) and display their expression in all species (for an example, see Supplementary Figure S20). Thereby, the total read counts or expression normalized as the reads per million (RPM) can be shown, as well as the expression of known similar miRNAs (known miRNAs with 90% similarity to the selected one). In addition, a species specificity index is shown for each entry, which indicates whether the displayed RNA is preferentially expressed in few species (values closer to 1) or ubiquitously in all species (values closer to 0). Third, each organism and considered database can be browsed separately; for example, for each organism we provide an overview of the number of reads and their mapped fraction, as well as their class distribution. In addition, detailed mapping information, such as total reads and average RPM, are displayed for the three analysed miRNA databases, the predicted miRNA candidates, the Rfam RNA families as well as their Gene Ontology terms. In particular, for Rfam RNA families, we provide coverage plots with the average RPM at each position of the 500 most expressed family members. All tables can be filtered according to their miRNA/RFAM IDs, their expression or the number of samples in which the sequence was found. Because Rfam families are composed of many sequences, we provide a detailed view for each family and species, which comprises the fourth usability feature. Users can then see if the detected parts of the family are common to many family members or if they are specific to few members. Furthermore, we enable the family coverage profiles to be directly compared amongst different species, which can highlight differences such as miRNA arm expression preferences (arm switches). Finally, users can search nucleotide sequences, either exactly or as part of a read, in all samples of the database and inspect their distribution amongst all species.

DISCUSSION

High-throughput sequencing in combination with microsampling devices allows the generation of data from species for which normal sample collection would be challenging. In our study, we collected blood from a variety of different species at German zoos and compared their small noncoding RNA profiles.

In the first steps of data analysis, quality filtering removed a considerable number of reads. This is probably due to two factors: as we used a minimally invasive method for sampling peripheral blood, the amount of RNA was indeed limited. We consequently chose a library preparation protocol suitable for low input amounts based on ligation-free template-switching cDNA generation. To this end, we used total small RNA fractions from precipitation-free isolation from dried blood without further size exclusion. As such, a high number of very small reads (shorter than 17 nt) were obtained and thus discarded. Next, we used 3′ polyadenylation of small RNAs before reverse transcription, which then requires the trimming of poly(A) stretches. Here, any small RNA with a poly(A) region is trimmed, as we cannot differentiate this from *in vitro* poly(A). For the unmapped fraction of reads and also for species for which, to date, no genome is available, it is unlikely that we sequenced many RNA degradation products, as we omitted any dephosphorylation and therefore enriched the library for 3′-OH RNAs.

Analysing the similarity of the read profiles by computing the Mash distances revealed that most of the samples of the same superorders and orders clustered together. Even at the species level, we still found two groups (birds and primates) that clustered in a way that was comparable to the phylogenetic taxonomy in NCBI. To the best of our knowledge, this is the first study showing that *k*-mer profiles derived from small RNA reads across many species still maintain the known evolutionary relationships.

Upon considering the distribution of RNA classes across species, we could not observe a clear pattern. As expected, rRNA constituted the dominant fraction in most species, with some exceptions. The number of reads that could be mapped to the Rfam classes varied enormously amongst the species. Human had the best coverage, but is also amongst the best annotated and most researched organisms. Relating these distributions to the differing amount of annotations known for many species, it seems reasonable that RNA classes are distributed heterogeneously. However, this is certainly also related to the fact that some organisms are more in the research focus than others. As the other animals in our study are not model organisms, it is possible that their unmapped reads belong to RNA families that have not yet been annotated in Rfam or otherwise present sequencing artefacts. Astonishingly, we found that the tRNA fraction was incredibly high in both lynx samples. As we found similar extreme distributions for both samples, this reduces the likelihood of sequencing or library preparation errors.

Therefore, we hypothesize that this could be related to the physiological or even pathophysiological condition of the Lynx that has not been diagnosed so far, especially since tRNA overexpression has often been associated with various cancer types in human (41–43). Interestingly, we found that miRNAs and YRNA levels were positively correlated, suggesting that even though their biogenesis pathways are different (44) they might share, potentially complementary, functions. We also found that the levels of snoRNAs and snRNAs correlated positively, which is not surprising, as they both belong to the upper class of small nuclear RNAs that guide RNA processing proteins.

The evaluation of the expression of sncRNAs in the context of their phylogeny highlighted that large differences that can be observed between some superorders, and in particular between Neognathae and Chelonia in comparison to the others of this study. We even found examples of potential precursors that showed preferential arm expressions depending on their superorders. Nevertheless, these findings are of course limited by the size of groups, and more samples would be needed for higher confidence. In particular, arm expression comparisons can be difficult, due to the fact that precursors containing the same or similar miRNAs do not necessarily exist in all species. Further evidence, in particular via genome assemblies, would help to reduce this limitation.

The recovery of deposited miRNA sequences from three miRNA databases highlighted that miRBase contains the highest number of unique sequences, but also include numerous redundant variations of sequences belonging to the same family. We showed that known miRNAs are available in more species than previously assumed and other ones might be expressed predominantly as different isoforms.

For the prediction of novel miRNAs from NGS data, we chose mirnovo (33) because this tool does not require a reference genome. To obtain an estimate of how well this prediction works, we counted how many known miRNA sequences can be recovered with the prediction. Although we used a very lenient mapping strategy, a median of only about 40.5 miRNAs were found per organism. In contrast, the tool predicted more than 10 times as many novel candidates per organism. By applying a filtering approach and thus reducing the predictions by 4-fold, we expect to have increased the ratio of true positives considerably. Because we cannot verify these results experimentally, it remains unclear how many true positive findings the predictions actually contain.

While our study describes expression patterns of sncRNAs in blood cells for a large collection of animals and provides fascinating new insights into the distribution and conservation of sncRNAs, certain limitations of the present study need to be considered and discussed. First, the samples were collected during veterinary examinations, including routine examinations but also blood collection of animals with pathologies. These factors might be reflected in the patterns of sncRNAs, but according to our experience from human samples, such effects are rather moderate compared with the variations that we observe here. A more important factor may be variations between representatives of the same species; we thus aim to obtain more specimens, in terms of collecting more samples from the same species but also adding more species. Another limitation stems from the focus of our study. We focus exclusively on circulating sncRNAs in blood cells and thus miss sncRNAs which might be specific to other cell types. In order to reach a comprehensive description of the sncRNAs present in the analysed species, more tissues and specimens will be needed.

CONCLUSION

The detection, annotation and validation of sncRNAs, especially miRNAs, is still a growing field. To understand their function and their potential as biomarkers for diseases, we must first understand how to distinguish actually expressed and valid miRNAs from false positive findings. Conservation is a widely applied feature for identifying miRNAs in related species. Such analyses are often only performed via homology- and sequence-based *in silico* approaches. With our study, we provide a large collection of small RNA NGS expression data for species that have not been analysed before in great detail. We created a comprehensive publicly available online resource for researchers in the field to facilitate the assessment of evolutionarily conserved small RNA sequences.

DATA AVAILABILITY

All sequencing data have been deposited in the Sequence Read Archive with the accession SRP162759.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Edanz (www.edanzediting.com) for editing the English text of a draft of this manuscript.

FUNDING

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

Conflict of interest statement. None declared.

REFERENCES

1. Crick,F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561–563.
2. Lee,R.C., Feinbaum,R.L. and Ambros,V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
3. Mercer,T.R., Dinger,M.E. and Mattick,J.S. (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, **10**, 155–159.
4. Veneziano,D., Di Bella,S., Nigita,G., Laganà,A., Ferro,A. and Croce,C.M. (2016) Noncoding RNA: Current deep sequencing data analysis approaches and challenges. *Human Mutat.*, **37**, 1283–1298.
5. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2018) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
6. 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.
7. Weber,M.J. (2005) New human and mouse microRNA genes found by homology search. *FEBS J.*, **272**, 59–73.

8. Yue,J., Sheng,Y. and Orwig,K.E. (2008) Identification of novel homologous microRNA genes in the rhesus macaque genome. *BMC Genomics*, **9**, 8.
9. Artzi,S., Kiezun,A. and Shomron,N. (2008) miRNAminer: a tool for homologous microRNA gene search. *BMC Bioinformatics*, **9**, 39.
10. Baev,V., Daskalova,E. and Minkov,I. (2009) Computational identification of novel microRNA homologs in the chimpanzee genome. *Comput. Biol. Chem.*, **33**, 62–70.
11. Long,J.-E. and Chen,H.-X. (2009) Identification and characteristics of cattle MicroRNAs by homology searching and small RNA cloning. *Biochem. Genet.*, **47**, 329–343.
12. Ludwig,N., Leidinger,P., Becker,K., Backes,C., Fehlmann,T., Pallasch,C., Rheinheimer,S., Meder,B., Stähler,C., Meese,E. *et al.* (2016) Distribution of miRNA expression across human tissues. *Nucleic Acids Res.*, **44**, 3865–3877.
13. Backes,C., Meese,E. and Keller,A. (2016) Specific miRNA disease biomarkers in blood, serum and plasma: challenges and prospects. *Mol. Diagn. Ther.*, **20**, 509–518.
14. Keller,A., Fehlmann,T., Ludwig,N., Kahraman,M., Laufer,T., Backes,C., Vogelmeier,C., Diener,C., Biertz,F., Herr,C. *et al.* (2018) Genome-wide MicroRNA expression profiles in COPD: Early predictors for cancer development. *Genomics Proteomics Bioinformatics*, **16**, 162–171.
15. 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 microRNAs using next-generation sequencing. *Alzheimers Demen.*, **12**, 565–576.
16. Fehlmann,T., Ludwig,N., Backes,C., Meese,E. and Keller,A. (2016) Distribution of microRNA biomarker candidates in solid tissues and body fluids. *RNA Biol.*, **13**, 1084–1088.
17. Keller,A., Leidinger,P., Bauer,A., Elsharawy,A., Haas,J., Backes,C., Wendschlag,A., Giese,N., Tjaden,C., Ott,K. *et al.* (2011) Toward the blood-borne miRNome of human diseases. *Nat. Methods*, **8**, 841–843.
18. Keller,A., Leidinger,P., Vogel,B., Backes,C., ElSharawy,A., Galata,V., Mueller,S.C., Marquart,S., Schrauder,M.G., Strick,R. *et al.* (2014) miRNAs can be generally associated with human pathologies as exemplified for miR-144. *BMC Med.*, **12**, 224.
19. Kahraman,M., Laufer,T., Backes,C., Schrörs,H., Fehlmann,T., Ludwig,N., Kohlhaas,J., Meese,E., Wehler,T., Bals,R. *et al.* (2017) Technical stability and biological variability in microRNAs from dried blood spots: a lung cancer therapy-monitoring showcase. *Clin. Chem.*, **63**, 1476–1488.
20. Pirritano,M., Fehlmann,T., Laufer,T., Ludwig,N., Gasparoni,G., Li,Y., Meese,E., Keller,A. and Simon,M. (2018) NGS analysis of total small non coding RNAs from low input RNA from dried blood sampling. *Anal. Chem.*, **90**, 11791–11796.
21. Fehlmann,T., Backes,C., Kahraman,M., Haas,J., Ludwig,N., Posch,A.E., Würstle,M.L., Hübenthal,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.
22. R Core Team (2018) *R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing Vienna*, Austria, https://www.r-project.org.
23. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* , **57**, 289–300.
24. Mangiafico,S. (2019) rcompanion: Functions to Support Extension Education Program Evaluation. *R package version 2.0.10* . https://rcompanion.org.
25. Wickham,H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, NY, https://ggplot2.tidyverse.org.
26. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
27. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
28. Schliep,K.P. (2011) phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**, 592.
29. Letunic,I. and Bork,P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.*, **44**, W242–W245.
30. Weese,D., Holtgrewe,M. and Reinert,K. (2012) RazerS 3: faster, fully sensitive read mapping. *Bioinformatics*, **28**, 2592–2599.
31. 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.
32. Rognes,T., Flouri,T., Nichols,B., Quince,C. and Mahé,F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.
33. Vitsios,D.M., Kentepozidou,E., Quintais,L., Benito-Gutiérrez,E., van Dongen,S., Davis,M.P. and Enright,A.J. (2017) Mirnovo: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. *Nucleic Acids Res.*, **45**, e177.
34. Backes,C., Meder,B., Hart,M., Ludwig,N., Leidinger,P., Vogel,B., Galata,V., Roth,P., Menegatti,J., Grässer,F. *et al.* (2015) Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic Acids Res.*, **44**, e53.
35. Alles,J., Fehlmann,T., Fischer,U., Backes,C., Galata,V., Minet,M., Hart,M., Abu-Halima,M., Grässer,F.A., Lenhof,H.-P. *et al.* (2019) An estimate of the total number of true human miRNAs. *Nucleic Acids Res.*, doi:10.1093/nar/gkz097.
36. Kalvari,I., Argasinska,J., Quinones-Olvera,N., Nawrocki,E.P., Rivas,E., Eddy,S.R., Bateman,A., Finn,R.D. and Petrov,A.I. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, **46**, D335–D342.
37. Fromm,B., Domanska,D., Hackenberg,M., Mathelier,A., Hoye,E., Johansen,M., Hovig,E., Flatmark,K. and Peterson,K.J. (2018) MirGeneDB2.0: the curated microRNA Gene Database. bioRxiv doi: https://doi.org/10.1101/258749, 05 February 2018, preprint: not peer reviewed.
38. Meder,B., Backes,C., Haas,J., Leidinger,P., Stähler,C., Großmann,T., Vogel,B., Frese,K., Giannitsis,E., Katus,H.A. *et al.* (2014) Influence of the confounding factors age and sex on microRNA profiles from peripheral blood. *Clin. Chem.*, **60**, 1200–1208.
39. Keller,A., Leidinger,P., Vogel,B., Backes,C., ElSharawy,A., Galata,V., Mueller,S.C., Marquart,S., Schrauder,M.G., Strick,R. *et al.* (2014) MiRNAs can be generally associated with human pathologies as exemplified for miR-144. *BMC Med.*, **12**, 224.
40. Muñoz-Culla,M., Irizar,H., Sáenz-Cuesta,M., Castillo-Triviño,T., Osorio-Querejeta,I., Sepúlveda,L., De Munain,A.L., Olascoaga,J. and Otaegui,D. (2016) SncRNA (microRNA & snoRNA) opposite expression pattern found in multiple sclerosis relapse and remission is sex dependent. *Sci. Rep.*, **6**, 20126.
41. Goodarzi,H., Nguyen,H.C., Zhang,S., Dill,B.D., Molina,H. and Tavazoie,S.F. (2016) Modulated expression of specific tRNAs drives gene expression and cancer progression. *Cell*, **165**, 1416–1427.
42. Huang,S.-q., Sun,B., Xiong,Z.-p., Shu,Y., Zhou,H.-h., Zhang,W., Xiong,J. and Li,Q. (2018) The dysregulation of tRNAs and tRNA derivatives in cancer. *J. Experiment. Clin. Cancer Res.*, **37**, 101.
43. Zhou,Y., Goodenbour,J.M., Godley,L.A., Wickrema,A. and Pan,T. (2009) High levels of tRNA abundance and alteration of tRNA charging by bortezomib in multiple myeloma. *Biochem. Biophys. Res. Commun.*, **385**, 160–164.
44. Nicolas,F.E., Hall,A.E., Csorba,T., Turnbull,C. and Dalmay,T. (2012) Biogenesis of Y RNA-derived small RNAs is independent of the microRNA pathway. *FEBS Lett.*, **586**, 1226–1230.

3.5 *Spring is in the air: seasonal profiles indicate vernal change of miRNA activity*

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

3.6 *Evaluating the use of circulating microRNA profiles for lung cancer detection in symptomatic patients*

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

3.7 *Deregulated microRNA and mRNA expression profiles in the peripheral blood of patients with Marfan syndrome*

Abu-Halima *et al. J Transl Med (2018) 16:60*
https://doi.org/10.1186/s12967-018-1429-3

Journal of
Translational Medicine

RESEARCH **Open Access**

Deregulated microRNA and mRNA expression profiles in the peripheral blood of patients with Marfan syndrome

Masood Abu-Halima[1*], Mustafa Kahraman[2], Dominic Henn[3], Tanja Rädle-Hurst[4], Andreas Keller[2], Hashim Abdul-Khaliq[4†] and Eckart Meese[1†]

Abstract

Background: MicroRNAs (miRNAs) are small RNAs regulating gene expression post-transcriptionally. While acquired changes of miRNA and mRNA profiles in cancer have been extensively studied, little is known about expression changes of circulating miRNAs and messenger RNAs (mRNA) in monogenic constitutional anomalies affecting several organ systems, like Marfan syndrome (MFS). We performed integrated miRNA and mRNA expression profiling in blood samples of Marfan patients in order to investigate deregulated miRNA and mRNA networks in these patients which could serve as potential diagnostic and prognostic tools for MFS therapy.

Methods: MiRNA and mRNA expression profiles were determined in blood samples from MFS patients ($n = 7$) and from healthy volunteer controls ($n = 7$) by microarray analysis. Enrichment analyses of altered mRNA expression were identified using bioinformatic tools.

Results: A total of 28 miRNAs and 32 mRNAs were found to be significantly altered in MFS patients compared to controls (> 2.0-fold change, adjusted $P < 0.05$). The expression of 11 miRNA and 6 mRNA candidates was validated by RT-qPCR in an independent cohort of 26 MFS patients and 26 matched HV controls. Significant inverse correlations were evident between 8 miRNAs and 5 mRNAs involved in vascular pathology, inflammation and telomerase regulation. Significant positive correlations were present for 7 miRNAs with age, for 2 miRNAs with the MFS aortic root status (Z-score) and for 7 miRNAs with left ventricular end-diastolic diameter in MFS patients. In addition, miR-331-3p was significantly up-regulated in MFS patients without mitral valve prolapse (MVP) as compared with patients with MVP.

Conclusions: Our data show deregulated gene and miRNA expression profiles in the peripheral blood of MFS patients, demonstrating several candidates for prognostic biomarkers for cardiovascular manifestations in MFS as well as targets for novel therapeutic approaches. A deregulation of miRNA expression seems to play an important role in MFS, highlighting the plethora of effects on post-transcriptional regulation of miRNAs and mRNAs initiated by constitutional mutations in single genes.

Trial registration Nr: EA2/131/10. Registered 28 December, 2010

Keywords: MicroRNA, mRNA, Integration analysis, Marfan syndrome, Fibrillin

---

*Correspondence: masood@daad-alumni.de
†Hashim Abdul-Khaliq and Eckart Meese contributed equally to this work
[1] Institute of Human Genetics, Saarland University, 66421 Homburg/Saar, Germany
Full list of author information is available at the end of the article

© The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

Marfan syndrome (MFS, OMIM #154700) is a connective tissue disorder with an estimated incidence of 1:5000 individuals, across all ethnic backgrounds [1, 2]. Although an autosomal dominant inheritance of MFS typically appears in affected multi-generation families, approximately 25% of cases have the disorder as the result of a de novo mutation [3]. The phenotypic variability of this disorder ranges from minor stigmata to life-threatening manifestations, typically involving the cardiovascular (thoracic aortic aneurysms (TAA) and dissections), ocular (ectopia lentis), and musculoskeletal system (tall stature with arachnodactyly) [4–6]. Cardiovascular manifestations are responsible for the high morbidity in individuals with MFS and can include dilation of the ascending aorta, pulmonary artery dilation, and mitral valve prolapse [6]. Mutations in Fibrillin-1 (FBN1) were identified as the cause of MFS [5]. As FBN1 is a constituent of the connective tissue in a wide range of organs, decreased mechanical stability caused by mutations in FBN1 has pleiotropic effects. Pleiotropy was introduced by Plate in 1910 to describe multiple phenotypic effects of a single genetic trait [7]. Although there are 3077 known mutations in the FBN1 gene (UMD-FBN1 listed in the database: http://www.umd.be/FBN1/, updated on August 28, 2014) and more than 1500 different disease-causing FBN1 mutations, there is no single FBN1 genotype feature that qualifies as a reliable predictor of the clinical severity and long-term course of MFS. Even within a given family with an identical FBN1 mutation, there is considerable variation as to the severity of manifestation, pointing towards complex interactions of FBN1 with other genes and their products [4, 6, 8]. Currently, methods for predicting the prognosis of Marfan-related cardiovascular manifestations are limited. However, in several pathologies, microRNAs (miRNAs) have emerged in recent years as a promising new type of biomarker for the prognosis of disease, including initial data on MFS and aortic disease [9, 10]. MiRNAs are a class of non-coding RNAs of 18–22 nucleotides in length that regulate gene expression post-transcriptionally via sequence-specific interaction with the 3′ untranslated region (UTR) of a target gene's mRNA, resulting in inhibition of translation and/or mRNA degradation [7]. Altered expression of miRNA has been associated with many human diseases, including MFS [9, 11]. Recently, it was reported that miR-29b is associated with vascular remodeling and subsequent aneurysm development characteristic of MFS and that this miRNA plays an important role in regulating aortic wall apoptosis and extracellular matrix abnormalities in MFS [11]. In addition to miRNA expression analysis, genome-wide mRNA expression analyses of skin fibroblast cultures from individuals with known FBN1 mutations and controls has been performed [12]. In tissue of MFS patients, however, investigations of miRNAs as well as mRNAs are still lacking. Thus, it is conceivable that in addition to an entire miRNome expression profiling, the search for miRNAs whose expression inversely correlates with the expression of mRNA targets may demonstrate another layer of the molecular diversity of this pleiotropic syndrome and may potentially be a useful diagnostic and prognostic tool for MFS therapy and treatment. A crucial clinical challenge are still insufficient indication criteria for preventive aortic replacement that call for biological parameters beyond the current restriction to ultrasonographic and magnetic resonance imaging (MRI) measurements. Therefore, we investigated differences in miRNA and mRNA expression patterns between MFS patients and healthy volunteer [13] controls. We furthermore performed an integrated analysis across all samples to identify mRNA targets of deregulated miRNAs. To our knowledge, this is the first large-scale investigation of the association between miRNA-related mRNAs in patients with MFS.

