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

Algorithms and Applications for non-coding RNAs in Aging

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

der UNIVERSITÄT DES SAARLANDES
2021

vorgelegt von Fabian Michael Kern
geb. am 13.07.1993 in Speyer, Deutschland

Tag der Promotion:	16. Dezember 2021
Dekan:	Univ.-Prof. Dr. med. Michael D. Menger
Berichterstatter:	Prof. Andreas Keller
	Prof. Eckart Meese
	Prof. Oliver Kohlbacher

"Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry." — Richard P. Feynman

Abstract

Gene expression is a complex molecular process governing fate and function of most eukaryotic cells. The fundamental mechanism, namely that genetic material of a cell is compactly stored on chromosomal DNA and at times being transcribed into messenger-RNA to facilitate on-demand protein biosynthesis, is widely known. However, the interplay of biochemical regulatory pathways underlying an individual's disease phenotype development remains incompletely understood. Intriguingly, the $\sim$ 20.000 protein-coding genes only account for 2% of the human genome, triggering profound questions on the purpose of remaining segments. In recent years it became apparent that non-coding RNAs essentially tune the observed gene expression circuits. In particular the small non-coding RNAs such as microRNAs, turned out to be regulatory players by switching on and off protein translation of target messenger-RNAs. Several thousand mammalian microRNAs have been discovered so far but little is known about their impact on the transcriptome, which likely depends on contextual variables like cell type identity, cellular and tissue environment or phase of activation.

Previous efforts demonstrated that gene expression programs in human and mouse undergo gradual changes along the life trajectory with amplification at higher ages. In parallel, age-related diseases are currently accumulating in our globally aging population, posing a serious challenge to our society and healthcare systems. Neurodegenerative disorders such as Alzheimer's disease and Parkinson's disease show steadily rising incidence rates with several million people already affected. Both are caused by pathological protein accumulation in selectively vulnerable neurons and brain regions. Notably, these neurological disorders do not appear all of a sudden in an individual but are believed to originate after long asymptomatic phases of subtle aberrant changes on the cellular level, turning early diagnosis into an intricate affair. Yet, no single comprehensive model to explain aging associated changes in gene expression exists and certainly any such model must take into account the role of microRNAs and other important non-coding RNAs.

With the advent of ultra-high-throughput sequencing techniques and unprecedented computational power, the screening of microRNAs and their targets from human biofluids and tissues became not only affordable but scalable. To deal with the increasing complexity of molecular studies, novel bioinformatics-driven approaches are needed to generate reproducible and comprehensive conclusions from large-scale data sets. Here, the role of small non-coding RNAs in governing gene expression changes observed in complex age-related diseases is explored with the aid of new methods and databases as well as several thousand RNA profiling samples.

This cumulative doctoral thesis comprises eight peer-reviewed publications. Basic research covers a comprehensive review on most target prediction tools and a novel experimental and computational workflow for microRNA-target pathway identification. In addition, with miRPathDB 2.0 the so-far largest database on enriched microRNA pathways for human and mouse is presented. Moreover, the new versatile web tool miEAA 2.0 allows rapid annotation of statistically enriched molecular properties and functions for large lists of microRNAs from ten species. The lessons learned from web-based tool development were condensed in an invited summary and survey article on scientific web server availability along with best practices for developers.

The here presented toolkit was used in three applied research studies to investigate the association between microRNAs and their target pathways in the context of aging as well as the to date largest Parkinson's disease biomarker discovery framework. Circulating microRNAs obtained low-invasively from whole-blood samples bear diagnostic and prognostic value in Alzheimer's and Parkinson's disease patients, which was discovered using machine learning models. Furthermore, selected microRNA families were found to systematically target entire signaling pathways as to effectively silence gene expression. Indeed, these pathways are affected in prevalent neurodegenerative disorders.