Methods

Patient samples

The study was conducted in accordance with the Declaration of Helsinki and approved by the locally appointed Ethics committee [Institutional Review Board (Number: EA2/131/10)]. Informed consent was obtained from all patients and HV controls. A cohort of 34 patients in whom the clinical diagnosis of classical MFS was made according to the current Ghent nosology [6] was assessed for the ocular, musculoskeletal, and cardiovascular features by an ophthalmologist, a pediatrician, a cardiologist, and a clinical geneticist. Two-dimensional echocardiography was used to measure the diameter of the ascending aorta which was used to determine the patients' Z-score. Moreover the left ventricular end-diastolic diameter (LVEDD) and the presence of a mitral valve prolapse (MVP) were assessed by echocardiography. The patient cohort included 15 males and 19 females with a mean age of 27.62 years (standard deviation ± 15.66 years) and confirmed FBN1 mutation which were compared with age- and sex matched HV controls ($n = 34$). All Marfan patients were on Angiotensin receptor blockers. Beta blockers or ACE inhibitors had been added to the medication depending on the level of arterial hypertension or the presence of other cardiovascular morbidities. In all HV controls, a physical examination including measurement of blood pressure and transcutaneous oxygen saturation as well as two-dimensional echocardiography was performed to rule out any confounding cardiac and extracardiac abnormalities. At the time of enrolment, none of the controls took any medication or had elevated blood pressure. Additionally, none of them had any heart abnormality on the echocardiogram. In all patients and HV controls 2.5 mL of venous blood from the cubital vein was collected in PAXgene blood tubes (BD Biosciences, San Jose, California, United States). All PAXgene blood tubes were stored at room temperature for 2 h to ensure complete lysis of the blood cells before they were stored at − 20 °C until RNA isolation. MiRNA raw data were acquired from samples which had previously been used for a related study published by our group [9].

RNA isolation

Total RNA including miRNAs of all MFS patients and HV controls was isolated with the PAXgene miRNA blood kit using the QIAcube™ automated isolation instrument according to the manufacturer's instructions (Qiagen, Hilden, Germany). The RNA concentration and purity were confirmed by the spectrophotometric ratio using absorbance measurements at wavelengths of 260 and 280 nm on a NanoDrop-2000 (Thermo Scientific, Waltham, Massachusetts, United States). The integrity of the isolated RNA was analyzed on a RNA Nano 6000 chip using an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, California, United States). Genomic DNA contamination was removed, and conventional polymerase chain reaction (PCR) was carried out to exclude any residual DNA in the samples as previously described [14]. Moreover, the RNA-based RT-qPCRs were carried out using predesigned exon spanning primers (Qiagen).

Gene expression microarray assay analysis

MRNA expression profiles of MFS (n = 8) and HV controls (n = 8) samples were performed with SurePrint G3 Human Gene Expression v2 8x60K microarrays containing 50,599 biological features (Agilent Technologies, Santa Clara, CA, United States). All procedures were carried out according to the manufacturer's protocol. Briefly, 100 ng total RNA from each sample was reversely transcribed, amplified and labeled using the LowInput QuickAmp Labeling Kit (Agilent). Quantification and specific activity of labeled complementary RNA (cRNA) was evaluated using the NanoDrop-2000 spectrophotometer (Thermo Scientific) to ensure that labeled cRNA was of sufficient quality for hybridization. A total of 600 ng of cRNA was then applied to the microarray slide per the manufacturer's instructions and hybridized in a rotating oven for 17 h at 65 °C and 10 rpm. Arrays were washed and then scanned using a DNA Microarray Scanner (Agilent). Feature extraction software was utilized to extract gene expression data (Agilent).

MiRNA microarray assay analysis

We used the Sureprint G3 Human miRNA 8x60K microarrays raw data of 8 MFS and 8 HV controls [9]. MiRNA expression profiles were performed with Sureprint G3 Human miRNA 8x60K microarrays containing 40 replicates of 1205 miRNAs of miRBase v16 (Agilent). All procedures were carried out according to the manufacturer's protocol. Briefly, 100 ng total RNA from each sample was processed using the miRNA Complete Labeling and Hyb Kit (Agilent) to generate fluorescently labeled miRNA. The labeled RNA was then applied to the microarray slide per the manufacturer's instructions and hybridized in a rotating oven for 20 h at 55 °C and 20 rpm. Arrays were washed and then scanned using a DNA Microarray Scanner (Agilent). Feature extraction software was utilized to extract gene expression data (Agilent).

Reverse transcription and quantitative real-time PCR

Expression of selected mRNAs and miRNAs in MFS and HV controls was determined by real-time quantitative PCR (RT-qPCR) using the StepOnePlus™ Real-Time PCR System (Applied Biosystems, Foster City, CA, United States) and the miScript PCR System that contain *mi*Script RT II Kit with 5× miScript HiFlex Buffer and SYBR Green PCR along with the QuantiTect and miScript Primer Assays (Qiagen). All procedures were carried out according to the manufacturer's recommendations. Using a cohort of independent MFS patients (n = 26) and HV controls (n = 26), 13 differentially expressed mRNAs (CLU, CRYAA, CTNNA1, DYSF, GBP2, ITGB3, LIMK2, MFN2, MMP9, MX1, SIRPB1, POT1 and SOCS3) and 18 differentially expressed miRNAs (miR-1228, miR-1234-3p, miR-1275, miR-139-3p, miR-151-5p, miR-200c, miR-24, miR-30e, miR-324-5p, miR-940, miR-3616-3p, miR-362-5p, miR-500b, miR-502-3p, miR-564, miR-627, miR-874 and miR-331-3p) were selected to validate the array results. In brief, 400 ng of total RNA were converted into complementary DNA (cDNA). The resulting cDNA was then diluted to have 1.5 and 0.5 ng/μL for mRNA and miRNA, respectively. All RT-qPCR experiments were performed using the Liquid Handling Robot QIAgility™ (Qiagen) before performing RT-qPCR. *β*-Actin and RNU6B small nuclear RNA (snRNA) primer assays (Qiagen) were chosen as endogenous references for mRNA and miRNA normalization. Moreover, a no template control (NTC) and no reverse transcriptase control (RT negative) were included in each mRNA and miRNA in each run, and a miRNA reverse transcription control (miRTC) was performed to assess the performance of the reverse transcription reaction by detecting template synthesized from the kit's built-in control RNA

(Qiagen). Melting curve analysis was used to control for the specificity of RT-qPCR products.

Overrepresentation analysis

To evaluate the significance of the identified differentially expressed genes, the Protein ANalysis THrough Evolutionary Relationships (PANTHER) Classification System was used to categorize the differentially expressed genes according to PANTHER protein class, Gene Ontology (GO) Molecular Function, GO Biological Process and GO cellular components annotations [15]. For each biological pathway and/or process, the difference between the observed fraction of genes in that pathway and/or process and the number expected by chance was tested using Fisher's exact test. *P* values were adjusted using a Bonferroni correction.

Statistical analysis

The statistical analysis was performed using R (version 3.4.0) to analyze the differences in mRNA and miRNA expression patterns in the MFS patients and HV controls. Raw data generated by the Agilent Feature Extraction image analysis software was normalized by variance stabilizing normalization (vsn) [16] and quantile normalization methods for mRNAs and miRNAs, respectively and uploaded to the NCBI GEO database (Accession ID: GSE110966). The significance level of mRNAs and miRNAs was determined by applying an unpaired two-tailed t test. Then the median values of each miRNA and mRNA were log2 transformed and the resulting miRNA P values were adjusted for multiple testing using Benjamini–Hochberg adjustment. In addition, the area under the receiver operating characteristic curve values for each miRNA were computed. For the significantly deregulated miRNAs and protein coding genes with $P<0.05$ and fold change > 2 or < 1/2 in MFS patients compared to HV controls, we computed a Pearson correlation coefficient of expression for each mRNA–miRNA pair. Spearman's correlations coefficient was used to correlate the clinical parameters of MFS and the expression level of both validated miRNAs and mRNAs. Using the DataAssist™ Software v3.0 (Applied Biosystems), the fold-change and *P* value (unpaired t test with Welch's correction) of each mRNA and miRNA was calculated.

Results

Patient characteristics

Among 19 females and 15 males included in the study, there were 22 patients with MVP, 6 patients with ectopia lentis and 14 patients who underwent aortic root replacement because of aortic dissection or an aortic aneurysm (aortic root > 53 mm). Additional file 3: Table S1 shows the clinical features of the MFS patients. The presented data refer to the largest diameters of the aortic root before surgery.

Correlation analysis of miRNA and mRNA between MFS patients and HV controls

As an initial analysis, we calculated the degree of correlation based on Pearson's correlation coefficient across samples from each group, i.e., MFS patients and HV controls. The correlation plots are presented in Additional file 1: Figure S1 for miRNA and Additional file 2: Figure S2 for mRNA. In general, the correlation heatmaps illustrated that the correlation was strong (Pearson correlation coefficient r of mostly > 0.80 and > 0.92 for miRNA and mRNA, respectively) in both MFS patients and HV controls, except for two samples which we identified as outliers in the mRNA data by applying Hampel's rule for outlier detection [17]. These two samples and the corresponding miRNA samples of the same MFS patients and HV controls were excluded from further analyses.

Differentially expressed miRNAs between MFS patients and HV controls

Using the high-throughput SurePrint G3 Human v16 miRNA microarray platform, we screened the expression level of 1205 human mature miRNAs of miRBase v16. Following background correction and normalization, the miRNA expression levels from MFS patients and HV controls were identified. After excluding outliers, 7 MFS patients and 7 HV controls were considered for further analysis. Using quantile normalization, a total of 277 miRNAs were detected in at least 25% of the samples in at least one group (filtering). By applying an un-paired two-tailed t test for miRNAs that showed a significant change in the considered groups, 63 miRNAs showed statistical significance in MFS patients versus HV controls (Table 1) ($P<0.05$. FDR adjusted). By considering only the differentially expressed miRNAs with twofold or greater change in MFS patients versus HV controls, a total of 28 miRNAs were identified including 15 down-regulated and 13 up-regulated miRNAs (*P* value < 0.05, fold change ≥ 2.0) (Table 1). To compare the relative expression level of the differentially expressed miRNAs, we used the hierarchical clustering of miRNAs and hierarchical clustering of samples based on average linkage and Euclidian distance of the significantly deregulated 63 miRNAs out of 1205 miRNAs in MFS patients versus HV controls (Fig. 1). In general, hierarchical clustering revealed that MFS patients and HV controls were grouped into two distinct clusters, except for only one MFS (fell into the wrong cluster). Moreover, the heatmap showed that some miRNAs were expressed only in the MFS patients group and/or expressed at a low level in HV controls and vice versa (Fig. 1).

Table 1 Significantly expressed miRNAs in the blood of patients with MFS (n = 7) compared HVs controls (n = 7) as determined by microarray (unpaired two-tailed t test. P < 0.05, FDR adjusted)

miRNA	Median log2 MFS	Median log2 HVS	Fold change	Log2 fold change	Regulation	P value	Adjusted P value	AUC
hsa-miR-4271	− 3.32	4.40	0.0047	− 7.72	Down	0.0037	0.0689	0.95
hsa-miR-3616-3p	− 3.32	4.15	0.0056	− 7.47	Down	0.0003	0.0149	0.96
hsa-miR-1228	− 3.32	3.80	0.0072	− 7.12	Down	0.0004	0.0149	0.98
hsa-miR-1238	− 3.32	3.17	0.0111	− 6.49	Down	0.0169	0.1552	0.84
hsa-miR-191	− 3.32	3.03	0.0123	− 6.35	Down	0.0450	0.2177	0.80
hsa-miR-1234	0.94	4.02	0.12	− 3.09	Down	0.0004	0.0149	0.98
hsa-miR-4313	0.69	3.39	0.15	− 2.70	Down	0.0468	0.2177	0.78
hsa-miR-139-3p	0.65	3.15	0.18	− 2.50	Down	0.0001	0.0149	1.00
hsa-miR-874	2.83	4.76	0.26	− 1.92	Down	0.0042	0.0731	0.92
hsa-miR-564	4.70	5.90	0.43	− 1.20	Down	0.0003	0.0149	1.00
hsa-miR-940	3.06	4.21	0.45	− 1.15	Down	0.0004	0.0149	1.00
hsa-miR-1207-5p	5.80	6.93	0.46	− 1.13	Down	0.0319	0.1878	0.83
hsa-miR-1280	1.64	2.73	0.47	− 1.10	Down	0.0048	0.0790	0.98
hsa-miR-181d	3.93	4.99	0.48	− 1.06	Down	0.0277	0.1845	0.80
hsa-miR-1275	3.77	4.80	0.49	− 1.03	Down	0.0014	0.0323	0.94
hsa-miR-3653	7.04	8.01	0.51	− 0.97	Down	0.0484	0.2177	0.85
hsa-miR-1268	5.18	6.07	0.54	− 0.88	Down	0.0239	0.1739	0.89
hsa-miR-320b	10.66	11.48	0.57	− 0.82	Down	0.0085	0.1242	0.92
hsa-miR-532-3p	8.69	9.46	0.59	− 0.77	Down	0.0130	0.1444	0.84
hsa-miR-642b	3.17	3.93	0.59	− 0.76	Down	0.0179	0.1552	0.87
hsa-miR-4323	2.73	3.44	0.61	− 0.70	Down	0.0218	0.1675	0.87
hsa-miR-93	5.67	6.31	0.64	− 0.64	Down	0.0297	0.1871	0.80
hsa-miR-3162	6.83	7.46	0.65	− 0.63	Down	0.0068	0.1050	0.89
hsa-miR-3679-5p	4.53	5.13	0.66	− 0.59	Down	0.0278	0.1845	0.87
hsa-miR-423-5p	9.11	9.62	0.70	− 0.51	Down	0.0124	0.1433	0.86
hsa-miR-1225-5p	6.16	6.67	0.70	− 0.51	Down	0.0287	0.1852	0.84
hsa-miR-3651	4.99	5.44	0.73	− 0.45	Down	0.0478	0.2177	0.81
hsa-miR-638	5.98	6.41	0.74	− 0.43	Down	0.0193	0.1575	0.86
hsa-miR-766	4.84	5.25	0.76	− 0.40	Down	0.0238	0.1739	0.91
hsa-miR-191	5.10	5.49	0.76	− 0.39	Down	0.0350	0.1978	0.83
hsa-miR-762	4.55	4.73	0.88	− 0.18	Down	0.0403	0.2106	0.78
hsa-miR-221	2.77	− 3.32	68.19	6.09	Up	0.0499	0.2177	0.18
hsa-miR-1288	2.53	− 3.32	57.95	5.86	Up	0.0344	0.1978	0.21
hsa-miR-3125	2.35	− 3.32	50.81	5.67	Up	0.0481	0.2177	0.21
hsa-miR-500b	2.27	− 3.32	48.35	5.60	Up	0.0142	0.1462	0.14
hsa-miR-200c	2.13	− 3.32	43.66	5.45	Up	0.0142	0.1462	0.14
hsa-miR-3200-3p	2.07	− 3.32	41.98	5.39	Up	0.0280	0.1845	0.17
hsa-miR-3667-5p	1.81	− 3.32	35.05	5.13	Up	0.0380	0.2067	0.14
hsa-miR-627	1.74	− 3.32	33.47	5.06	Up	0.0311	0.1878	0.14
hsa-miR-664	1.33	− 3.32	25.09	4.65	Up	0.0395	0.2106	0.20
hsa-miR-223	13.58	11.98	3.04	1.60	Up	0.0496	0.2177	0.15
hsa-miR-660	4.28	2.92	2.57	1.36	Up	0.0192	0.1575	0.08
hsa-miR-29c	4.59	3.33	2.41	1.27	Up	0.0177	0.1552	0.00
hsa-miR-7	4.62	3.56	2.09	1.06	Up	0.0357	0.1978	0.12
hsa-miR-29a	6.93	5.98	1.93	0.95	Up	0.0117	0.1407	0.04
hsa-miR-500a	5.70	4.80	1.87	0.90	Up	0.0008	0.0202	0.04
hsa-miR-23a	9.46	8.57	1.85	0.89	Up	0.0103	0.1364	0.06
hsa-miR-151-5p	9.11	8.33	1.73	0.79	Up	0.0175	0.1552	0.09
hsa-miR-324-5p	6.20	5.41	1.72	0.79	Up	0.0005	0.0149	0.05

Table 1 (continued)

miRNA	Median log2 MFS	Median log2 HVS	Fold change	Log2 fold change	Regulation	P value	Adjusted P value	AUC
hsa-miR-4306	12.57	11.84	1.66	0.73	Up	0.0161	0.1552	0.08
hsa-miR-186	6.78	6.10	1.60	0.68	Up	0.0261	0.1845	0.12
hsa-miR-502-3p	5.53	4.87	1.58	0.66	Up	0.0005	0.0149	0.00
hsa-miR-23b	6.34	5.70	1.56	0.64	Up	0.0468	0.2177	0.18
hsa-miR-629	4.76	4.12	1.56	0.64	Up	0.0466	0.2177	0.20
hsa-miR-362-5p	6.34	5.74	1.52	0.60	Up	0.0005	0.0149	0.02
hsa-miR-652	9.07	8.47	1.51	0.60	Up	0.0116	0.1407	0.16
hsa-miR-24	7.95	7.38	1.48	0.57	Up	0.0018	0.0392	0.01
hsa-miR-501-3p	4.31	3.74	1.48	0.56	Up	0.0099	0.1364	0.04
hsa-miR-30e	5.97	5.49	1.39	0.48	Up	0.0027	0.0542	0.08
hsa-miR-331-3p	10.09	9.62	1.38	0.46	Up	0.0205	0.1621	0.16
hsa-miR-451	16.51	16.05	1.38	0.46	Up	0.0004	0.0149	0.06
hsa-miR-532-5p	5.94	5.53	1.33	0.41	Up	0.0318	0.1878	0.13
hsa-miR-103	10.66	10.25	1.33	0.41	Up	0.0478	0.2177	0.18

Each value represents the median of 7 MFS patients and 7 HV controls and ± standard deviation (STDV). Statistical analysis was performed with unpaired-two-tailed t test ($P < 0.05$). MFS Marfan syndrome, HVs healthy volunteers; AUC area under the receiver operating characteristic curve

Differentially expressed genes between MFS patients and HV controls

Using the high-throughput SurePrint G3 Human Gene Expression v2 microarray platform, we screened the expression level of human 50,599 biological features of ENSEMBL release 52. Following background correction and normalization, the gene expression levels from MFS patients and HV controls were identified. Using variance stabilizing normalization (vsn) to the generated gene expression data, 58,717 transcripts were detected (no filtering). By applying an un-paired two-tailed t test for the transcripts that showed a significant change in the considered groups, we found 1662 transcripts with significant differences of MFS patients versus HV controls ($P < 0.05$) (Additional file 3: Table S2). By considering only the differentially expressed transcripts with 1.5-fold or greater change in MFS patients versus HV controls, a total of 505 transcripts were identified including 15 down-regulated and 490 up-regulated transcripts (P value < 0.05, fold change ≥ 1.5). Considering only the protein coding genes and removing different transcript variants, 296 genes out of 505 transcripts were identified, including 5 down-regulated and 291 up-regulated protein coding genes (Additional file 3: Table S3). The number of significant protein coding genes with twofold or greater change in MFS patients versus HV controls was decreased including one down-regulated and 31 up-regulated genes (P value < 0.05) (Table 2). Using hierarchical clustering with the Euclidian distance measure, we analyzed how the MFS patients and HV controls relate to each other. For this task, we used the 65 transcripts with the highest expression variances out of the 50,599 biological features. Figure 2 shows the resulting heatmap of the hierarchical clustering. In general, we observed two distinct clusters without overlap, with the first cluster containing only HV controls and the second cluster containing only MFS patients.

Validation of candidate miRNAs and mRNAs by RT-qPCR

In order to validate the results obtained from the microarray analysis, RT-qPCR was performed using a larger separate cohort of MFS patients and HV controls (MFS patients, n = 26 and HV controls, n = 26). Eighteen miRNAs were selected based on their differential expression level in MFS patients versus HV controls, and some of them were selected based on their known associations with cardiovascular diseases and MFS (listed in Table 1). In addition, 13 mRNAs were selected based on their known associations with cardiovascular diseases like mitral valve stenosis, myocardial infarction, ischemia and

(See figure on next page.)
Fig. 1 Unsupervised hierarchical clustering (Euclidian distance, complete linkage) of the 14 samples based on expression of the 63 with significant highest variance out of the 1205 miRNAs. The heatmap shows miRNAs with high expression in red, miRNAs with low expression in green. The red lines indicate three main clusters of samples

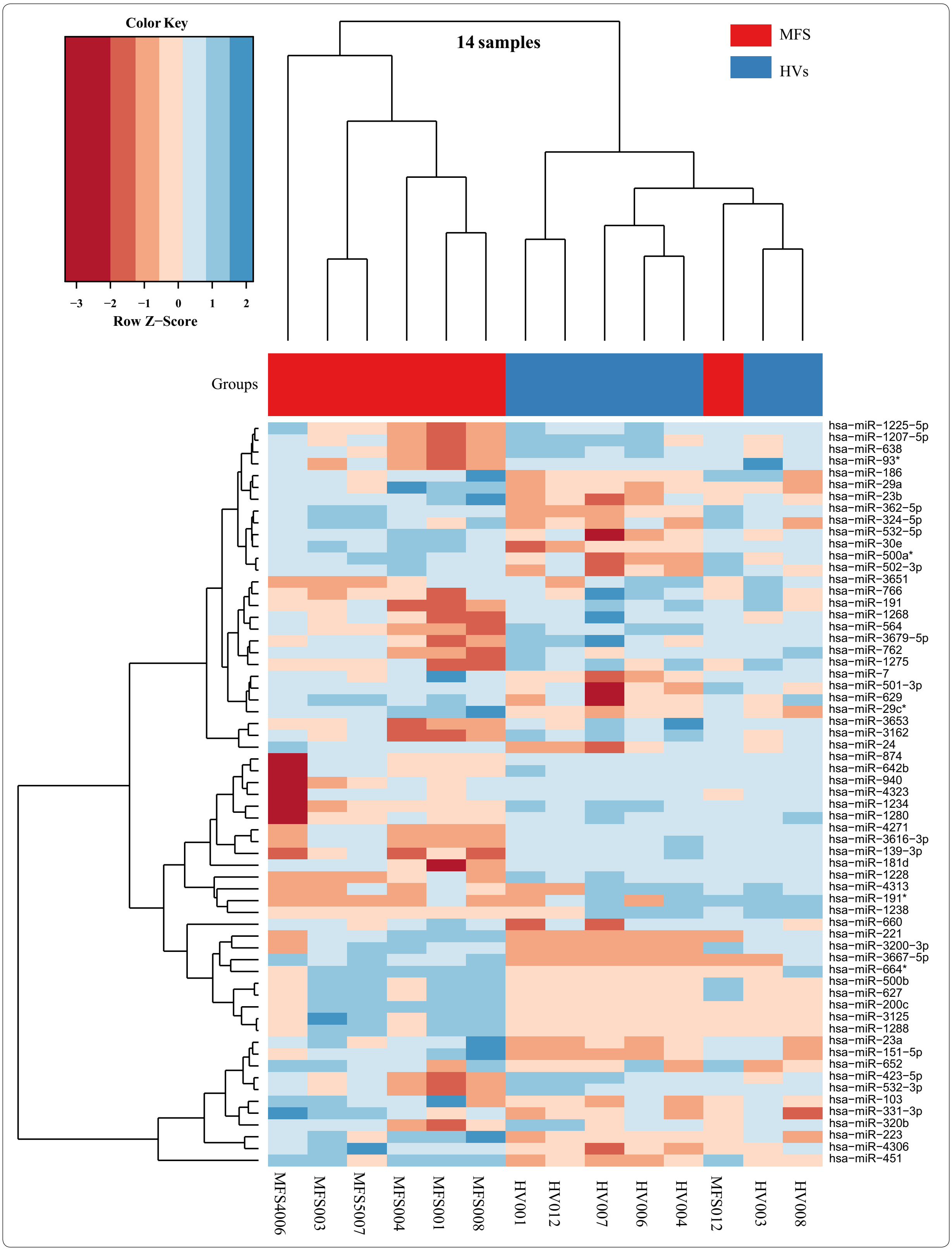

Color Key
−3
−2
−1
0
1
2
Row Z−Score
14 samples
MFS
HVs
Groups
hsa−miR−1225−5p
hsa−miR−1207−5p
hsa−miR−638
hsa−miR−93*
hsa−miR−186
hsa−miR−29a
hsa−miR−23b
hsa−miR−362−5p
hsa−miR−324−5p
hsa−miR−532−5p
hsa−miR−30e
hsa−miR−500a*
hsa−miR−502−3p
hsa−miR−3651
hsa−miR−766
hsa−miR−191
hsa−miR−1268
hsa−miR−564
hsa−miR−3679−5p
hsa−miR−762
hsa−miR−1275
hsa−miR−7
hsa−miR−501−3p
hsa−miR−629
hsa−miR−29c*
hsa−miR−3653
hsa−miR−3162
hsa−miR−24
hsa−miR−874
hsa−miR−642b
hsa−miR−940
hsa−miR−4323
hsa−miR−1234
hsa−miR−1280
hsa−miR−4271
hsa−miR−3616−3p
hsa−miR−139−3p
hsa−miR−181d
hsa−miR−1228
hsa−miR−4313
hsa−miR−191*
hsa−miR−1238
hsa−miR−660
hsa−miR−221
hsa−miR−3200−3p
hsa−miR−3667−5p
hsa−miR−664*
hsa−miR−500b
hsa−miR−627
hsa−miR−200c
hsa−miR−3125
hsa−miR−1288
hsa−miR−23a
hsa−miR−151−5p
hsa−miR−652
hsa−miR−423−5p
hsa−miR−532−3p
hsa−miR−103
hsa−miR−331−3p
hsa−miR−320b
hsa−miR−223
hsa−miR−4306
hsa−miR−451
MFS4006
MFS003
MFS5007
MFS004
MFS001
MFS008
HV001
HV012
HV007
HV006
HV004
MFS012
HV003
HV008

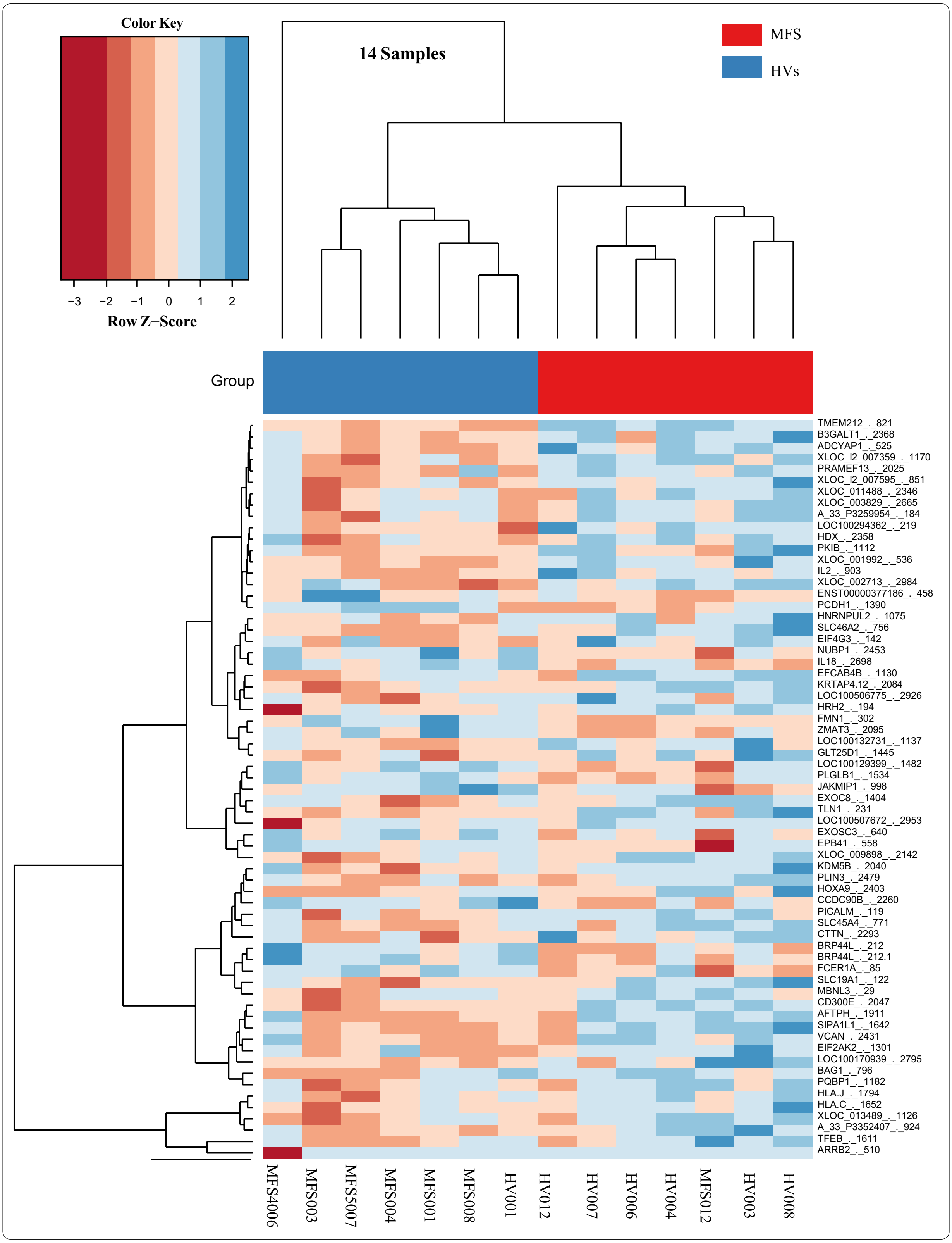
Table 2 Significantly expressed protein coding genes in the blood of patients with MFS ($n = 7$) compared HVs controls ($n = 7$) as determined by microarray (unpaired two-tailed t test. > 2.0-fold difference. $P < 0.05$)

Gene Name	NCBI accession code	Median log2 MFS	Median log2 HVS	Fold change	log2 Fold change	Regulation	P value	AUC
POT1 (protection of telomeres 1)	NM_015450	7.96	9.05	0.47	−1.09	Down	0.0002	1.00
SIRPB1 (signal regulatory protein beta 1)	NM_001135844	11.30	8.96	5.08	2.34	Up	0.0474	0.16
HBZ (hemoglobin subunit zeta)	NM_005332	8.71	6.47	4.73	2.24	Up	0.0132	0.16
MYOM2 (myomesin 2)	NM_003970	8.38	6.50	3.67	1.87	Up	0.0372	0.14
ITGB3 (integrin subunit beta 3)	NM_000212	9.23	7.38	3.62	1.86	Up	0.0392	0.08
MX1 (MX dynamin like GTPase 1)	NM_002462	12.89	11.36	2.89	1.53	Up	0.0383	0.16
DYSF (dysferlin)	NM_003494	10.49	9.04	2.74	1.45	Up	0.0149	0.08
CCR1 (C-C motif chemokine receptor 1)	NM_001295	12.91	11.50	2.66	1.41	Up	0.0217	0.12
IFIT2 (interferon induced protein with tetratricopeptide repeats 2)	NM_001547	11.56	11.33	1.17	0.23	Up	0.0317	0.14
LRG1 (leucine rich alpha-2-glycoprotein 1)	NM_052972	11.14	9.85	2.45	1.29	Up	0.0181	0.16
TRANK1 (tetratricopeptide repeat and ankyrin repeat containing 1)	NM_014831	10.86	9.62	2.37	1.25	Up	0.0013	0.02
ZAN (zonadhesin (gene/pseudogene))	NM_173059	9.30	8.08	2.34	1.23	Up	0.0423	0.18
CRYAA (crystallin alpha A)	NM_000394	9.30	8.08	2.34	1.23	Up	0.0329	0.18
NRGN (neurogranin)	NM_006176	12.69	11.52	2.25	1.17	Up	0.0356	0.12
RNF213 (ring finger protein 213)	NM_020914	8.24	7.70	1.46	0.55	Up	0.0377	0.18
LIMK2 (LIM domain kinase 2)	NM_001031801	9.31	8.16	2.22	1.15	Up	0.0181	0.14
MFN2 (mitofusin 2)	NM_014874	5.49	5.18	1.24	0.31	Up	0.0288	0.20
CTTN (cortactin)	NM_005231	9.51	8.36	2.21	1.14	Up	0.0268	0.16
MX2 (MX dynamin like GTPase 2)	NM_002463	12.60	11.52	2.11	1.07	Up	0.0157	0.12
CD14 (CD14 molecule)	NM_001174104	15.04	13.97	2.10	1.07	Up	0.0332	0.16
GRN (granulin precursor)	NM_002087	15.83	14.76	2.10	1.07	Up	0.0258	0.16
RNF222 (ring finger protein 222)	NM_001146684	5.29	5.01	1.22	0.28	Up	0.0336	0.14
MVP (major vault protein)	NM_017458	14.55	13.49	2.08	1.06	Up	0.0046	0.08
CXCL5 (C-X-C motif chemokine ligand 5)	NM_002994	8.88	7.83	2.07	1.05	Up	0.0455	0.18
CTNNA1 (catenin alpha 1)	NM_001903	9.08	8.05	2.05	1.04	Up	0.0309	0.16
MMP9 (matrix metallopeptidase 9)	NM_004994	13.20	12.14	2.07	1.05	Up	0.0456	0.20
SOCS3 (suppressor of cytokine signaling 3)	NM_003955	9.47	8.43	2.05	1.03	Up	0.0472	0.20
GBP2 (guanylate binding protein 2)	NM_004120	11.49	10.46	2.04	1.03	Up	0.0097	0.08
GNL3L (G protein nucleolar 3 like)	NM_019067	8.33	7.93	1.32	0.41	Up	0.0325	0.20
CLU (clusterin)	NM_001831	9.71	8.69	2.02	1.02	Up	0.0423	0.22
SELL (selectin L)	NM_000655	14.93	13.93	2.00	1.00	Up	0.0011	0.04
OSM (oncostatin M)	NM_020530	9.18	8.19	2.00	1.00	Up	0.0414	0.14

Each value represents the median of 7 MFS patients and 7 HV controls and ± standard deviation (STDV). Statistical analysis was performed with unpaired-two-tailed t test ($P < 0.05$). *MFS* Marfan syndrome, *HVs* healthy volunteers; AUC area under the receiver operating characteristic curve

(See figure on next page.)

Fig. 2 Unsupervised hierarchical clustering (Euclidian distance, complete linkage) of the 14 samples based on expression of the 65 transcripts with the highest expression variances out of the 50,599 biological features. The heatmap shows transcripts with high expression in red, transcript with low expression in green. The red lines indicate three main clusters of samples


Color Key
−3
−2
−1
0
1
2
Row Z−Score
14 Samples
MFS
HVs
Group
TMEM212_._821
B3GALT1_._2368
ADCYAP1_._525
XLOC_l2_007359_._1170
PRAMEF13_._2025
XLOC_l2_007595_._851
XLOC_011488_._2346
XLOC_003829_._2665
A_33_P3259954_._184
LOC100294362_._219
HDX_._2358
PKIB_._1112
XLOC_001992_._536
IL2_._903
XLOC_002713_._2984
ENST00000377186_._458
PCDH1_._1390
HNRNPUL2_._1075
SLC46A2_._756
EIF4G3_._142
NUBP1_._2453
IL18_._2698
EFCAB4B_._1130
KRTAP4.12_._2084
LOC100506775_._2926
HRH2_._194
FMN1_._302
ZMAT3_._2095
LOC100132731_._1137
GLT25D1_._1445
LOC100129399_._1482
PLGLB1_._1534
JAKMIP1_._998
EXOC8_._1404
TLN1_._231
LOC100507672_._2953
EXOSC3_._640
EPB41_._558
XLOC_009898_._2142
KDM5B_._2040
PLIN3_._2479
HOXA9_._2403
CCDC90B_._2260
PICALM_._119
SLC45A4_._771
CTTN_._2293
BRP44L_._212
BRP44L_._212.1
FCER1A_._85
SLC19A1_._122
MBNL3_._29
CD300E_._2047
AFTPH_._1911
SIPA1L1_._1642
VCAN_._2431
EIF2AK2_._1301
LOC100170939_._2795
BAG1_._796
PQBP1_._1182
HLA.J_._1794
HLA.C_._1652
XLOC_013489_._1126
A_33_P3352407_._924
TFEB_._1611
ARRB2_._510
MFS4006
MFS003
MFS5007
MFS004
MFS001
MFS008
HV001
HV012
HV007
HV006
HV004
MFS012
HV003
HV008

acute coronary syndrome, and some had been observed to be differentially expressed with twofold or greater change in MFS patients versus HV controls in the microarray data (listed in Table 2). Considering the direction of expression changes, there was a general accordance between microarray and RT-qPCR data for the miRNAs and mRNAs tested. In detail, RT-qPCR validated the results of the microarray analysis for 11 out of 18 miRNAs with regards both to the direction of expression changes and to the significance of different expressions between MFS patients and HV controls, including one significantly down-regulated miRNA (miR-1234-3p) and 10 significantly up-regulated miRNAs (miR-151-5p, miR-200c, miR-24, miR-30e, miR-324-5p, miR-362-5p, miR-500b, miR-502-3p, miR-627, and miR-331-3p) (Fig. 3). Likewise for the mRNA analysis, 11 out of 13 mRNAs showed the same direction of expression changes in the RT-qPCR and in the microarray analysis. Of these 11 miRNAs, 6 mRNAs were validated with regard to both the direction of expression changes and to the significance of different expressions between MFS patients and HV controls including 5 significantly up-regulated mRNAs (DYSF, GBP2, LIMK2, MMP9, and MX1) and one significantly down-regulated mRNA (POT1) (Fig. 4).

Inverse correlation between miRNA and target mRNA

To further understand the relationship between miRNA and mRNA changes, and to specifically identify potentially relevant miRNA–mRNA target interactions, we calculated the Pearson correlation coefficient for every stably expressed miRNA and every protein coding gene that was significantly deregulated with a fold change of < 0.5/> 1.5. This computation yielded 11 significant negative combinations ($P < 0.05$) with a correlation range between − 0.62 and − 0.89 (Table 3). MiR-1234-3p showed the highest number of significant correlations followed by an intermediate group consisting of miR-324-5p, miR-151-5p, miR-200c-3p, miR-200c-3p, miR-362-5p, miR-502-3p and miR-627-5p. A number of genes involved in mitral valve stenosis, myocardial infection, ischemia and acute coronary syndrome exhibited a statistically significantly correlation between miRNA and mRNA expression. In addition, based on microarray data and RT-qPCR, miR-1234-3p was down-regulated and LIMK2, DYSF, GBP2, and MMP9 were up-regulated whereas one mRNA (POT1) was down-regulated and miR-324-5p, miR-151-5p, miR-200c-3p, miR-200c-3p, miR-362-5p, miR-502-3p and miR-627-5p were up-regulated.

Correlation between the clinical parameters and expression levels of miRNA and mRNA

We further analyzed the correlations between the validated 11 miRNAs and 6 mRNAs by RT-qPCR and various clinical parameters of MFS. We found that the expression levels of 7 miRNAs including miR-151-5p,

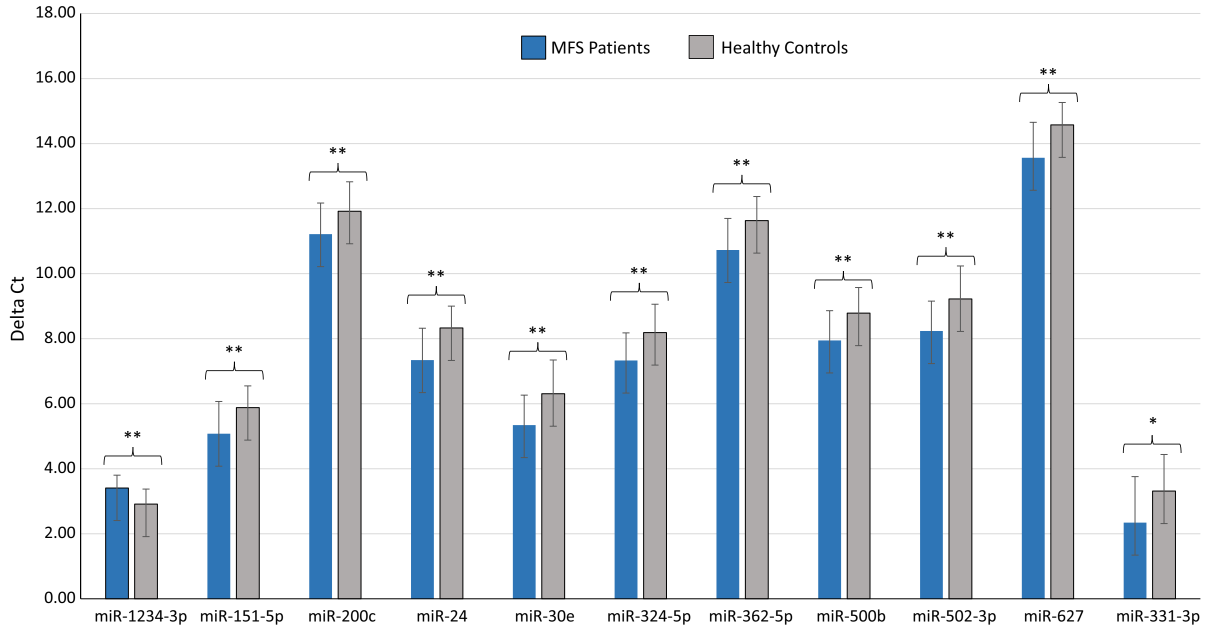


Fig. 3 Validation of differentially expressed miRNAs in the blood of MFS patients (n = 26) compared to HV controls (n = 26) as determined by RT-qPCR ($P < 0.05$). Mean ΔCt MFS and HV controls (Lower ΔCt, higher expression level). RNAU6B as an endogenous control for normalization, Unpaired-two-tailed t tests and ± standard deviation (STDV) were used to evaluate differences in expression. *$P \leq 0.05$; **$P \leq 0.01$; ***$P \leq 0.001$