Taken together, the published candidate signatures and validated targets are pivotal for subsequent experimental perturbation in microRNA or gene knockout studies. In future efforts, large-scale single-cell studies will be required to further dissect disease and cell-type specificity of aging disease biomarker candidates and their long-term effect on gene expression, possibly indicating early neuropathological hallmarks.

Zusammenfassung

Genexpression ist ein komplexer molekularer Prozess, der das Überleben und die Funktion der meisten eukaryotischen Zellen entscheidend beeinflusst. Der zugrunde liegende Mechanismus, nämlich, dass das genetische Material einer Zelle kompakt in chromosomaler DNA vorliegt und je nach Bedarf in messenger-RNA zur Proteinbiosynthese genutzt wird, ist weitgehend bekannt. Allerdings ist das Zusammenspiel der regulatorischen Pfade im Hintergrund der phenotypischen Veränderungen von erkrankten Individuen nur wenig verstanden. Interessanterweise machen die fast 20.000 protein-kodierenden Gene nur in etwa 2% des menschlichen Erbgutes aus. In den letzten Jahren hat man festgestellt, dass nicht-kodierende RNAs eine essentielle Rolle bei der Einstellung der beobachteten Genexpressionsschaltkreise spielen. Insbesondere kleine nicht-kodierende RNAs wie microRNAs, stellten sich als zuvor unterschätzte regulatorische Einheiten heraus, die die Translation von Ziel-messenger-RNA in Proteine an und ausschalten. Mehrere tausend microRNAs wurden bisher bei Säugetieren entdeckt, trotzdem ist immer noch wenig über ihren Einfluss auf das Transkriptom bekannt, ein Zusammenhang der wahrscheinlich vom Kontext wie Zelltypidentität, dem zelluären Umfeld sowie dem umgebenden Gewebe, und den Aktivierungsphasen abhängt.

Frühere Forschungsarbeiten haben bereits gezeigt, dass das Genexpressionsprogramm im Menschen und in der Maus sukzessiven Änderungen im Laufe des Lebens unterworfen ist, welche sich im höheren Alter verstärken. Zur gleichen Zeit akkumulieren Fälle von altersbedingten Krankheiten in unserer immer älter werdenden, globalen Population, was ernstzunehmende Herausforderungen für unsere Gesellschaft sowie unser Gesundheitssystem mit sich bringt. Neurodegenerative Krankheiten wie Morbus Alzheimer und Morbus Parkinson zeigen eine kontinuierlich ansteigende Inzidenz, wobei bereits mehrere millionen Menschen weltweit betroffen sind. Besonders für diese Krankheiten ist, dass sie bei einem Menschen nicht spontan oder plötzlich entstehen, sondern vermutlich nach langer Zeit der asymptomatischen Phase aufgrund schleichender, abnormaler Veränderungen auf zellulärer Ebene entstehen, was eine frühe Diagnose überaus schwierig gestaltet. Bisher existiert noch kein verständliches Modell das die altersassoziierten Veränderungen der Genexpression erklären kann, wobei jedes darauf ausgerichtete Modell mit Bestimmtheit die Rolle der microRNAs und anderen wichtigen nicht-kodierenden RNAs zwangsläufig in Betracht ziehen muss.

Mit dem Aufkommen der Sequenzierung im Ultrahochdurchsatzverfahren und der unübertroffenen Leistung moderner Computersysteme, wurde die Untersuchung von microRNAs und ihren Zielgenen anhand von Proben menschlicher Flüssigkeiten und Geweben nicht nur möglich gemacht, sondern kann entsprechend hochskaliert werden. Um mit der zunehmenden Komplexität molekularer Studien Schritt zu halten, braucht es neue Ansätze der Bioinformatik um reproduzierbare und nachvollziehbare Schlüsse aus großen Datensätzen gewinnen zu können. Im Rahmen dieser Arbeit wurden kleine nicht-kodierende RNAs hinsichtlich ihrer Rolle der Genregulation in komplexen altersbedingten Krankheiten anhand neuer Methoden und Datenbanken sowie mehreren tausend Proben der RNA-Sequenzierung untersucht.