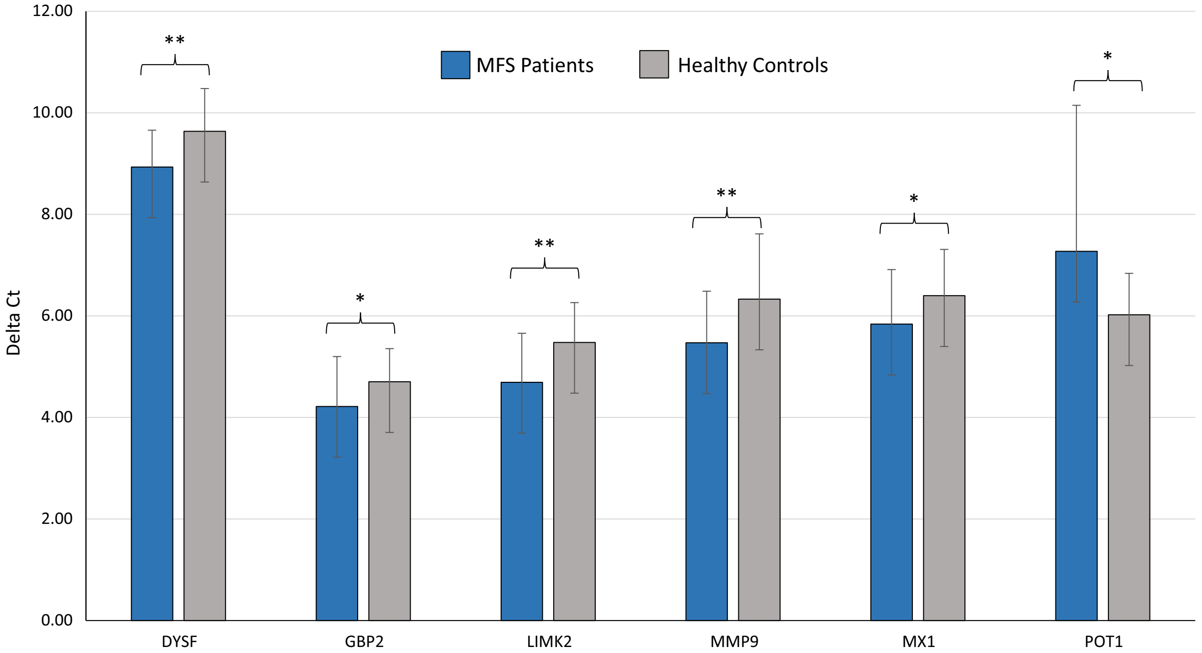


Fig. 4 Validation of differentially expressed mRNAs in the blood of MFS patients (n = 26) compared to HV controls (n = 26) as determined by RT-qPCR (P < 0.05). Mean ΔCt MFS and HV controls (Lower ΔCt, higher expression level). β-Actin as an endogenous housekeeping gene for normalization, Unpaired-two-tailed t tests and ± standard deviation (STDV) were used to evaluate differences in expression. *P ≤ 0.05; **P ≤ 0.01; ***P ≤ 0.001

Table 3 Significant negative correlation between the identified miRNA and mRNA by microarray (un-paired two-tailed t test < 0.5/> 1.5-fold difference, P < 0.05)

miRNA	mRNA	P value correlation	Correlation	Fold change miRNA	P value miRNA	Fold change mRNA	P value mRNA
hsa-miR-1234	LIMK2	0.00001	− 0.89	0.12	0.00038	2.22	0.01813
hsa-miR-1234	DYSF	0.00018	− 0.85	0.12	0.00038	2.74	0.01491
hsa-miR-1234	GBP2	0.00143	− 0.78	0.12	0.00038	2.04	0.00973
hsa-miR-1234	MMP9	0.00382	− 0.74	0.12	0.00038	2.15	0.03885
hsa-miR-324-5p	POT1	0.00670	− 0.69	1.72	0.00050	0.47	0.00017
hsa-miR-151-5p	POT1	0.00728	− 0.68	1.73	0.01754	0.47	0.00017
hsa-miR-200c	POT1	0.01132	− 0.65	43.66	0.01425	0.47	0.00017
hsa-miR-362-5p	POT1	0.01281	− 0.64	1.52	0.00045	0.47	0.00017
hsa-miR-502-3p	POT1	0.01590	− 0.63	1.58	0.00054	0.47	0.00017
hsa-miR-500b	POT1	0.01868	− 0.62	48.35	0.01419	0.47	0.00017
hsa-miR-627	POT1	0.01868	− 0.62	33.47	0.03109	0.47	0.00017

miR-24, miR-30e, miR-324-5p, miR-362-5p, miR-500b, and miR-502-3p significantly correlated with the age of patients with MFS (Table 4) ($P < 0.05$). In contrast, no significant correlation was observed between the expression levels of these 7 miRNAs and the aortic root status of patients with MFS. However, there was a significant correlation between the expression level of miR-200c (P value = 0.015) and a borderline significant correlation with miR-151-5p, miR-324-5p and miR-500b and aortic root status (Z-score) of patients with MFS. Statistically significant correlations were observed between 7 miRNAs including miR-151-5p, miR-24, miR-30e, miR-324-5p, miR-500b, miR-502-3p, and miR-627) and the LVEDD. There was no significant correlation between

Table 4 Correlation between clinical parameters and validated miRNA and mRNA expression levels by RT-qPCR in patients with MFS P < 0.05)

Parameters	**Marfan syndrome patients**							**Healthy volunteers**	
	Age (MFS patients)		**Aortic root status (Z score)**		**LVEDD**		**MVP**	**Age**	
miRNA	**Correlation**	**P value**	**Correlation**	**P value**	**Correlation**	**P value**	**P value***	**Correlation**	**P value**
hsa-miR-1234	0.003	0.988	0.063	0.761	− 0.075	0.716	0.732	− 0.122	0.554
hsa-miR-151-5p	0.519	*0.007*	0.375	***0.059***	0.496	*0.010*	0.979	0.055	0.789
hsa-miR-200c	0.370	***0.063***	0.473	*0.015*	0.213	0.295	0.060	− 0.220	0.279
hsa-miR-24	0.496	*0.010*	0.281	0.164	0.420	*0.033*	0.654	− 0.194	0.342
hsa-miR-30e	0.565	*0.003*	0.230	0.258	0.472	*0.015*	0.517	− 0.030	0.883
hsa-miR-324-5p	0.510	*0.008*	0.350	***0.079***	0.457	*0.019*	0.391	− 0.019	0.927
hsa-miR-362-5p	0.413	*0.036*	0.216	0.289	0.383	***0.053***	0.673	0.024	0.906
hsa-miR-500b	0.440	*0.024*	0.384	***0.053***	0.456	*0.019*	0.812	− 0.003	0.987
hsa-miR-502-3p	0.505	*0.008*	0.260	0.200	0.425	*0.030*	0.816	0.036	0.860
hsa-miR-627	0.310	0.123	0.291	0.149	0.423	*0.032*	0.816	0.010	0.961
hsa-miR-331-3p	0.307	0.127	0.333	0.097	0.015	0.942	0.041	− 0.243	0.231

Parameters	**Marfan syndrome patients**							**Healthy volunteers**	
	Age		**Aortic root status (Z score)**		**LVEDD**		**MVP**	**Age**	
mRNA	**Correlation**	**P value**	**Correlation**	**P value**	**Correlation**	**P value**	**P value**	**Correlation**	**P value**
DYSF	0.123	0.550	0.087	0.672	− 0.010	0.963	0.958	0.186	0.385
GBP2	0.200	0.326	0.071	0.729	− 0.198	0.332	0.816	0.080	0.710
LIMK2	0.003	0.989	− 0.142	0.489	− 0.241	0.235	0.916	0.135	0.530
MMP9	0.276	0.173	0.272	0.178	− 0.042	0.840	0.460	0.295	0.162
MX1	0.071	0.732	− 0.142	0.489	0.019	0.926	0.897	0.073	0.736
POT1	− 0.028	0.891	− 0.041	0.842	− 0.306	0.129	0.510	− 0.075	0.727

LVEDD left ventricular end diastolic diameter, MVP mitral valve prolapse. P values were calculated using unpaired-two-tailed t test (P < 0.05). *P values were calculated using Wilcoxon test (P < 0.05)

Italic—significant with adjusted P-value

Bold Italic—borderline significant with adjusted P-value

the validated target mRNAs and the clinical parameters of MFS. Furthermore, we also assessed the significance of the differences of the validated 11 miRNAs in MFS patients without MVP compared with patients with MVP using the Wilcoxon test. Out of these 11 miRNAs, miR-331-3p showed a significant down-regulation in patients with MVP compared to patients without MVP ($P < 0.05$). MiR-200c showed a borderline significant decrease in MFS patients with MVP compared with patients without MVP ($P = 0.060$).

Classification and overrepresentation analysis

Considering only the protein coding genes and removing of different transcript variants, 292 genes out of 296 were grouped according to PANTHER protein class, GO Molecular Function, GO Biological Process and GO cellular components annotations. The complete classifications can be found in the Additional file 3: Table S4. In detail, after applying Bonferroni correction for multiple testing, there were statistically significant pathways within the 292 genes differentially expressed in the MFS patients, displaying apoptosis signaling pathway (P value = 1.54E−03), JAK/STAT signaling pathway (P value = 1.78E−02), integrin signaling pathway (P value = 1.92E−02) and angiogenesis (P value = 3.50E−02) (Table 5).

Discussion

In this study, we found 13 miRNAs and 31 mRNAs with significantly increased expression levels and 15 miRNAs and one single mRNA (POT1) with significantly decreased expression level in patients with MFS compared with HV controls. In a cohort of independent MFS patients and HV controls, 11 miRNAs and 6 mRNAs were validated. These data show that miRNA and mRNA expression levels in the blood of patients with MFS differ

Table 5 Pathways significantly enriched for the identified protein coding genes in the blood of patients with MFS compared to HV controls (adjusted P value < 0.05)

PANTHER classification pathways	Number of genes			Over-/under-represented (±)	Fold enrichment	P value
	Reference list[a]	Target list[b]	Expected[c]			
Apoptosis signaling pathway (P00006)	119	10	1.68	+	5.96	0.00154
JAK/STAT signaling pathway (P00038)	17	4	0.24	+	16.69	0.00210
Integrin signalling pathway (P00034)	194	11	2.73	+	4.02	0.01920
Angiogenesis (P00005)	174	10	2.45	+	4.08	0.03500

Pathways resulted significantly over-represented by the identified protein coding genes. P values were tested using Fisher exact test and adjusted using a Bonferroni correction test. MFS Marfan syndrome, HVs healthy volunteers

[a] Number of genes in the reference list that map to this PANTHER classification category

[b] Number of genes in the target genes list that map to this PANTHER classification category

[c] Expected value is the number of genes that could be expected in target genes list for this PANTHER category based on the reference list

from HV controls and that distinct differences in specific miRNA expression patterns can be further explored as potential biomarkers for differentiating between patients with MFS and HV controls. A distinctive non-invasive surrogate biomarker for MFS would be of high clinical value, as mutation analysis of the huge (65 exons) FBN1 gene is still relatively expensive and time-consuming and therefore restricted to phenotypically recognized Ghent-positive patients. An affordable screening test for MFS would likely detect a considerable number of atypical MFS who currently remain undiagnosed. Moreover, our investigation provides a comprehensive analysis of the gene expression pattern in patients with MFS as compared to HV controls, suggesting that non-pathogenic variants of other genes than FBN1 may significantly influence the phenotype, and explain the often striking clinical variation among members of a given MFS family. The identification of these genes may lead to a possible novel signature related to MFS, provide new prognostic parameters and ultimately even generate targets for novel approaches to chemoprevention of complications beyond currently unsatisfying medical treatment options [18]. Intriguingly, many of the biological pathways identified, such as apoptosis signaling [19], JAK/STAT signaling [20], integrin signaling [21] and angiogenesis pathways [22], have been associated with development of cardiovascular complications in MFS and its related diseases including aortic and pulmonary artery dilation as well as mitral valve prolapse. Among the identified deregulated mRNAs, some genes play a role in cardiomyocyte differentiation and remodeling during acute myocardial infarction and in dilated cardiomyopathy (DCM). For example, patients suffering from DCM show a strong and lasting increase of *oncostatin M* (OSM) gene expression level and signaling [23]. Moreover, significant changes in *clusterin* [24] gene level have been detected in patients with acute myocardial infarction (AMI) [25] and increased levels of *selectin L* (SELL) are associated with ischemic stroke [26]. The JAK/STAT pathway is negatively regulated by the suppressor of cytokine signaling (SOCS) protein, and the myocardium-specific suppressor of *cytokine signaling 3* (SOCS3) gene plays a key role in the development of left ventricular (LV) remodeling after AMI [27]. In agreement with the higher expression level of *C-X-C motif chemokine ligand 5* (CXCL5) in the blood of patients with MFS, CXCL5 showed an increased expression level in the plasma of patients with coronary artery disease. Recent studies have showed that CXCL5 and its receptors are implicated in congestive heart failure and ischemic stroke, making CXCL5 a candidate gene for potential future therapy strategies in cardiovascular diseases [28–30]. CXCL5 has also been reported to be up-regulated in abdominal aortic aneurysm (AAA) [31]. *Matrix metalloproteinase 9* (MMP) was shown to be up-regulated in the blood of MFS patients by microarray and RT-qPCR in our analysis. MMP9 showed a significant inverse correlation with hsa-miR-1234, which also was identified in the MFS patients [9]. A proteolytic degradation of the extracellular matrix of the aortic wall by an upregulation MMPs has been shown to be involved in the pathogenesis of TAA and AAA and also contributes to the histologic changes found in the aortic wall of patients with MFS [32, 33]. The expression of MMP9 has been shown to be up-regulated in the vascular wall of human AAA [34, 35] and also in aneurysm tissue in a mouse model of MFS [32]. Interestingly, Balistreri et al., found potential associations of SNPs in the MMP9 gene [rs3918242 (−1562C/T MMP-9)] degenerative forms of mitral valve diseases (MVDs), concluding that genetic variants in MMP9 play a role in MVD in MFS patients [36]. Together with our data, showing an up-regulation of MMP9 in the blood of MFS patients compared to controls, indicate that MMP9 may represent a potential biomarker and therapeutic target to reduce the growth rate of TAAs in MFS patients. *Doxycyclin* and statins have proven to be effective inhibitors of MMPs [37, 38] and have shown therapeutic benefits in both TAA and AAA patients [39, 40]. However, data on MFS patients as well as large randomized trials are still lacking, making these drugs promising candidates for future investigations in MFS. Our data, showing a significant inverse correlation of miR-1234 and MMP9 indicate that a down-regulation of this miRNA may be involved in the up-regulation MMP9 in MFS. We demonstrate a significantly up-regulated expression of the *LIM kinase 2* (LIMK2) which also inversely correlated with miR-1234. LIMK2 regulates dynamic changes of the actin cytoskeleton by phosphorylating *cofilin* and thereby inactivating its F-actin depolymerizing activity [41]. It was shown in mouse models that an activation of LIMK2 is associated with a disturbed flow in the aortic arch and disturbs endothelial cell (EC) barrier function, which was reversed by inhibition of LIMK2 with *m-calpain* [42]. An up-regulation of LIMK2 likely linked to a down-regulation of miR-1234 in the blood of MFS patients, which was demonstrated in our study, therefore may be related to elevated levels of vascular wall shear stress in the thoracic aorta of MFS patients [43] and be associated with endothelial dysfunction in MFS. Since effective LIMK2 inhibition has already been shown to improve endothelial function in animal models [42]. LIMK2 may represent a promising target for future investigations in MFS patients. Our data shown a significant up-regulation of *guanylate binding protein-2* (GBP-2) and a significant inverse correlation with miR-1234 in the blood of MFS patients compared to controls. Human *guanylate binding proteins* (GBPs) are a class of large GTPases which are induced by cytokines like Interferon alpha/gamma, Interleukin-1 and TNF-alpha [44]. GBP-2 has not yet been investigated as comprehensively as GBP-1, but shares 75% sequence identity with this isoform [45] which has been shown to be actively secreted by ECs [46]. Patients with rheumatic diseases like rheumatoid arthritis, systemic lupus erythematosus [22], and systemic sclerosis, which are characterized by a chronic inflammatory vessel activation, show reduced levels of GBP-1 in their peripheral blood [44]. In a rat arteriovenous (AV) loop model, it has been shown that GBP-1 inhibits endothelial cell progenitor migration and leads to endothelial cell dysfunction [44]. An up-regulation of GBP-2 in the blood of MFS patients is likely reflects the vascular pathology and disturbed endothelial cell function in these patients. A significant up-regulation and inverse correlation to miR-1234 in the blood of MFS patients compared to controls was also shown for *dysferlin* (DYSF). Mutations in DYSF lead to limb-girdle muscular dystrophy type 2B and Miyoshi myopathy. DYSF, which is expressed in human ECs, has been shown to form a complex with *platelet endothelial cellular adhesion molecule-1* (PECAM-1), thereby preventing its proteosomal degradation [47]. Since PECAM-1 is a ligand of $\alpha_V\beta_3$-integrin and a promotor of angiogenesis, these data are in line with our observation of an enrichment of gene sets for angiogenesis and integrin signaling in the blood of MFS patients. DYSF is induced in vitro by TNF-alpha and has also been shown to be up-regulated in the blood vessels of patients with multiple sclerosis representing increased vascularinflammation and a disturbed blood–brain barrier [48]. It seems likely that an overexpression of DYSF in the blood of MFS patients, as with GBP-2, represents vascular inflammation in these patients and may be a potential biomarker for the severity of vascular pathologies in MFS warranting further investigations. The gene encoding for *protection of telomeres 1* (POT1) was the only significantly down-regulated gene in the blood of MFS patients (fold-change ≤ 2) compared to controls and exhibited inverse correlations with 7 miRNAs. POT1 binds single-stranded DNA as a heterodimer with *tripeptidyl peptidase 1* (TPP1) and promotes telomerase-mediated telomere extension. Reduced telomere length is recognized as a hallmark of cardiovascular aging and as a biomarker for and TAA and dissections [49, 50]. It has to date not been investigated whether telomere length plays a role in the pathogenesis of MFS. The reduced expression of POT1 in the blood of MFS patients demonstrated in our study, however, indicates, that accelerated cardiac ageing may be present in MFS, which may be reflected in reduced telomere length and POT1 expression. Among the miRNAs inversely correlating with POT1, miR-362 has been linked to the degree of inflammation in samples from abdominal aortic aneurysms [51]. Moreover, miR-500 was shown to be deregulated in degenerative mitral valve disease [52] and miR-502 was also up-regulated in the sera of patients with congestive heart failure [53]. Some genes which we found to be up-regulated in the blood of MFS patients compared to controls play a role in cardiomyocyte differentiation and remodeling during AMI as well as in DCM. Patients with DCM show a strong and lasting increase of *oncostatin M* (OSM) gene expression [23]. Moreover, significant changes in *clusterin* [24] have been detected in patients with AMI [25] and increased levels of *selectin L* (SELL) are associated with ischemic stroke [26]. The JAK/STAT pathway is negatively regulated by the suppressor of cytokine signaling (SOCS) protein, and the myocardium-specific suppressor of *cytokine signaling 3* (SOCS3) gene plays a key role in the development of left ventricular (LV) remodeling after AMI [27]. In agreement with the higher expression level of *C-X-C motif chemokine ligand 5* (CXCL5) in the blood of patients with MFS, CXCL5 showed an increased expression level in the plasma of patients with coronary artery disease. Recent studies have shown that CXCL5 is up-regulated in abdominal aortic aneurysms (AAA) [31], making it a candidate for potential future anti-inflammatory therapy strategies in MFS. We identified three miRNAs, namely miR-151-5p, miR-324-5p, and miR-500b, which correlated significantly with the Z-score and the LVEDD of MFS patients. MiR-24 and miR-30e correlated only with the LVEDD of MFS patients in our study. MiR-24 has been reported to be up-regulated in tissue from thoracic aortic aneurysms [54] and the miR-30-family was shown to be up-regulated in tissue from thoracic aortic dissections and abdominal aortic aneurysms [55, 56]. Interestingly miR-331-3p, which has been linked to cardiac hypertrophy [57] and miR-200, which also has been linked to cardiovascular disease [58], were down-regulated in MFS patients with MVP compared with patients without MVP. MiR-200c-3p also showed an inverse correlation to POT1 in our study. These deregulated miRNAs may serve as potential future biomarkers in MFS after conformational analysis in studies with larger sample sizes. MiRNA and mRNA profiles measured in PAXgene blood samples comes to a greater extent from the cellular components of the blood, i.e. leukocytes and erythrocytes, and only to a lesser extent from cell-free RNA. Therefore, the expression changes we identified in our study presumably reflect rather changes in the blood cells of the patients rather than expression changes in solid tissue, i.e. bone. One of the MFS clinical manifestations is the musculoskeletal system (typically tall stature with arachnodactyly) and patients with MFS also have significant musculoskeletal phenotypes which may affect the marrow cavity and subsequently influence the hematopoiesis process. Therefore, it is conceivable that changes in miRNA and mRNA expression profile in the blood of MFS patients might be the results of differences in the hematopoiesis process in MFS patients compared to healthy controls. Another line of thinking is, that differences in the mRNA and miRNA profiles might originate from differences in blood flow kinetics between MFS patients and controls. It is known that altered blood pressure has an impact on miRNA expression profile, as shown by Neth et al. [59]. Aortic dilatation and structural cardiac anomalies like MVP in MFS patients exposes the vascular endothelium to altered hemodynamic forces, which may indirectly influence miRNA and mRNA profiles in the blood cells of MFS patients due to different blood flow velocities compared to healthy individuals.

Limitations of our study are related to a relatively small sample size. Moreover, our analysis focused only on the main diagnostic criteria such as FBN1 positivity, aortic root dilatation and lens dislocation and revealed correlations of miRNA expression to cardiovascular features such as aortic root dilatation and mitral valve prolapse. Skeletal features which are characterized by a highly variable age of onset are heterogeneous and have been considered as secondary diagnostic criteria according to the modified Ghent criteria. Certainly, the skeletal features are important leading diagnostic criteria for further evaluation of the patients suspected to have Marfan disease. Correlation of miRNA expression to skeletal abnormalities has to be performed in future studies with larger cohorts of patients with definitive and highly characterized main skeletal features. Future studies also have to investigate whether the observed miRNA expression profiles are specific to MFS or also relate to other syndromes with familial thoracic aortic aneurysm like Loeys–Dietz syndrome, Shprintzen–Goldberg syndrome or mutations in ACTA2.

Conclusions

We present the first study investigating miRNA and mRNA expression patterns in the peripheral blood of MFS patients in comparison with HV controls. A strong deregulation of both miRNA and mRNA expression profiles was present in MFS patients including multiple genes with high relevance to cardiovascular pathogenesis and diseases. Four genes associated with vascular pathology and inflammation namely MMP9, LIMK2, GBP-2, and DYSF were up-regulated in MFS patients and showed inverse correlations with miR-1234. POT-1 was down-regulated and inversely correlated with 7 miRNAs indicating a potential role of telomere length in the pathogenesis of MFS. These genes represent promising candidates for future investigations aiming at prognostic biomarkers for cardiovascular manifestations in MFS as well as targets for novel therapeutic approaches. Apart from the particular considerations as to the value of the observed distinctive miRNA/mRNA patterns for diagnosis and prognosis of MFS patients, our study fundamentally highlights the extreme breadth of molecular downstream effects initiated by a constitutional single point mutation in a monogenic heritable condition. Pleiotropy also has an as yet underestimated molecular dimension that may provide insights into how complex seemingly "simple" monogenic traits actually are.

Additional files

Additional file 1: Figure S1. Pearson correlation coefficient-based heat map representation between samples. Samples are clustered by the Euclidean distance between rows and columns based on miRNA expression level.

Additional file 2: Figure S2. Pearson correlation coefficient-based heat map representation between samples. Samples are clustered by the Euclidean distance between rows and columns based on mRNA expression level.

Additional file 3: Table S1. Clinical characteristics of patients. **Table S2.** Significantly expressed transcripts s in the blood of patients with MFS (n = 7) compared HVs controls (n = 7) as determined by microarray (*P*-value <0.05). **Table S3.** Significantly expressed protein coding genes in the blood of patients with MFS (n = 7) compared HVs controls (n = 7) as determined by microarray (*P*-value <0.05). **Table S4.** Over-representation analysis of target genes list.

Abbreviations

AAA: abdominal aortic aneurysms; AMI: acute myocardial infarction; cDNA: complementary DNA; cRNA: complementary RNA; DCM: dilated cardiomyopathy; EC: endothelial cell; FBN1: Fibrillin-1; GO: Gene Ontology; HV: healthy volunteer; LV: left ventricular; LVEDD: left ventricular end-diastolic diameter; MGS: Marfan syndrome; miRNAs: microRNAs; miRTC: miRNA reverse transcription control; MRI: magnetic resonance imaging; mRNA: messenger RNA; MVP: mitral valve prolapse; NTC: no template control; PANTHER: Protein ANalysis THrough Evolutionary Relationships; PCR: polymerase chain reaction; RT-qPCR: real-time quantitative PCR; snRNA: small nuclear RNA; TAA: thoracic aortic aneurysms; 3′UTR: 3′ untranslated region; vsn: variance stabilizing normalization.

Authors' contributions

MA performed experimental work, particularly the miRNA and mRNA, array experiment, RT-qPCR validation and wrote the manuscript; MK, performed bioinformatics analysis and helped in article writing, DH, helped in the interpretation of clinical data and article writing; AK, performed bioinformatics analysis, EM, designed the study; TRH, recruited, examined matched controls and collected blood samples; HAK, designed the study, diagnosed patients and edited the manuscript; EM, designed the study, coordinated the molecular biology experiment and edited the manuscript. All authors read and approved the final manuscript.

Author details

[1] Institute of Human Genetics, Saarland University, 66421 Homburg/Saar, Germany. [2] Chair for Clinical Bioinformatics, Saarland University, 66041 Saarbrücken, Germany. [3] Department of Hand, Plastic and Reconstructive Surgery, BG Trauma Center Ludwigshafen, University of Heidelberg, 67071 Ludwigshafen, Germany. [4] Department of Pediatric Cardiology, Saarland University Medical Center, 66421 Homburg/Saar, Germany.

Acknowledgements

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The data sets during and/or analyzed during the current study available from the corresponding author on reasonable request.

Consent for publication

Subjects were recruited and evaluated in accordance with relevant guidelines and regulations following the approval of the local ethics committee (Ethical vote Nr: EA2/131/10). All participants or their legal guardians gave written informed consent before enrolment.

Ethics approval and consent to participate

Institutional Review Board approval, Ethical vote (Nr: EA2/131/10).

Funding

This study was funded by the Hedwig-Stalter-Stiftung, Competence Network for Congenital Heart Defects, which received funding from the Federal Ministry of Education and Research, Grant Number 01GI0601 (2014), and the German Centre for Cardiovascular Research, Grant Number 81X2800112 (2015).

Publisher's Note

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

Received: 30 October 2017 Accepted: 28 February 2018
Published online: 12 March 2018

References

1. Pyeritz RE. Marfan syndrome: current and future clinical and genetic management of cardiovascular manifestations. Semin Thorac Cardiovasc Surg. 1993;5:11–6.
2. Tsang AK, Taverne A, Holcombe T. Marfan syndrome: a review of the literature and case report. Spec Care Dent. 2013;33:248–54.
3. Dietz HC. Marfan syndrome. In: Pagon RA, Adam MP, Ardinger HH, Wallace SE, Amemiya A, Bean LJH, Bird TD, Ledbetter N, Mefford HC, Smith RJH, Stephens K, editors. GeneReviews(R). Seattle: University of Washington; 1993.
4. Collod-Beroud G, Le Bourdelles S, Ades L, Ala-Kokko L, Booms P, Boxer M, Child A, Comeglio P, De Paepe A, Hyland JC, et al. Update of the UMD-FBN1 mutation database and creation of an FBN1 polymorphism database. Hum Mutat. 2003;22:199–208.
5. Judge DP, Dietz HC. Marfan's syndrome. Lancet. 2005;366:1965–76.
6. von Kodolitsch Y, De Backer J, Schuler H, Bannas P, Behzadi C, Bernhardt AM, Hillebrand M, Fuisting B, Sheikhzadeh S, Rybczynski M, et al. Perspectives on the revised Ghent criteria for the diagnosis of Marfan syndrome. Appl Clin Genet. 2015;8:137–55.
7. Stearns FW. One hundred years of pleiotropy: a retrospective. Genetics. 2010;186:767–73.
8. Faivre L, Collod-Beroud G, Loeys BL, Child A, Binquet C, Gautier E, Callewaert B, Arbustini E, Mayer K, Arslan-Kirchner M, et al. Effect of mutation type and location on clinical outcome in 1,013 probands with Marfan syndrome or related phenotypes and FBN1 mutations: an international study. Am J Hum Genet. 2007;81:454–66.
9. Abu-Halima M, Ludwig N, Radle-Hurst T, Keller A, Motsch L, Marsollek I, El Rahman MA, Abdul-Khaliq H, Meese E. Characterization of micro-RNA profile in the blood of patients with Marfan's syndrome. Thorac Cardiovasc Surg. 2018;66:116–24.
10. Ikonomidis JS, Ivey CR, Wheeler JB, Akerman AW, Rice A, Patel RK, Stroud RE, Shah AA, Hughes CG, Ferrari G, et al. Plasma biomarkers for distinguishing etiologic subtypes of thoracic aortic aneurysm disease. J Thorac Cardiovasc Surg. 2013;145:1326–33.
11. Merk DR, Chin JT, Dake BA, Maegdefessel L, Miller MO, Kimura N, Tsao PS, Iosef C, Berry GJ, Mohr FW, et al. miR-29b participates in early aneurysm development in Marfan syndrome. Circ Res. 2012;110:312–24.
12. Yao Z, Jaeger JC, Ruzzo WL, Morale CZ, Emond M, Francke U, Milewicz DM, Schwartz SM, Mulvihill ER. A Marfan syndrome gene expression phenotype in cultured skin fibroblasts. BMC Genom. 2007;8:319.
13. Schwedler G, Lindinger A, Lange PE, Sax U, Olchvary J, Peters B, Bauer U, Hense HW. Frequency and spectrum of congenital heart defects among live births in Germany: a study of the Competence Network for Congenital Heart Defects. Clin Res Cardiol. 2011;100:1111–7.
14. Abu-Halima M, Ludwig N, Hart M, Leidinger P, Backes C, Keller A, Hammadeh M, Meese E. Altered micro-ribonucleic acid expression profiles of extracellular microvesicles in the seminal plasma of patients with oligoasthenozoospermia. Fertil Steril. 2016;106(5):1061–9.e3.
15. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 2013;41:D377–86.

16. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. 2002;18(Suppl 1):S96–104.
17. Abu-Halima M, Ludwig N, Hart M, Leidinger P, Backes C, Keller A, Hammadeh M, Meese E. Altered micro-ribonucleic acid expression profiles of extracellular microvesicles in the seminal plasma of patients with oligoasthenozoospermia. Fertil Steril. 2016;106:1061–9.
18. Milleron O, Arnoult F, Ropers J, Aegerter P, Detaint D, Delorme G, Attias D, Tubach F, Dupuis-Girod S, Plauchu H, et al. Marfan Sartan: a randomized, double-blind, placebo-controlled trial. Eur Heart J. 2015;36:2160–6.
19. Nataatmadja M, West M, West J, Summers K, Walker P, Nagata M, Watanabe T. Abnormal extracellular matrix protein transport associated with increased apoptosis of vascular smooth muscle cells in marfan syndrome and bicuspid aortic valve thoracic aortic aneurysm. Circulation. 2003;108(Suppl 1):II329–34.
20. Kishore R, Verma SK. Roles of STATs signaling in cardiovascular diseases. JAKSTAT. 2012;1:118–24.
21. Mariko B, Ghandour Z, Raveaud S, Quentin M, Usson Y, Verdetti J, Huber P, Kielty C, Faury G. Microfibrils and fibrillin-1 induce integrin-mediated signaling, proliferation and migration in human endothelial cells. Am J Physiol Cell Physiol. 2010;299:C977–87.
22. Kessler K, Borges LF, Ho-Tin-Noe B, Jondeau G, Michel JB, Vranckx R. Angiogenesis and remodelling in human thoracic aortic aneurysms. Cardiovasc Res. 2014;104:147–59.
23. Kubin T, Poling J, Kostin S, Gajawada P, Hein S, Rees W, Wietelmann A, Tanaka M, Lorchner H, Schimanski S, et al. Oncostatin M is a major mediator of cardiomyocyte dedifferentiation and remodeling. Cell Stem Cell. 2011;9:420–32.
24. Joy J, McClure N, Cooke IE. A comparison of spontaneously conceived twins and twins conceived by artificial reproductive technologies. J Obstet Gynaecol. 2008;28:580–5.
25. Cubedo J, Padro T, Garcia-Moll X, Pinto X, Cinca J, Badimon L. Proteomic signature of Apolipoprotein J in the early phase of new-onset myocardial infarction. J Proteome Res. 2011;10:211–20.
26. Wei YS, Lan Y, Meng LQ, Nong LG. The association of L-selectin polymorphisms with L-selectin serum levels and risk of ischemic stroke. J Thromb Thrombolysis. 2011;32:110–5.
27. Oba T, Yasukawa H, Hoshijima M, Sasaki K, Futamata N, Fukui D, Mawatari K, Nagata T, Kyogoku S, Ohshima H, et al. Cardiac-specific deletion of SOCS-3 prevents development of left ventricular remodeling after acute myocardial infarction. J Am Coll Cardiol. 2012;59:838–52.
28. Damas JK, Eiken HG, Oie E, Bjerkeli V, Yndestad A, Ueland T, Tonnessen T, Geiran OR, Aass H, Simonsen S, et al. Myocardial expression of CC- and CXC-chemokines and their receptors in human end-stage heart failure. Cardiovasc Res. 2000;47:778–87.
29. Damas JK, Gullestad L, Ueland T, Solum NO, Simonsen S, Froland SS, Aukrust P. CXC-chemokines, a new group of cytokines in congestive heart failure—possible role of platelets and monocytes. Cardiovasc Res. 2000;45:428–36.
30. Zineh I, Beitelshees AL, Welder GJ, Hou W, Chegini N, Wu J, Cresci S, Province MA, Spertus JA. Epithelial neutrophil-activating peptide (ENA-78), acute coronary syndrome prognosis, and modulatory effect of statins. PLoS ONE. 2008;3:e3117.
31. Yoshimura K, Nagasawa A, Kudo J, Onoda M, Morikage N, Furutani A, Aoki H, Hamano K. Inhibitory effect of statins on inflammation-related pathways in human abdominal aortic aneurysm tissue. Int J Mol Sci. 2015;16:11213–28.
32. Chung AW, Au Yeung K, Sandor GG, Judge DP, Dietz HC, van Breemen C. Loss of elastic fiber integrity and reduction of vascular smooth muscle contraction resulting from the upregulated activities of matrix metalloproteinase-2 and -9 in the thoracic aortic aneurysm in Marfan syndrome. Circ Res. 2007;101:512–22.
33. Segura AM, Luna RE, Horiba K, Stetler-Stevenson WG, McAllister HA Jr, Willerson JT, Ferrans VJ. Immunohistochemistry of matrix metalloproteinases and their inhibitors in thoracic aortic aneurysms and aortic valves of patients with Marfan's syndrome. Circulation. 1998;98:II331–7 **(discussion II337–338)**.
34. Armstrong PJ, Johanning JM, Calton WC Jr, Delatore JR, Franklin DP, Han DC, Carey DJ, Elmore JR. Differential gene expression in human abdominal aorta: aneurysmal versus occlusive disease. J Vasc Surg. 2002;35:346–55.
35. McMillan WD, Patterson BK, Keen RR, Shively VP, Cipollone M, Pearce WH. In situ localization and quantification of mRNA for 92-kD type IV collagenase and its inhibitor in aneurysmal, occlusive, and normal aorta. Arterioscler Thromb Vasc Biol. 1995;15:1139–44.
36. Balistreri CR, Allegra A, Crapanzano F, Pisano C, Triolo OF, Argano V, Candore G, Lio D, Ruvolo G. Associations of rs3918242 and rs2285053 MMP-9 and MMP-2 polymorphisms with the risk, severity, and short- and long-term complications of degenerative mitral valve diseases: a 4.8-year prospective cohort study. Cardiovasc Pathol. 2016;25:362–70.
37. Luan Z, Chase AJ, Newby AC. Statins inhibit secretion of metalloproteinases-1, -2, -3, and -9 from vascular smooth muscle cells and macrophages. Arterioscler Thromb Vasc Biol. 2003;23:769–75.
38. Xiong W, Knispel RA, Dietz HC, Ramirez F, Baxter BT. Doxycycline delays aneurysm rupture in a mouse model of Marfan syndrome. J Vasc Surg. 2008;47:166–72 **(discussion 172)**.
39. Mosorin M, Juvonen J, Biancari F, Satta J, Surcel HM, Leinonen M, Saikku P, Juvonen T. Use of doxycycline to decrease the growth rate of abdominal aortic aneurysms: a randomized, double-blind, placebo-controlled pilot study. J Vasc Surg. 2001;34:606–10.
40. Odajima H, Baba M. The relationship between respiratory threshold to acetylcholine and prognosis for asthma. Arerugi. 1990;39:526–31.
41. Khurana T, Khurana B, Noegel AA. LIM proteins: association with the actin cytoskeleton. Protoplasma. 2002;219:1–12.
42. Miyazaki T, Honda K, Ohata H. m-Calpain antagonizes RhoA overactivation and endothelial barrier dysfunction under disturbed shear conditions. Cardiovasc Res. 2010;85:530–41.
43. Geiger J, Arnold R, Herzer L, Hirtler D, Stankovic Z, Russe M, Langer M, Markl M. Aortic wall shear stress in Marfan syndrome. Magn Reson Med. 2013;70:1137–44.
44. Hammon M, Herrmann M, Bleiziffer O, Pryymachuk G, Andreoli L, Munoz LE, Amann KU, Mondini M, Gariglio M, Airo P, et al. Role of guanylate binding protein-1 in vascular defects associated with chronic inflammatory diseases. J Cell Mol Med. 2011;15:1582–92.
45. Abdullah N, Balakumari M, Sau AK. Dimerization and its role in GMP formation by human guanylate binding proteins. Biophys J. 2010;99:2235–44.
46. Naschberger E, Lubeseder-Martellato C, Meyer N, Gessner R, Kremmer E, Gessner A, Sturzl M. Human guanylate binding protein-1 is a secreted GTPase present in increased concentrations in the cerebrospinal fluid of patients with bacterial meningitis. Am J Pathol. 2006;169:1088–99.
47. Sharma A, Yu C, Leung C, Trane A, Lau M, Utokaparch S, Shaheen F, Sheibani N, Bernatchez P. A new role for the muscle repair protein dysferlin in endothelial cell adhesion and angiogenesis. Arterioscler Thromb Vasc Biol. 2010;30:2196–204.
48. Hochmeister S, Grundtner R, Bauer J, Engelhardt B, Lyck R, Gordon G, Korosec T, Kutzelnigg A, Berger JJ, Bradl M, et al. Dysferlin is a new marker for leaky brain blood vessels in multiple sclerosis. J Neuropathol Exp Neurol. 2006;65:855–65.
49. Balistreri CR, Pisano C, Merlo D, Fattouch K, Caruso M, Incalcaterra E, Colonna-Romano G, Candore G. Is the mean blood leukocyte telomere length a predictor for sporadic thoracic aortic aneurysm? Data from a preliminary study. Rejuvenation Res. 2012;15:170–3.
50. Yan J, Yang Y, Chen C, Peng J, Ding H, Wen Wang D. Short leukocyte telomere length is associated with aortic dissection. Intern Med. 2011;50:2871–5.
51. Busch A, Busch M, Scholz CJ, Kellersmann R, Otto C, Chernogubova E, Maegdefessel L, Zernecke A, Lorenz U. Aneurysm miRNA signature differs, depending on disease localization and morphology. Int J Mol Sci. 2016;17:81.
52. Chen YT, Wang J, Wee AS, Yong QW, Tay EL, Woo CC, Sorokin V, Richards AM, Ling LH. Differential microRNA expression profile in myxomatous mitral valve prolapse and fibroelastic deficiency valves. Int J Mol Sci. 2016;17:753.
53. Cakmak HA, Coskunpinar E, Ikitimur B, Barman HA, Karadag B, Tiryakioglu NO, Kahraman K, Vural VA. The prognostic value of circulating microRNAs in heart failure: preliminary results from a genome-wide expression study. J Cardiovasc Med (Hagerstown). 2015;16:431–7.
54. Patuzzo C, Pasquali A, Malerba G, Trabetti E, Pignatti P, Tessari M, Faggian G. A preliminary microRNA analysis of non syndromic thoracic aortic aneurysms. Balkan J Med Genet. 2012;15:51–5.

55. Biros E, Moran CS, Wang Y, Walker PJ, Cardinal J, Golledge J. microRNA profiling in patients with abdominal aortic aneurysms: the significance of miR-155. Clin Sci (Lond). 2014;126:795–803.
56. Liao M, Zou S, Weng J, Hou L, Yang L, Zhao Z, Bao J, Jing Z. A microRNA profile comparison between thoracic aortic dissection and normal thoracic aorta indicates the potential role of microRNAs in contributing to thoracic aortic dissection pathogenesis. J Vasc Surg. 2011;53(1341–1349):e1343.
57. Calvier L, Chouvarine P, Legchenko E, Hoffmann N, Geldner J, Borchert P, Jonigk D, Mozes MM, Hansmann G. PPARgamma links BMP2 and TGF-beta1 pathways in vascular smooth muscle cells, regulating cell proliferation and glucose metabolism. Cell Metab. 2017;25(1118–1134):e1117.
58. Magenta A, Ciarapica R, Capogrossi MC. The emerging role of miR-200 family in cardiovascular diseases. Circ Res. 2017;120:1399–402.
59. Neth P, Nazari-Jahantigh M, Schober A, Weber C. MicroRNAs in flow-dependent vascular remodelling. Cardiovasc Res. 2013;99:294–303.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

4
Discussion, outlook and conclusion

4.1 Discussion

In this thesis, we approached the miRNA field from different perspectives. We implemented miRMaster to discover new potential high confidence miRNAs candidates, we worked on four different projects to gain insights into the miRNA characteristics which can be supportive for the clinical usability of these molecules, and we conducted two different studies with lung cancer and Marfan syndrome as main diseases to investigate the biomarker potential of miRNAs. However, the herein presented works have limitations, and respective improvements are suggested in the following.

The implemented tool miRMaster is based on one classifier which is selected out of 180 various constellations of classification, feature scaling and subset selection methods. It is state of the art to test plenty of combinations for finding the right model. Nevertheless, the specific training and test set that we used succeeded in the chosen model. Perhaps a different dataset would favour and yield a different model. Alternatively to this single-model procedure, an ensemble learning model could be used [201]. This one consists of multiple weak learners, and the final classification decision is based on majority voting of uncorrelated learners to reduce variance in prediction [202]. In our classical machine learning approach, we manually derived the features from sequence and structure for the classification. In deep learning approaches, the feature derivation from sequences and structures is done automatically, which can reveal hidden features that cannot be observed manually and could thus lead to further performance improvements [203].

While many studies evaluate the potential biomarker capabilities of miRNAs [204], researchers could also show the impact of confounding factors such as gender, age and smoking behavior on miRNA activities [205, 206]. In our project about change of miRNA activity in seasonal profiles, we could show an up-regulation for hsa-miR-106b-5p and hsa-let-7c-5p in spring by analyzing two independent groups. All samples were collected only from healthy volunteers. Retrospectively, it would be interesting from a pathological point of view to include two patient groups differing in the disease (e.g. common diseases like lung cancer and cardiovascular disease). With these additional disease cohorts, we could investigate whether we can confirm the seasonal observations made in the healthy groups. However, one should be aware that it is also possible that biological expression changes related to the health state can overlay seasonal effects in the body, which would make it more difficult to understand seasonal effects among diseased patients.

With regard to our work with sequencing data derived from 21 animals, further time points of blood collection would allow to see what kind of common seasonal effects in miRNA expression exist. This could be interesting because many animals share similar seasonal behavior (e.g. hibernation or flight to more southern regions). Additionally, more replicates of the same species would be helpful to make more reliable conclusions.

Alternatives to the venous blood-based PAXgene like DBS can have promising results regarding technical stability and biological variation. However, the handling of this collection device can be quite challenging which results in different sized dried blood spots and RNA amounts. Microsampling devices (Mitra) as further option can ensure constant blood volumes, which could have an impact for measurement of robust miRNA expression. Inside the category of venous blood sampling, another common method for RNA stabilization is Tempus tube which showed higher RNA outputs compared to PAXgene [207].

With respect to the detection of biomarker candidates for diagnosis of diseases, we have published two articles. The lung cancer project included a substantially higher number of samples than the Marfan publication. Larger sample sizes better reflect the true population, but meta data can become more important too, especially in multi-center projects. To reduce confounding effects based on population, forming similar comparison groups (e.g. perfect matching regarding age, gender and smoking behavior), as shown in the publication, can be important for obtaining statistical correct conclusions [208]. Since the centers might differ in their main focus (e.g. therapy options) or patients may receive different forms of therapy depending on their health condition, it makes sense to include medication and possibly other factors in the analyses and grouping [209]. This could be helpful to optimize diagnosis signatures or to pursue additional biological questions in order to gain knowledge for therapy and clinical routine. Regarding profiling a further challenge for clinical studies can be the exclusive usage of the whole blood samples. The substantial higher amount of RBCs and their miRNAs can overlay the expression based on a potential disease-related immune response. Perhaps additional profiling of the blood cell composition would be helpful to deconvolute the complex whole blood signal into WBC-type-specific signals [210].

4.2 Outlook

In this thesis we focused mainly on miRNAs whose diagnostic potential was investigated in studies performed on microarrays which detect only known entries of miRBase. The usage of NGS enables the discovery of new miRNAs and other sncRNAs as well [155] whose potential for diagnostics and therapy is introduced in the following.

sncRNA biomarkers In the recent years, the clinical interest has grown for studies about circulating sncRNA biomarker candidates such as tRNAs and Y RNAs [211]. Both classes have a gene regulatory function like miRNAs [212, 213], and they were also analyzed for discrimination between case and control in several human cancer types. Thereby, Y RNAs are associated with deregulation in cancer (e.g. brain, bladder, blood, lung and postate cancer) [214]. Regarding tRNAs, the tRNA-derived small RNA fragments (tsRNAs, tsRF) identified in 2008 [215] came into the focus of clinical research, e.g. to be investigated as potential discriminator in breast cancer [216]. In comparison to the well studied miRNAs, other small non-coding RNAs (tRNA, Y RNA, snoRNAs, etc.) become increasingly well understood. The combined usage of markers derived from different RNA classes and families could advance research in cancer diagnostics and pathways by addressing multiple layers instead of limiting an analysis to a single sncRNA-class [217].

Alternative miRNA-based biomarkers Another advantage of sequencing is that base-modifications can be determined. These are typical for isomiRs [81], which were explained in the introduction of this theses (see Section 1.2.2). For RNAs in general, epitranscriptional modifications (e.g. 8-oxoguanine, oxidation of Guanin, o^8G) were observed [218]. In miRNAs, oxidation occurs mainly in the seed region (positions 2-9). For example, miR-1 has modifications on position 7 (oxidation to o^8G or substitution with Uracil) which happen during pathogenesis of cardiac hypertrophy in rat and mice models [219]. These position-specific modifications can be potential indicators for diseases and expand the options for diagnosis with miRNAs.

siRNA therapeutics - A fast growing RNA therapy In the previous paragraphs we list endogenously encoded sncRNA molecules. In the following, we outline the potential of exogenous sequences for therapeutics. In general, RNA therapies can be divided into three groups [220]: RNAs targeting RNA or DNA, RNAs targeting proteins and RNAs encoding proteins. We focus here only on the first category whose components (miRNAs and siRNAs) are part of the RNA interference (RNAi) system. Like miRNAs, siRNAs regulate gene activity and prevent mRNA translation by degradation [221]. While miRNA therapeutics are elaborated separately in the next paragraph, we consider here only siRNA therapeutics with which researchers knockdown genes that are causing the corresponding disease. In 2018, 2019 and 2020, the United States Food and Drug Adminstration (FDA) approved the siRNA drugs Patisiran, Givosiran and Lumasiran as the first RNAi-based pharmaceuticals. Patisiran is used for therapy of hereditary amyloidogenic transthyretin, Givosiran for treatment of adults with acute hepatic porphyria and Lumasiran for treatment of primary hyperoxaluria type 1 [222, 223]. Currently, siRNA drug development has over 30 therapeutical candidates (at different stages of clinical trials) and three approved therapeutic agents – which are much more successful compared to the miRNA therapeutics with several terminated candidates and no approvals [223, 224]. One important reason could be that each siRNA targets exactly one molecule which makes it favourable for single gene disorders, while miRNAs usually have multiple targets and form complex networks which are associated with multi-gene diseases such as human cancers or neurodegenerative diseases [225].

miRNA therapeutics Regarding miRNA therapeutics, researchers work on influencing the expression level of a miRNA. If the observed level of a miRNA does not meet the expected or desired measure, an up- or down-regulation the corresponding miRNA could be a solution. For down-regulation, anti-miRs are used to block the overexpressed miRNA so that it cannot bind to its target (gain of target function). In contrast, miRNA mimics bind to the same target of a down-regulated miRNA to increase its regulatory effect which (loss of target function) [226, 227]. Successfully developed exogenous regulation can become an important achievement in precision medicine, especially in cancer treatment to replace chemotherapy and its range of adverse effects. Nevertheless, the development of miRNA therapeutics remain difficult. To pass the different phases of the clinical trials, the revealed obstacles such as RNA instability, strong side effects or failed drug delivery to the target cells need to be overcome [228, 229]. The pharmaceutical agent which is most advanced in miRNA therapeutics is Miravirsen, currently under phase II clinical trials. This anti-miR down-regulates miR-122 for the treatment of hepatitis C virus infection [230, 231]. Although there is no miRNA-based drug approved yet, the growing market for miRNA therapeutics is relevant for human complex multi-target disorders such as different types of cancers and neurodegenerative diseases [224, 231]. In addition, it is also promising that three siRNA drugs for RNAi are already approved (see the previous paragraph about RNA therapy).

Synthesized biomarkers The discovery of natural-born biomarkers can be difficult due to technical, biological and stability reasons. Depending on which sample collection device is used, the expression of one miRNA can show different patterns, e.g. the found miRNA is up-regulated in serum and down-regulated in plasma of the same patients [232]. In addition, expression profiles can be influenced by population-related characteristics such as gender or age [233]. Another challenge of naturally occurring biomarkers is the RNA instability as result of the rapid degradation [234]. Kwong et al. showed that synthetic biomarkers can be an alternative to endogenous ones in mouse models. The disease-deregulated proteases cleave the substrates of the nanoparticles. These substrates can then be detected in urine by mass spectrometry [235]. The advantage of this bioengineering approach is the direct detection of the interaction between disease related components and the nanoscale agents. Thereby, a highly sensitive measurement could also reflect the disease state, which could be important for clinical therapy decisions. In summary, this approach can be a promising alternative option to overcome technical or population-driven effects on endogenous biomarker expressions.

Home-sampling In the clinical routine for screening studies, low-dose computed tomography (CT) and typical sampling devices (e.g. PAXgene or EDTA tubes) are part of the extended medical equipment to gauge a patient's health. Both CT and sampling tube methods are fairly inaccessible: both require trained medical professionals and CT devices are not available in every medical location. Ignoring the technological and personnel challenges, participation rates for screening (approaches) are lower due to the anxiety and reluctance of going to a doctor's office. Home-sampling already plays an increasing role in rural areas or in patients with periodic reading of blood or medication levels [236, 237]. It may also become an important way to raise the participation rates in different screening programs and ease the screening process itself. In the gynecological field of cervical screening, there have already been promising results [238]. That opens the way to expand the home-sampling process to other diseases. Using different blood tubes and measuring methods, such as high-throughput analysis or array measurement, can lead to different and even discordant profiling [239, 240]. As home-sampling becomes more and more useful and as miRNAs become more important players as molecular biomarkers, obtaining reproducible results is still a challenge for the future. We could already show promising results with dried blood spots (DBS) for diagnosis of disease in newborns. The sample collection itself can be quite challenging by squeezing blood drops from the fingers, which can lead to different volumes of the blood fluid. As their handling is rather complicated for a home-sampling process, the usage of microsampling devices (Mitra) can be an alternative. Aside from the easy handling, the main advantage of the Mitra is the constant blood volume per sample obtained with correct application. A further study to evaluate their potential as home-sampling tools seems to be reasonable and can be carried out like our project with DBS. In the suggested experiment, each patient would donate two samples in the form of one PAXgene tube as reference and of one Mitra as test device.

4.3 Conclusion

Because miRNAs regulate essential processes in cell development, are involved in immune system activities and are linked to tumorigenesis, the circulating cellular and cell-free fraction of these regulators has attracted considerable interest in research on blood-based biomarkers. Here, in this thesis, I have presented meaningful results toward the potential clinical use of miRNAs. Our study for lung cancer detection in a large cohort can improve strategies for early-stage detection. If further validated, it is conceivable that such signatures could either be used to "rule-in" participants for better adherence to present screening programs (e.g. LD-CT screening in US). Alternatively, the test could be used to inform on the malignancy of pulmonary nodules, especially in cases where it is difficult or dangerous to perform a biopsy or surgical procedure ("rule-out"). However, the current retrospective design of our work warrants additional validation in prospective cohorts collected in the asymptomatic screening setting before the optimal use of the test could be firmly established.

Regarding alternative sampling devices such as DBS and Mitra, our findings support to embark on new ways to simplify patient management for research on blood-based diagnostics of non-acute health questions. It is here that small RNAs offer a distinct stability advantage over long mRNAs. During emergency situations such as the current Covid-19 pandemic, the latter may enable home sampling in times of social distancing and or mandatory quarantine, as well as of offering inclusion in healthcare programs for those with restricted mobility.

Bibliography

[1] Kahraman M, Laufer T, Backes C, Schrörs H, Fehlmann T, Ludwig N, Kohlhaas J, Meese E, Wehler T, Bals R, Keller A (2017) Technical Stability and Biological Variability in MicroRNAs from Dried Blood Spots: A Lung Cancer Therapy-Monitoring Showcase. *Clinical Chemistry*, 63:1476–1488

[2] Fehlmann T, Kahraman M, Ludwig N, Backes C, Galata V, Keller V, Geffers L, Mercaldo N, Hornung D, Weis T, Kayvanpour E, Abu-Halima M, Deuschle C, Schulte C, Suenkel U, von Thaler AK, Maetzler W, Herr C, Fähndrich S, Vogelmeier C, Guimaraes P, Hecksteden A, Meyer T, Metzger F, Diener C, Deutscher S, Abdul-Khaliq H, Stehle I, Haeusler S, Meiser A, Groesdonk HV, Volk T, Lenhof HP, Katus H, Balling R, Meder B, Kruger R, Huwer H, Bals R, Meese E, Keller A (2020) Evaluating the Use of Circulating MicroRNA Profiles for Lung Cancer Detection in Symptomatic Patients. *JAMA oncology*, 6:714–723

[3] Fehlmann T, Backes C, Kahraman M, Haas J, Ludwig N, Posch AE, Würstle ML, Hübenthal M, Franke A, Meder B, Meese E, Keller A (2017) Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic Acids Research*, 45:8731–8744

[4] Fehlmann T, Backes C, Alles J, Fischer U, Hart M, Kern F, Langseth H, Rounge T, Umu SU, Kahraman M, Laufer T, Haas J, Staehler C, Ludwig N, Hübenthal M, Meder B, Franke A, Lenhof HP, Meese E, Keller A (2018) A high-resolution map of the human small non-coding transcriptome. *Bioinformatics* (*Oxford, England*), 34:1621–1628

[5] Fehlmann T, Backes C, Pirritano M, Laufer T, Galata V, Kern F, Kahraman M, Gasparoni G, Ludwig N, Lenhof HP, Gregersen HA, Francke R, Meese E, Simon M, Keller A (2019) The sncRNA Zoo: a repository for circulating small noncoding RNAs in animals. *Nucleic Acids Research*, 47:4431–4441

[6] Ludwig N, Hecksteden A, Kahraman M, Fehlmann T, Laufer T, Kern F, Meyer T, Meese E, Keller A, Backes C (2019) Spring is in the air: seasonal profiles indicate vernal change of miRNA activity. *RNA Biology*, 16:1034–1043

[7] Abu-Halima M, Kahraman M, Henn D, Rädle-Hurst T, Keller A, Abdul-Khaliq H, Meese E (2018) Deregulated microRNA and mRNA expression profiles in

the peripheral blood of patients with Marfan syndrome. *Journal of Translational Medicine*, 16:60

[8] Ha TY (2011) MicroRNAs in Human Diseases: From Cancer to Cardiovascular Disease. *Immune Network*, 11:135–154

[9] De Sousa VML, Carvalho L (2018) Heterogeneity in Lung Cancer. *Pathobiology: Journal of Immunopathology, Molecular and Cellular Biology*, 85:96–107

[10] Arab A, Karimipoor M, Irani S, Kiani A, Zeinali S, Tafsiri E, Sheikhy K (2017) Potential circulating miRNA signature for early detection of NSCLC. *Cancer Genetics*, 216-217:150–158

[11] Ellis PM, Vandermeer R (2011) Delays in the diagnosis of lung cancer. *Journal of Thoracic Disease*, 3:183–188

[12] Gray JR, Davies SJ (1996) Marfan syndrome. *Journal of Medical Genetics*, 33:403–408

[13] Spiro SG, Silvestri GA (2005) One Hundred Years of Lung Cancer. *American Journal of Respiratory and Critical Care Medicine*, 172:523–529

[14] World Health Organization (WHO). Cancer. URL: `https://www.who.int/news-room/fact-sheets/detail/cancer` (visited on 07/03/2021)

[15] Proctor RN (2012) The history of the discovery of the cigarette–lung cancer link: evidentiary traditions, corporate denial, global toll. *Tobacco Control*, 21:87–91

[16] History of the Surgeon General's Report on Smoking and Health | CDC, 2019

[17] Witschi H (2001) A short history of lung cancer. *Toxicological Sciences: An Official Journal of the Society of Toxicology*, 64:4–6

[18] Malhotra J, Malvezzi M, Negri E, La Vecchia C, Boffetta P (2016) Risk factors for lung cancer worldwide. *The European Respiratory Journal*, 48:889–902

[19] Keller A, Fehlmann T, Ludwig N, Kahraman M, Laufer T, Backes C, Vogelmeier C, Diener C, Biertz F, Herr C, Jörres RA, Lenhof HP, Meese E, Bals R, COSYCONET Study Group (2018) Genome-wide MicroRNA Expression Profiles in COPD: Early Predictors for Cancer Development. *Genomics, Proteomics & Bioinformatics*, 16:162–171

[20] National Cancer Institute. Cancer of the Lung and Bronchus - Cancer Stat Facts. URL: `https://seer.cancer.gov/statfacts/html/lungb.html` (visited on 07/03/2021)

[21] Moen TR, Chen B, Holmes DR, Duan X, Yu Z, Yu L, Leng S, Fletcher JG, McCollough CH (2021) Low-dose CT image and projection dataset. *Medical Physics*, 48:902–911

[22] Chassagnon G, Revel MP (2016) Lung cancer screening: Current status and perspective. *Diagnostic and Interventional Imaging*, 97:949–953

[23] National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England Journal of Medicine*, 365:395–409

[24] Hoffmann H, Heußel CP, Eichhorn M (2017) Screening for Lung Cancer: Current Status. *Zentralblatt Fur Chirurgie*, 142:S11–S16

[25] Bratulic S, Gatto F, Nielsen J (2019) The Translational Status of Cancer Liquid Biopsies. *Regenerative Engineering and Translational Medicine*

[26] Hyun KA, Gwak H, Lee J, Kwak B, Jung HI (2018) Salivary Exosome and Cell-Free DNA for Cancer Detection. *Micromachines*, 9:340

[27] Yu L, Shen J, Mannoor K, Guarnera M, Jiang F (2014) Identification of ENO1 As a Potential Sputum Biomarker for Early-Stage Lung Cancer by Shotgun Proteomics. *Clinical Lung Cancer*, 15:372–378.e1

[28] López-Sánchez LM, Jurado-Gámez B, Feu-Collado N, Valverde A, Cañas A, Fernández-Rueda JL, Aranda E, Rodríguez-Ariza A (2017) Exhaled breath condensate biomarkers for the early diagnosis of lung cancer using proteomics. *American Journal of Physiology. Lung Cellular and Molecular Physiology*, 313:L664–L676

[29] Miller MC, Doyle GV, Terstappen LWMM (2010) Significance of Circulating Tumor Cells Detected by the CellSearch System in Patients with Metastatic Breast Colorectal and Prostate Cancer. *Journal of Oncology*, 2010:617421

[30] Marquette CH, Boutros J, Benzaquen J, Ferreira M, Pastre J, Pison C, Padovani B, Bettayeb F, Fallet V, Guibert N, Basille D, Ilie M, Hofman V, Hofman P, AIR project Study Group (2020) Circulating tumour cells as a potential biomarker for lung cancer screening: a prospective cohort study. *The Lancet. Respiratory Medicine*, 8:709–716

[31] Alix-Panabières C, Schwarzenbach H, Pantel K (2012) Circulating tumor cells and circulating tumor DNA. *Annual Review of Medicine*, 63:199–215

[32] Chen K, Zhao H, Shi Y, Yang F, Wang LT, Kang G, Nie Y, Wang J (2019) Perioperative Dynamic Changes in Circulating Tumor DNA in Patients with Lung Cancer (DYNAMIC). *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 25:7058–7067

[33] Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, Douville C, Javed AA, Wong F, Mattox A, Hruban RH, Wolfgang CL, Goggins MG, Molin MD, Wang TL, Roden R, Klein AP, Ptak J, Dobbyn L, Schaefer J, Silliman N, Popoli M, Vogelstein JT, Browne JD, Schoen RE, Brand RE, Tie J, Gibbs P, Wong HL, Mansfield AS, Jen J, Hanash SM, Falconi M, Allen PJ, Zhou S, Bettegowda C, Diaz LA, Tomasetti C, Kinzler KW, Vogelstein B, Lennon AM, Papadopoulos N (2018) Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science (New York, N.Y.)*, 359:926–930

[34] Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV, CCGA Consortium (2020) Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, 31:745–759

[35] Roser AE, Caldi Gomes L, Schünemann J, Maass F, Lingor P (2018) Circulating miRNAs as Diagnostic Biomarkers for Parkinson's Disease. *Frontiers in Neuroscience*, 12:625

[36] Peng Y, Croce CM (2016) The role of MicroRNAs in human cancer. *Signal Transduction and Targeted Therapy*, 1:1–9

[37] Keller A, Backes C, Haas J, Leidinger P, Maetzler W, Deuschle C, Berg D, Ruschil C, Galata V, Ruprecht K, Stähler C, Würstle M, Sickert D, Gogol M, Meder B, Meese E (2016) Validating Alzheimer's disease micro RNAs using next-generation sequencing. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 12:565–576

[38] Colpaert RMW, Calore M (2019) MicroRNAs in Cardiac Diseases. *Cells*, 8:E737

[39] Jones GS, Baldwin DR (2018) Recent advances in the management of lung cancer. *Clinical Medicine*, 18:s41–s46

[40] Hirashima DE, Soriano ES, Meirelles RL, Alberti GN, Nosé W (2010) Outcomes of Iris-Claw Anterior Chamber versus Iris-Fixated Foldable Intraocular Lens in Subluxated Lens Secondary to Marfan Syndrome. *Ophthalmology*, 117:1479–1485

[41] Tsang AKL, Taverne A, Holcombe T (2013) Marfan syndrome: a review of the literature and case report. *Special Care in Dentistry*, 33:248–254

[42] Kainulainen K, Pulkkinen L, Savolainen A, Kaitila I, Peltonen L (1990) Location on chromosome 15 of the gene defect causing Marfan syndrome. *The New England Journal of Medicine*, 323:935–939

[43] Dietz HC, Cutting CR, Pyeritz RE, Maslen CL, Sakai LY, Corson GM, Puffenberger EG, Hamosh A, Nanthakumar EJ, Curristin SM, Stetten G, Meyers DA, Francomano CA (1991) Marfan syndrome caused by a recurrent de novo missense mutation in the fibrillin gene. *Nature*, 352:337–339

[44] Judge DP, Dietz HC (2005) Marfan's syndrome. *Lancet*, 366:1965–1976

[45] Bolar N, Van Laer L, Loeys BL (2012) Marfan syndrome: from gene to therapy. *Current Opinion in Pediatrics*, 24:498–504

[46] Verstraeten A, Alaerts M, Laer LV, Loeys B (2016) Marfan Syndrome and Related Disorders: 25 Years of Gene Discovery. *Human Mutation*, 37:524–531

[47] De Paepe A, Devereux RB, Dietz HC, Hennekam RC, Pyeritz RE (1996) Revised diagnostic criteria for the Marfan syndrome. *American Journal of Medical Genetics*, 62:417–426

[48] Loeys BL, Dietz HC, Braverman AC, Callewaert BL, Backer JD, Devereux RB, Hilhorst-Hofstee Y, Jondeau G, Faivre L, Milewicz DM, Pyeritz RE, Sponseller PD, Wordsworth P, Paepe AMD (2010) The revised Ghent nosology for the Marfan syndrome. *Journal of Medical Genetics*, 47:476–485

[49] Vaidyanathan B (2008) Role of beta-blockers in Marfan's syndrome and bicuspid aortic valve: A time for re-appraisal. *Annals of Pediatric Cardiology*, 1:149–152

[50] Lumban Tobing SDA, Akbar DL (2020) Challenges and experiences in correcting scoliosis of a patient with Marfan Syndrome: A case report. *International Journal of Surgery Case Reports*, 76:85–89

[51] Silverman DI, Burton KJ, Gray J, Bosner MS, Kouchoukos NT, Roman MJ, Boxer M, Devereux RB, Tsipouras P (1995) Life expectancy in the Marfan syndrome. *The American Journal of Cardiology*, 75:157–160

[52] Phylactou LA, Kilpatrick MW (1999) Potential therapy paradigms for Marfan syndrome. *Expert Opinion on Investigational Drugs*, 8:983–993

[53] Zeng Y, Li J, Li G, Huang S, Yu W, Zhang Y, Chen D, Chen J, Liu J, Huang X (2018) Correction of the Marfan Syndrome Pathogenic FBN1 Mutation by Base Editing in Human Cells and Heterozygous Embryos. *Molecular Therapy*, 26:2631–2637

[54] Carroll D (2017) Genome Editing: Past, Present, and Future. *The Yale Journal of Biology and Medicine*, 90:653–659

[55] Li H, Yang Y, Hong W, Huang M, Wu M, Zhao X (2020) Applications of genome editing technology in the targeted therapy of human diseases: mechanisms, advances and prospects. *Signal Transduction and Targeted Therapy*, 5:1–23

[56] Kubo T, Ohno Y, Takenaka D, Nishino M, Gautam S, Sugimura K, Kauczor HU, Hatabu H (2016) Standard-dose vs. low-dose CT protocols in the evaluation of localized lung lesions: Capability for lesion characterization—iLEAD study. *European Journal of Radiology Open*, 3:67–73

[57] Han J, Chen M, Wang Y, Gong B, Zhuang T, Liang L, Qiao H (2018) Identification of Biomarkers Based on Differentially Expressed Genes in Papillary Thyroid Carcinoma. *Scientific Reports*, 8:9912

[58] Aronson JK, Ferner RE (2017) Biomarkers—A General Review. *Current Protocols in Pharmacology*, 76:9.23.1–9.23.17

[59] Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75:843–854

[60] Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403:901–906

[61] Slack FJ (2006) Regulatory RNAs and the demise of 'junk' DNA. *Genome Biology*, 7:328

[62] Ambros V (2001) microRNAs: Tiny Regulators with Great Potential. *Cell*, 107:823–826

[63] Kowalczyk MS, Higgs DR, Gingeras TR (2012) RNA discrimination. *Nature*, 482:310–311

[64] Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biology*, 7:S4

[65] Bartel DP (2004) MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, 116:281–297

[66] Lewis BP, Burge CB, Bartel DP (2005) Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell*, 120:15–20

[67] Akhtar MM, Micolucci L, Islam MS, Olivieri F, Procopio AD (2016) Bioinformatic tools for microRNA dissection. *Nucleic Acids Research*, 44:24–44

[68] Huang W (2017) MicroRNAs: Biomarkers, Diagnostics, and Therapeutics. *Methods in Molecular Biology (Clifton, N.J.)*, 1617:57–67

[69] Winter J, Jung S, Keller S, Gregory RI, Diederichs S (2009) Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nature Cell Biology*, 11:228–234

[70] Ha M, Kim VN (2014) Regulation of microRNA biogenesis. *Nature Reviews Molecular Cell Biology*, 15:509–524

[71] Fang W, Bartel DP (2015) The menu of features that define primary microRNAs and enable de novo design of microRNA genes. *Molecular cell*, 60:131–145

[72] O'Brien J, Hayder H, Zayed Y, Peng C (2018) Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Frontiers in Endocrinology*, 0

[73] Zhang H, Kolb FA, Jaskiewicz L, Westhof E, Filipowicz W (2004) Single Processing Center Models for Human Dicer and Bacterial RNase III. *Cell*, 118:57–68

[74] Wostenberg C, Lary JW, Sahu D, Acevedo R, Quarles KA, Cole JL, Showalter SA (2012) The Role of Human Dicer-dsRBD in Processing Small Regulatory RNAs. *PLOS ONE*, 7:e51829

[75] Fabian MR, Sonenberg N (2012) The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nature Structural & Molecular Biology*, 19:586–593

[76] Zhang X, Zeng Y (2010) The terminal loop region controls microRNA processing by Drosha and Dicer. *Nucleic Acids Research*, 38:7689–7697

[77] Trotta E (2014) On the Normalization of the Minimum Free Energy of RNAs by Sequence Length. *PLOS ONE*, 9:e113380

[78] Wu JC, Gardner DP, Ozer S, Gutell RR, Ren P (2009) Correlation of RNA Secondary Structure Statistics with Thermodynamic Stability and Applications to Folding. *Journal of molecular biology*, 391:769–783

[79] Thakur V, Wanchana S, Xu M, Bruskiewich R, Quick WP, Mosig A, Zhu XG (2011) Characterization of statistical features for plant microRNA prediction. *BMC Genomics*, 12:108

[80] Lee LW, Zhang S, Etheridge A, Ma L, Martin D, Galas D, Wang K (2010) Complexity of the microRNA repertoire revealed by next-generation sequencing. *RNA*, 16:2170–2180

[81] Wu CW, Evans JM, Huang S, Mahoney DW, Dukek BA, Taylor WR, Yab TC, Smyrk TC, Jen J, Kisiel JB, Ahlquist DA (2018) A Comprehensive Approach to Sequence-oriented IsomiR annotation (CASMIR): demonstration with IsomiR profiling in colorectal neoplasia. *BMC Genomics*, 19:401

[82] Wu H, Neilson JR, Kumar P, Manocha M, Shankar P, Sharp PA, Manjunath N (2007) miRNA profiling of naïve, effector and memory CD8 T cells. *PloS One*, 2:e1020

[83] Cloonan N, Wani S, Xu Q, Gu J, Lea K, Heater S, Barbacioru C, Steptoe AL, Martin HC, Nourbakhsh E, Krishnan K, Gardiner B, Wang X, Nones K, Steen JA, Matigian NA, Wood DL, Kassahn KS, Waddell N, Shepherd J, Lee C, Ichikawa J, McKernan K, Bramlett K, Kuersten S, Grimmond SM (2011) MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biology*, 12:R126

[84] Loher P, Londin ER, Rigoutsos I (2014) IsomiR expression profiles in human lymphoblastoid cell lines exhibit population and gender dependencies. *Oncotarget*, 5:8790–8802

[85] Telonis AG, Magee R, Loher P, Chervoneva I, Londin E, Rigoutsos I (2017) Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. *Nucleic Acids Research*, 45:2973–2985

[86] Telonis AG, Loher P, Jing Y, Londin E, Rigoutsos I (2015) Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity. *Nucleic Acids Research*, 43:9158–9175

[87] Juzenas S, Venkatesh G, Hübenthal M, Hoeppner MP, Du ZG, Paulsen M, Rosenstiel P, Senger P, Hofmann-Apitius M, Keller A, Kupcinskas L, Franke A, Hemmrich-Stanisak G (2017) A comprehensive, cell specific microRNA catalogue of human peripheral blood. *Nucleic Acids Research*, 45:9290–9301

[88] Jovanovic M, Hengartner MO (2006) miRNAs and apoptosis: RNAs to die for. *Oncogene*, 25:6176–6187

[89] Boyerinas B, Park SM, Hau A, Murmann AE, Peter ME (2010) The role of let-7 in cell differentiation and cancer. *Endocrine-Related Cancer*, 17:F19–F36

[90] Shcherbata HR, Ward EJ, Fischer KA, Yu JY, Reynolds SH, Chen CH, Xu P, Hay BA, Ruohola-Baker H (2007) Stage-specific differences in the requirements for germline stem cell maintenance in the Drosophila ovary. *Cell stem cell*, 1:698–709

[91] Lizé M, Klimke A, Dobbelstein M (2011) MicroRNA-449 in cell fate determination. *Cell Cycle*, 10:2874–2882

[92] Hashimoto Y, Akiyama Y, Yuasa Y (2013) Multiple-to-Multiple Relationships between MicroRNAs and Target Genes in Gastric Cancer. *PLOS ONE*, 8:e62589

[93] Iwakawa Ho, Tomari Y (2015) The Functions of MicroRNAs: mRNA Decay and Translational Repression. *Trends in Cell Biology*, 25:651–665

[94] Djuranovic S, Nahvi A, Green R (2012) miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science (New York, N.Y.)*, 336:237–240

[95] Ruike Y, Ichimura A, Tsuchiya S, Shimizu K, Kunimoto R, Okuno Y, Tsujimoto G (2008) Global correlation analysis for micro-RNA and mRNA expression profiles in human cell lines. *Journal of Human Genetics*, 53:515–523

[96] Gu Y, Ampofo E, Menger MD, Laschke MW (2017) miR-191 suppresses angiogenesis by activation of NF-κB signaling. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 31:3321–3333

[97] Tan H, Huang S, Zhang Z, Qian X, Sun P, Zhou X (2019) Pan-cancer analysis on microRNA-associated gene activation. *EBioMedicine*, 43:82–97

[98] Xu P, Wu Q, Yu J, Rao Y, Kou Z, Fang G, Shi X, Liu W, Han H (2020) A Systematic Way to Infer the Regulation Relations of miRNAs on Target Genes and Critical miRNAs in Cancers. *Frontiers in Genetics*, 11:278

[99] Kuhn DE, Martin MM, Feldman DS, Terry AV, Nuovo GJ, Elton TS (2008) Experimental Validation of miRNA Targets. *Methods (San Diego, Calif.)*, 44:47–54

[100] Backes C, Meder B, Hart M, Ludwig N, Leidinger P, Vogel B, Galata V, Roth P, Menegatti J, Grässer F, Ruprecht K, Kahraman M, Grossmann T, Haas J, Meese E, Keller A (2016) Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic Acids Research*, 44:e53

[101] Backes C, Fehlmann T, Kern F, Kehl T, Lenhof HP, Meese E, Keller A (2018) miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Research*, 46:D160–D167

[102] Londin E, Loher P, Telonis AG, Quann K, Clark P, Jing Y, Hatzimichael E, Kirino Y, Honda S, Lally M, Ramratnam B, Comstock CES, Knudsen KE, Gomella L, Spaeth GL, Hark L, Katz LJ, Witkiewicz A, Rostami A, Jimenez SA, Hollingsworth MA, Yeh JJ, Shaw CA, McKenzie SE, Bray P, Nelson PT, Zupo S, Van Roosbroeck K, Keating MJ, Calin GA, Yeo C, Jimbo M, Cozzitorto J, Brody JR, Delgrosso K, Mattick JS, Fortina P, Rigoutsos I (2015) Analysis of 13 cell types reveals evidence for the

expression of numerous novel primate- and tissue-specific microRNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 112:E1106–1115

[103] Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellos I, Papadimitriou D, Kavakiotis I, Maniou S, Skoufos G, Vergoulis T, Dalamagas T, Hatzigeorgiou AG (2018) DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. *Nucleic Acids Research*, 46:D239–D245

[104] Huang HY, Lin YCD, Li J, Huang KY, Shrestha S, Hong HC, Tang Y, Chen YG, Jin CN, Yu Y, Xu JT, Li YM, Cai XX, Zhou ZY, Chen XH, Pei YY, Hu L, Su JJ, Cui SD, Wang F, Xie YY, Ding SY, Luo MF, Chou CH, Chang NW, Chen KW, Cheng YH, Wan XH, Hsu WL, Lee TY, Wei FX, Huang HD (2020) miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic Acids Research*, 48:D148–D154

[105] Dweep H, Gretz N (2015) miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nature Methods*, 12:697–697

[106] Kehl T, Kern F, Backes C, Fehlmann T, Stöckel D, Meese E, Lenhof HP, Keller A (2020) miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database. *Nucleic Acids Research*, 48:D142–D147

[107] Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T, Pallasch C, Rheinheimer S, Meder B, Stähler C, Meese E, Keller A (2016) Distribution of miRNA expression across human tissues. *Nucleic Acids Research*, 44:3865–3877

[108] Zhang W, Dahlberg JE, Tam W (2007) MicroRNAs in Tumorigenesis. *The American Journal of Pathology*, 171:728–738

[109] Uhr K, Smissen WJCPvd, Heine AAJ, Ozturk B, Jaarsveld MTMv, Boersma AWM, Jager A, Wiemer EAC, Smid M, Foekens JA, Martens JWM (2019) MicroRNAs as possible indicators of drug sensitivity in breast cancer cell lines. *PLOS ONE*, 14:e0216400

[110] Rosenfeld N, Aharonov R, Meiri E, Rosenwald S, Spector Y, Zepeniuk M, Benjamin H, Shabes N, Tabak S, Levy A, Lebanony D, Goren Y, Silberschein E, Targan N, Ben-Ari A, Gilad S, Sion-Vardy N, Tobar A, Feinmesser M, Kharenko O, Nativ O, Nass D, Perelman M, Yosepovich A, Shalmon B, Polak-Charcon S, Fridman E, Avniel A, Bentwich I, Bentwich Z, Cohen D, Chajut A, Barshack I (2008) MicroRNAs accurately identify cancer tissue origin. *Nature Biotechnology*, 26:462–469

[111] Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, Lin C, Socci ND, Hermida L, Fulci V, Chiaretti S, Foà R, Schliwka J, Fuchs U, Novosel A, Müller RU, Schermer B, Bissels U, Inman J, Phan Q, Chien M, Weir DB, Choksi R, Vita GD, Frezzetti D, Trompeter HI, Hornung V, Teng G, Hartmann G, Palkovits M, Lauro RD, Wernet P, Macino G, Rogler CE, Nagle JW, Ju J, Papavasiliou FN, Benzing T, Lichter P, Tam W, Brownstein MJ, Bosio A, Borkhardt A, Russo JJ, Sander C, Zavolan M, Tuschl T (2007) A

Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. *Cell*, 129:1401–1414

[112] Grasedieck S, Sorrentino A, Langer C, Buske C, Döhner H, Mertens D, Kuchenbauer F (2013) Circulating microRNAs in hematological diseases: principles, challenges, and perspectives. *Blood*, 121:4977–4984

[113] Fabbri M, Paone A, Calore F, Galli R, Gaudio E, Santhanam R, Lovat F, Fadda P, Mao C, Nuovo GJ, Zanesi N, Crawford M, Ozer GH, Wernicke D, Alder H, Caligiuri MA, Nana-Sinkam P, Perrotti D, Croce CM (2012) MicroRNAs bind to Toll-like receptors to induce prometastatic inflammatory response. *Proceedings of the National Academy of Sciences*, 109:E2110–E2116

[114] Tsitsiou E, Lindsay MA (2009) microRNAs and the immune response. *Current Opinion in Pharmacology*, 9:514–520

[115] Mehta A, Baltimore D (2016) MicroRNAs as regulatory elements in immune system logic. *Nature Reviews Immunology*, 16:279–294

[116] Kihm A, Kaestner L, Wagner C, Quint S (2018) Classification of red blood cell shapes in flow using outlier tolerant machine learning. *PLOS Computational Biology*, 14:e1006278

[117] Juzenas S, Lindqvist CM, Ito G, Dolshanskaya Y, Halfvarson J, Franke A, Hemmrich-Stanisak G (2020) Depletion of erythropoietic miR-486-5p and miR-451a improves detectability of rare microRNAs in peripheral blood-derived small RNA sequencing libraries. *NAR Genomics and Bioinformatics*, 2:lqaa008

[118] Jolliffe IT, Cadima J (2016) Principal component analysis: a review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 374:20150202

[119] Li W, Cerise JE, Yang Y, Han H (2017) Application of t-SNE to human genetic data. *Journal of Bioinformatics and Computational Biology*, 15:1750017

[120] McInnes L, Healy J, Saul N, Großberger L (2018) UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3:861

[121] McInnes L, Healy J, Melville J (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

[122] McLachlan GJ, Bean RW, Ng SK (2017) Clustering. *Methods in Molecular Biology (Clifton, N.J.)*, 1526:345–362

[123] Vetter TR (2017) Fundamentals of Research Data and Variables: The Devil Is in the Details. *Anesthesia and Analgesia*, 125:1375–1380

[124] Du Prel JB, Röhrig B, Hommel G, Blettner M (2010) Choosing statistical tests: part 12 of a series on evaluation of scientific publications. *Deutsches Arzteblatt International*, 107:343–348

[125] Brownie C, Boos DD (1994) Type I error robustness of ANOVA and ANOVA on ranks when the number of treatments is large. *Biometrics*, 50:542–549

[126] Chen X, Robinson DG, Storey JD (2021) The functional false discovery rate with applications to genomics. *Biostatistics (Oxford, England)*, 22:68–81

[127] Nayarisseri A, Khandelwal R, Tanwar P, Madhavi M, Sharma D, Thakur G, Speck-Planche A, Singh SK (2021) Artificial Intelligence, Big Data and Machine Learning Approaches in Precision Medicine & Drug Discovery. *Current Drug Targets*, 22:631–655

[128] Ambros V (2004) The functions of animal microRNAs. *Nature*, 431:350–355

[129] Riolo G, Cantara S, Marzocchi C, Ricci C (2020) miRNA Targets: From Prediction Tools to Experimental Validation. *Methods and Protocols*, 4:1

[130] Shirdel EA, Xie W, Mak TW, Jurisica I (2011) NAViGaTing the micronome–using multiple microRNA prediction databases to identify signalling pathway-associated microRNAs. *PloS One*, 6:e17429

[131] Kern F, Backes C, Hirsch P, Fehlmann T, Hart M, Meese E, Keller A (2020) What's the target: understanding two decades of in silico microRNA-target prediction. *Briefings in Bioinformatics*, 21:1999–2010

[132] Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS (2003) MicroRNA targets in Drosophila. *Genome Biology*, 5:R1

[133] Agarwal V, Bell GW, Nam JW, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4

[134] Krüger J, Rehmsmeier M (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Research*, 34:W451–454

[135] Koutrouli M, Karatzas E, Paez-Espino D, Pavlopoulos GA (2020) A Guide to Conquer the Biological Network Era Using Graph Theory. *Frontiers in Bioengineering and Biotechnology*, 8:34

[136] Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP (2003) The microRNAs of Caenorhabditis elegans. *Genes & Development*, 17:991–1008

[137] Lau NC, Lim LP, Weinstein EG, Bartel DP (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science (New York, N.Y.)*, 294:858–862

[138] Lee RC, Ambros V (2001) An extensive class of small RNAs in Caenorhabditis elegans. *Science (New York, N.Y.)*, 294:862–864

[139] Lai EC, Tomancak P, Williams RW, Rubin GM (2003) Computational identification of DrosophilamicroRNA genes. *Genome Biology*, 4:R42

[140] Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31:3406–3415

[141] Mathelier A, Carbone A (2010) MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics*, 26:2226–2234

[142] Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly*, 125:167–188

[143] Sewer A, Paul N, Landgraf P, Aravin A, Pfeffer S, Brownstein MJ, Tuschl T, van Nimwegen E, Zavolan M (2005) Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, 6:267

[144] Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Research*, 35:W339–W344

[145] Batuwita R, Palade V (2009) microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics (Oxford, England)*, 25:989–995

[146] Hackenberg M, Sturm M, Langenberger D, Falcón-Pérez JM, Aransay AM (2009) miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Research*, 37:W68–W76

[147] Aparicio-Puerta E, Lebrón R, Rueda A, Gómez-Martín C, Giannoukakos S, Jaspez D, Medina JM, Zubkovic A, Jurak I, Fromm B, Marchal JA, Oliver J, Hackenberg M (2019) sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression. *Nucleic Acids Research*, 47:W530–W535

[148] Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*, 40:37–52

[149] Hu G, Kurgan L (2019) Sequence Similarity Searching. *Current Protocols in Protein Science*, 95:e71

[150] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43:e47

[151] Cooper A, Reimann R, Cronin D, Noessel C (2014) About Face: The Essentials of Interaction Design. Wiley Publishing

[152] Balasubramanee V, Wimalasena C, Singh R, Pierce M (2013) Twitter bootstrap and AngularJS: Frontend frameworks to expedite science gateway development. *2013 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 1–1

[153] Kern F, Fehlmann T, Keller A (2020) On the lifetime of bioinformatics web services. *Nucleic Acids Research*, 48:12523–12533

[154] Keller A, Backes C, Leidinger P, Kefer N, Boisguerin V, Barbacioru C, Vogel B, Matzas M, Huwer H, Katus HA, Stähler C, Meder B, Meese E (2011) Next-generation sequencing identifies novel microRNAs in peripheral blood of lung cancer patients. *Molecular bioSystems*, 7:3187–3199

[155] Lopez JP, Diallo A, Cruceanu C, Fiori LM, Laboissiere S, Guillet I, Fontaine J, Ragoussis J, Benes V, Turecki G, Ernst C (2015) Biomarker discovery: quantification of microRNAs and other small non-coding RNAs using next generation sequencing. *BMC Medical Genomics*, 8:35

[156] Wolff A, Bayerlová M, Gaedcke J, Kube D, Beißbarth T (2018) A comparative study of RNA-Seq and microarray data analysis on the two examples of rectal-cancer patients and Burkitt Lymphoma cells. *PLOS ONE*, 13:e0197162

[157] Freitas FCP, Depintor TS, Agostini LT, Luna-Lucena D, Nunes FMF, Bitondi MMG, Simões ZLP, Lourenço AP (2019) Evaluation of reference genes for gene expression analysis by real-time quantitative PCR (qPCR) in three stingless bee species (Hymenoptera: Apidae: Meliponini). *Scientific Reports*, 9:17692

[158] Garrido J, Aguilar M, Prieto P (2020) Identification and validation of reference genes for RT-qPCR normalization in wheat meiosis. *Scientific Reports*, 10:2726

[159] Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17:333–351

[160] Miyamoto M, Motooka D, Gotoh K, Imai T, Yoshitake K, Goto N, Iida T, Yasunaga T, Horii T, Arakawa K, Kasahara M, Nakamura S (2014) Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics*, 15:699

[161] Slatko BE, Gardner AF, Ausubel FM (2018) Overview of Next Generation Sequencing Technologies. *Current protocols in molecular biology*, 122:e59

[162] Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D (2011) Systematic Comparison of Three Methods for Fragmentation of Long-Range PCR Products for Next Generation Sequencing. *PLOS ONE*, 6:e28240

[163] Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P (2014) Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, 56:61–passim

[164] Fehlmann T, Reinheimer S, Geng C, Su X, Drmanac S, Alexeev A, Zhang C, Backes C, Ludwig N, Hart M, An D, Zhu Z, Xu C, Chen A, Ni M, Liu J, Li Y, Poulter M, Li Y, Stähler C, Drmanac R, Xu X, Meese E, Keller A (2016) cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clinical Epigenetics*, 8:123

[165] Qin D (2019) Next-generation sequencing and its clinical application. *Cancer Biology & Medicine*, 16:4–10

[166] Morganti S, Tarantino P, Ferraro E, D'Amico P, Duso BA, Curigliano G (2019) Next Generation Sequencing (NGS): A Revolutionary Technology in Pharmacogenomics and Personalized Medicine in Cancer. *Advances in Experimental Medicine and Biology*, 1168:9–30

[167] Motameny S, Wolters S, Nürnberg P, Schumacher B (2010) Next Generation Sequencing of miRNAs – Strategies, Resources and Methods. *Genes*, 1:70–84

[168] Cho N, Hwang B, Yoon Jk, Park S, Lee J, Seo HN, Lee J, Huh S, Chung J, Bang D (2015) De novo assembly and next-generation sequencing to analyse full-length gene variants from codon-barcoded libraries. *Nature Communications*, 6:8351

[169] Kircher M, Kelso J (2010) High-throughput DNA sequencing – concepts and limitations. *BioEssays*, 32:524–536

[170] Chatterjee A, Leichter AL, Fan V, Tsai P, Purcell RV, Sullivan MJ, Eccles MR (2015) A cross comparison of technologies for the detection of microRNAs in clinical FFPE samples of hepatoblastoma patients. *Scientific Reports*, 5:10438

[171] nCounter Nanostring Analysis System

[172] Kulkarni MM (2011) Digital Multiplexed Gene Expression Analysis Using the NanoString nCounter System. *Current Protocols in Molecular Biology*, 94:25B.10.1–25B.10.17

[173] Bondar G, Xu W, Elashoff D, Li X, Faure-Kumar E, Bao TM, Grogan T, Moose J, Deng MC (2020) Comparing NGS and NanoString platforms in peripheral blood mononuclear cell transcriptome profiling for advanced heart failure biomarker development. *Journal of Biological Methods*, 7:e123

[174] Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell HP, Ferree S, George RD, Grogan T, James JJ, Maysuria M, Mitton JD, Oliveri P, Osborn JL, Peng T, Ratcliffe AL, Webster PJ, Davidson EH, Hood L, Dimitrov K (2008) Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature Biotechnology*, 26:317–325

[175] Bumgarner R (2013) DNA microarrays: Types, Applications and their future. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, 0 22:Unit–22.1.

[176] Miller MB, Tang YW (2009) Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology. *Clinical Microbiology Reviews*, 22:611–633

[177] Barczak A, Rodriguez MW, Hanspers K, Koth LL, Tai YC, Bolstad BM, Speed TP, Erle DJ (2003) Spotted Long Oligonucleotide Arrays for Human Gene Expression Analysis. *Genome Research*, 13:1775–1785

[178] Zahurak M, Parmigiani G, Yu W, Scharpf RB, Berman D, Schaeffer E, Shabbeer S, Cope L (2007) Pre-processing Agilent microarray data. *BMC Bioinformatics*, 8:142

[179] Lamy P, Grove J, Wiuf C (2011) A review of software for microarray genotyping. *Human Genomics*, 5:304–309

[180] Reimers M, Weinstein JN (2005) Quality assessment of microarrays: Visualization of spatial artifacts and quantitation of regional biases. *BMC Bioinformatics*, 6:166

[181] Koshy L, Anju AL, Harikrishnan S, Kutty VR, Jissa VT, Kurikesu I, Jayachandran P, Jayakumaran Nair A, Gangaprasad A, Nair GM, Sudhakaran PR (2017) Evaluating genomic DNA extraction methods from human whole blood using endpoint and real-time PCR assays. *Molecular Biology Reports*, 44:97–108

[182] Wong ML, Medrano JF (2005) Real-time PCR for mRNA quantitation. *BioTechniques*, 39:75–85

[183] Narrandes S, Xu W (2018) Gene Expression Detection Assay for Cancer Clinical Use. *Journal of Cancer*, 9:2249–2265

[184] RAYMOND CK, ROBERTS BS, GARRETT-ENGELE P, LIM LP, JOHNSON JM (2005) Simple, quantitative primer-extension PCR assay for direct monitoring of microRNAs and short-interfering RNAs. *RNA*, 11:1737–1744

[185] Andersen CL, Jensen JL, Ørntoft TF (2004) Normalization of Real-Time Quantitative Reverse Transcription-PCR Data: A Model-Based Variance Estimation Approach to Identify Genes Suited for Normalization, Applied to Bladder and Colon Cancer Data Sets. *Cancer Research*, 64:5245–5250

[186] Karlen Y, McNair A, Perseguers S, Mazza C, Mermod N (2007) Statistical significance of quantitative PCR. *BMC Bioinformatics*, 8:131

[187] Chiu YW, Kao YH, Simoff MJ, Ost DE, Wagner O, Lavin J, Culbertson RA, Smith DG (2021) Costs of Biopsy and Complications in Patients with Lung Cancer. *ClinicoEconomics and Outcomes Research: CEOR*, 13:191–200

[188] Maleki F, Ovens K, McQuillan I, Kusalik AJ (2019) Size matters: how sample size affects the reproducibility and specificity of gene set analysis. *Human Genomics*, 13:42

[189] Carrillo-Ávila JA, de la Puente R, Catalina P, Rejón JD, Espín-Vallejo L, Valdivieso V, Aguilar-Quesada R (2020) Evaluation of RNA purification methods by using different blood stabilization tubes: identification of key features for epidemiological studies. *BMC Research Notes*, 13:77

[190] Macfarlane DE, Dahle CE (1997) Isolating RNA from clinical samples with Catrimox-14 and lithium chloride. *Journal of Clinical Laboratory Analysis*, 11:132–139

[191] Grüner N, Stambouli O, Ross RS (2015) Dried Blood Spots - Preparing and Processing for Use in Immunoassays and in Molecular Techniques. *Journal of Visualized Experiments : JoVE*:52619

[192] CHAPMAN OD (1924) THE COMPLEMENT-FIXATION TEST FOR SYPHILIS: USE OF PATIENT'S WHOLE BLOOD DRIED ON FILTER PAPER. *Archives of Dermatology and Syphilology*, 9:607–611

[193] Demirev PA (2013) Dried Blood Spots: Analysis and Applications. *Analytical Chemistry*, 85:779–789

[194] Jurado C (2013) Blood. Siegel JA, Saukko PJ, Houck MM, editors, *Encyclopedia of Forensic Sciences (Second Edition)*, pages 336–342: Academic Press, Waltham

[195] Segundo GRS, Nguyen ATV, Thuc HT, Nguyen LNQ, Kobayashi RH, Le HT, Le HTM, Torgerson TR, Ochs HD (2018) Dried Blood Spots, an Affordable Tool to Collect, Ship, and Sequence gDNA from Patients with an X-Linked Agammaglobulinemia Phenotype Residing in a Developing Country. *Frontiers in Immunology*, 9:289

[196] Protti M, Mandrioli R, Mercolini L (2019) Tutorial: Volumetric absorptive microsampling (VAMS). *Analytica Chimica Acta*, 1046:32–47

[197] Harahap Y, Diptasaadya R, Purwanto DJ (2020) Volumetric Absorptive Microsampling as a Sampling Alternative in Clinical Trials and Therapeutic Drug Monitoring During the COVID-19 Pandemic: A Review. *Drug Design, Development and Therapy*, 14:5757–5771

[198] Prinsenberg T, Rebers S, Boyd A, Zuure F, Prins M, Valk Mvd, Schinkel J (2020) Dried blood spot self-sampling at home is a feasible technique for hepatitis C RNA detection. *PLOS ONE*, 15:e0231385

[199] Rudge J, Kushon S () Volumetric absorptive microsampling: its use in COVID-19 research and testing. *Bioanalysis*:10.4155/bio–2021–0102

[200] Kahraman M, Röske A, Laufer T, Fehlmann T, Backes C, Kern F, Kohlhaas J, Schrörs H, Saiz A, Zabler C, Ludwig N, Fasching PA, Strick R, Rübner M, Beckmann MW, Meese E, Keller A, Schrauder MG (2018) MicroRNA in diagnosis and therapy monitoring of early-stage triple-negative breast cancer. *Scientific Reports*, 8:11584

[201] Afolabi LT, Saeed F, Hashim H, Petinrin OO (2018) Ensemble learning method for the prediction of new bioactive molecules. *PLoS ONE*, 13:e0189538

[202] Mehta P, Wang CH, Day AGR, Richardson C, Bukov M, Fisher CK, Schwab DJ (2019) A high-bias, low-variance introduction to Machine Learning for physicists. *Physics reports*, 810:1–124

[203] Zheng X, Fu X, Wang K, Wang M (2020) Deep neural networks for human microRNA precursor detection. *BMC Bioinformatics*, 21:17

[204] Condrat CE, Thompson DC, Barbu MG, Bugnar OL, Boboc A, Cretoiu D, Suciu N, Cretoiu SM, Voinea SC (2020) miRNAs as Biomarkers in Disease: Latest Findings Regarding Their Role in Diagnosis and Prognosis. *Cells*, 9:276

[205] Ameling S, Kacprowski T, Chilukoti RK, Malsch C, Liebscher V, Suhre K, Pietzner M, Friedrich N, Homuth G, Hammer E, Völker U (2015) Associations of circulating plasma microRNAs with age, body mass index and sex in a population-based study. *BMC Medical Genomics*, 8:61

[206] Takahashi K, Yokota SI, Tatsumi N, Fukami T, Yokoi T, Nakajima M (2013) Cigarette smoking substantially alters plasma microRNA profiles in healthy subjects. *Toxicology and Applied Pharmacology*, 272:154–160

[207] Skogholt AH, Ryeng E, Erlandsen SE, Skorpen F, Schønberg SA, Sætrom P (2017) Gene expression differences between PAXgene and Tempus blood RNA tubes are highly reproducible between independent samples and biobanks. *BMC Research Notes*, 10:136

[208] Heidel RE (2016) Causality in Statistical Power: Isomorphic Properties of Measurement, Research Design, Effect Size, and Sample Size. *Scientifica*, 2016:8920418

[209] Hippen AA, Greene CS (2021) Expanding and Remixing the Metadata Landscape. *Trends in Cancer*, 7:276–278

[210] Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carré C, Burdin N, Visan L, Ceccarelli M, Poidinger M, Zippelius A, Pedro de Magalhães J, Larbi A (2019) RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Reports*, 26:1627–1640.e7

[211] Dhahbi J, Nunez Lopez YO, Schneider A, Victoria B, Saccon T, Bharat K, McClatchey T, Atamna H, Scierski W, Golusinski P, Golusinski W, Masternak MM (2019) Profiling of tRNA Halves and YRNA Fragments in Serum and Tissue From Oral Squamous Cell Carcinoma Patients Identify Key Role of 5′ tRNA-Val-CAC-2-1 Half. *Frontiers in Oncology*, 9:959

[212] Kowalski MP, Krude T (2015) Functional roles of non-coding Y RNAs. *The International Journal of Biochemistry & Cell Biology*, 66:20–29

[213] Nguyen TTM, van der Bent ML, Wermer MJH, van den Wijngaard IR, van Zwet EW, de Groot B, Quax PHA, Kruyt ND, Nossent AY (2020) Circulating tRNA Fragments as a Novel Biomarker Class to Distinguish Acute Stroke Subtypes. *International Journal of Molecular Sciences*, 22:135

[214] Gulìa C, Signore F, Gaffi M, Gigli S, Votino R, Nucciotti R, Bertacca L, Zaami S, Baffa A, Santini E, Porrello A, Piergentili R (2020) Y RNA: An Overview of Their Role as Potential Biomarkers and Molecular Targets in Human Cancers. *Cancers*, 12:1238

[215] Li Y, Luo J, Zhou H, Liao JY, Ma LM, Chen YQ, Qu LH (2008) Stress-induced tRNA-derived RNAs: a novel class of small RNAs in the primitive eukaryote Giardia lamblia. *Nucleic Acids Research*, 36:6048–6055

[216] Wang J, Ma G, Ge H, Han X, Mao X, Wang X, Veeramootoo JS, Xia T, Liu X, Wang S (2021) Circulating tRNA-derived small RNAs (tsRNAs) signature for the diagnosis and prognosis of breast cancer. *npj Breast Cancer*, 7:1–5

[217] Jain G, Stuendl A, Rao P, Berulava T, Pena Centeno T, Kaurani L, Burkhardt S, Delalle I, Kornhuber J, Hüll M, Maier W, Peters O, Esselmann H, Schulte C, Deuschle C, Synofzik M, Wiltfang J, Mollenhauer B, Maetzler W, Schneider A, Fischer A (2019) A combined miRNA-piRNA signature to detect Alzheimer's disease. *Translational Psychiatry*, 9:250

[218] Simms CL, Zaher HS (2016) Quality control of chemically damaged RNA. *Cellular and Molecular Life Sciences*, 73:3639–3653

[219] Seok H, Lee H, Lee S, Ahn SH, Lee HS, Kim GWD, Peak J, Park J, Cho YK, Jeong Y, Gu D, Jeong Y, Eom S, Jang ES, Chi SW (2020) Position-specific oxidation of miR-1 encodes cardiac hypertrophy. *Nature*, 584:279–285

[220] DeWeerdt S (2019) RNA therapies explained. *Nature*, 574:S2–S3

[221] Valencia-Sanchez MA, Liu J, Hannon GJ, Parker R (2006) Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes & Development*, 20:515–524

[222] Hu B, Zhong L, Weng Y, Peng L, Huang Y, Zhao Y, Liang XJ (2020) Therapeutic siRNA: state of the art. *Signal Transduction and Targeted Therapy*, 5:101

[223] Zhang MM, Bahal R, Rasmussen TP, Manautou JE, Zhong Xb (2021) The growth of siRNA-based therapeutics: Updated clinical studies. *Biochemical Pharmacology*. RNA Therapeutics, 189:114432

[224] Lam JKW, Chow MYT, Zhang Y, Leung SWS (2015) siRNA Versus miRNA as Therapeutics for Gene Silencing. *Molecular Therapy - Nucleic Acids*, 4

[225] Friedmann T (1992) Approaches to Gene Therapy of Complex Multigenic Diseases: Cancer as a Model and Implications for Cardiovascular Disease and Diabetes. *Annals of Medicine*, 24:411–417

[226] Zhao Y, Alexandrov PN, Lukiw WJ (2016) Anti-microRNAs as Novel Therapeutic Agents in the Clinical Management of Alzheimer's Disease. *Frontiers in Neuroscience*, 10:59

[227] Xiao S, Chen YC, Betenbaugh MJ, Martin SE, Shiloach J (2015) MiRNA mimic screen for improved expression of functional neurotensin receptor from HEK293 cells. *Biotechnology and Bioengineering*, 112:1632–1643

[228] Burgess DJ (2012) Remember your driver. *Nature Reviews Genetics*, 13:73–73

[229] Hanna J, Hossain GS, Kocerha J (2019) The Potential for microRNA Therapeutics and Clinical Research. *Frontiers in Genetics*, 10:478

[230] Gebert LFR, Rebhan MAE, Crivelli SEM, Denzler R, Stoffel M, Hall J (2014) Miravirsen (SPC3649) can inhibit the biogenesis of miR-122. *Nucleic Acids Research*, 42:609–621

[231] Chakraborty C, Sharma AR, Sharma G, Lee SS (2021) Therapeutic advances of miRNAs: A preclinical and clinical update. *Journal of Advanced Research*, 28:127–138

[232] Mompeón A, Ortega-Paz L, Vidal-Gómez X, Costa TJ, Pérez-Cremades D, Garcia-Blas S, Brugaletta S, Sanchis J, Sabate M, Novella S, Dantas AP, Hermenegildo C (2020) Disparate miRNA expression in serum and plasma of patients with acute myocardial infarction: a systematic and paired comparative analysis. *Scientific Reports*, 10:5373

[233] Meder B, Backes C, Haas J, Leidinger P, Stähler C, Großmann T, Vogel B, Frese K, Giannitsis E, Katus HA, Meese E, Keller A (2014) Influence of the confounding factors age and sex on microRNA profiles from peripheral blood. *Clinical Chemistry*, 60:1200–1208

[234] Haun JB, Castro CM, Wang R, Peterson VM, Marinelli BS, Lee H, Weissleder R (2011) Micro-NMR for Rapid Molecular Analysis of Human Tumor Samples. *Science Translational Medicine*, 3:71ra16–71ra16

[235] Kwong GA, von Maltzahn G, Murugappan G, Abudayyeh O, Mo S, Papayannopoulos IA, Sverdlov DY, Liu SB, Warren AD, Popov Y, Schuppan D, Bhatia SN (2013) Mass-encoded synthetic biomarkers for multiplexed urinary monitoring of disease. *Nature Biotechnology*, 31:63–70

[236] Merchant RC, Clark MA, Liu T, Romanoff J, Rosenberger JG, Bauermeister J, Mayer KH (2018) Comparison of Home-Based Oral Fluid Rapid HIV Self-Testing Versus Mail-in Blood Sample Collection or Medical/Community HIV Testing By Young Adult Black, Hispanic, and White MSM: Results from a Randomized Trial. *AIDS and Behavior*, 22:337–346

[237] Lenaghan E, Holland R, Brooks A (2007) Home-based medication review in a high risk elderly population in primary care—the POLYMED randomised controlled trial. *Age and Ageing*, 36:292–297

[238] Wikström I, Lindell M, Sanner K, Wilander E (2011) Self-sampling and HPV testing or ordinary Pap-smear in women not regularly attending screening: a randomised study. *British Journal of Cancer*, 105:337–339

[239] Bowen RA, Remaley AT (2014) Interferences from blood collection tube components on clinical chemistry assays. *Biochemia Medica*, 24:31–44

[240] Backes C, Sedaghat-Hamedani F, Frese K, Hart M, Ludwig N, Meder B, Meese E, Keller A (2016) Bias in High-Throughput Analysis of miRNAs and Implications for Biomarker Studies. *Analytical Chemistry*, 88:2088–2095

Acknowledgement

First and foremost, I would like to express my sincere gratitude to my research supervisor, Andreas Keller, for giving me the opportunity to do my PhD in his group. Without his guidance, support and dedicated involvement in every step throughout the process, this work and thesis would have never been accomplished.

Furthermore, I am very grateful to Eckart Meese, whose scientific support was important to successfully complete this project.

I also wish to express my acknowledgement to the dissertation committee for their insightful comments and for spending their time to evaluate my thesis.

Furthermore, I am also much obliged to my colleagues and friends Tobias, Christina, Valentina, and Fabian who helped me with their encouragement and advises to complete this thesis, to become a better scientist and unforgettable memories during and outside of work. Hereby, I would like to single out Tobias, and thank him for the time we spent together at the university and on the long journey doing research on microRNAs.

Additionally, I appreciate the help of our secretary, Sabine, to accomplish all administrative related tasks.

Special thanks goes to my employer Hummingbird Diagnostics and all my colleagues there. I would like to mention following persons by name: Jurgita, Thomas, Ioannis and Rasti. In addition, I would like to thank in particular Bruno who helped me to finalize my thesis with his always available support. As last and most important person for me at Hummingbird, I am deeply thankful to my boss Jochen who allowed me to do this academic step and made this possible with his great backing and encouragement.

A huge thanks goes to my friend Silas who took his time to proofread my thesis.

Further, I would like to acknowledge all collaboration partners and coauthors I had the opportunity to work with and for the publications that we have accomplished together.

I would like to thank also my friends Sebastian, Mawlan, Velik, Beto, Ilknur and Beate, who helped me to recharge my batteries in the time besides my dissertation and work.

Finally, I would like to give my greatest thanks to my family for their unconditional support, especially my parents Gülyaz and Ismail who have made my career possible in all aspects and supported me with all their strength. Big thanks also to my siblings - Hatice, Yavuz and Hasan - who accompanied me from early age and are always available to help and advice.

Curriculum Vitae

Aus datenschutzrechtlichen Gründen wird der Lebenslauf in der elektronischen Fassung der Dissertation nicht veröffentlicht.

Tag der Promotion: 30. Mai 2022

Dekan: Prof. Dr. Michael D. Menger

Berichterstatter: Prof. Dr. Andreas Keller
Prof. Dr. Eckart Meese