Diese kumulative Dissertationsarbeit umfasst acht von unabhängigen Experten begutachtete (peer-reviewed), wissenschaftliche Publikationen. Die Grundlagenforschung enthält einen umfassenden Übersichtsartikel zu fast allen Methoden der Vorhersage von microRNA Zielgenen sowie ein neuartiges Protokoll bestehend aus Labormethoden und computergestützen Berechnungen zur Identifikation von durch microRNAs regulierte Genpfade. Zusätzlich wird mit miRPathDB 2.0 die bisher größte Datenbank zu signifikant angereicherten microRNA Zielpfaden präsentiert. Des Weiteren, bietet die neue und vielseitige, web-basierte Software miEAA 2.0 die Möglichkeit der rasanten Annotation statistisch angereicherter, molekularer Eigenschaften sowie bekannter Funktionen einer gegebenen Liste an microRNAs von zehn Spezies. Die durch web-basierte Softwareentwicklung zuvor angelernten Fähigkeiten sowie daraus resultierende Empfehlungen für nachfolgende Entwickler wurden kurz und bündig in einem eingeladenen Übersichtsartikel zum Thema *Verfügbarkeit wissenschaftlicher Software im Internet* veröffentlicht.

Die hier präsentierten Werkzeuge wurden gezielt in drei Studien zur angewandten Forschung genutzt um die Assoziation zwischen microRNAs und ihren Zielpfaden im Kontext der allgemeinen Altersforschung sowie im Rahmen der bisher größten Studie zur Entdeckung von Biomarkern der Parkinson Krankheit zu untersuchen. Im Blutkreislauf zirkulierende microRNAs, die anhand von Vollblutproben extrahiert wurden, zeigen diagnostisches und prognostisches Potential bei Alzheimer und Parkinson Patienten, was mit Methoden des maschinellen Lernens entdeckt werden konnte. Überdies konnte herausgefunden werden, dass bestimmte microRNA Familien systematisch Signalwege blockieren können, um die Genexpression herunterzufahren. Tatsächlich sind diese Pfade auch in neurodegenerativen Krankheiten betroffen.

Insgesamt sind die hier publizierten Signaturen von Kandidaten-microRNAs und einiger validierter Zielgene herausragend dazu geeignet in weiteren Studien anhand von gezielter Ausschaltung im Labor genauer untersucht zu werden. In zukünftigen Forschungsprojekten sollten groß angelegte Untersuchungen vieler einzelner Zellen im Vordergrund stehen, um zu verstehen wie spezifisch für

Krankheit oder Zelltyp die hier genannten Biomarker-Kandidaten für altersbedingte Krankheiten sind. Auch wird es wichtig sein die Langzeiteffekte von dysregulierten microRNAs auf die Genexpression zu verstehen, die möglicherweise frühzeitig neuropathologische Kennzeichen widerspiegeln.

Scientific publications

This is a cumulative thesis based on the following peer-reviewed journal papers [1–8]. The publications included herein are identical to the published versions if not indicated otherwise. Equally contributing first authors are denoted by a superscript dagger (†) symbol.

1. **F. Kern**, C. Backes, P. Hirsch, T. Fehlmann, M. Hart, E. Meese, A. Keller. What's the target: understanding two decades of *in silico* microRNA-target prediction. *Brief Bioinform*, 21(6):1999–2010, 2020. ISSN 1467-5463. doi: 10.1093/bib/bbz111. URL `https://doi.org/10.1093/bib/bbz111`.

2. T. Kehl†, **F. Kern**†, C. Backes, T. Fehlmann, D. Stöckel, E. Meese, H. P. Lenhof, A. Keller. miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database. *Nucleic Acids Res*, 48(D1): D142–D147, 2020. ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gkz1022. URL `https://doi.org/10.1093/nar/gkz1022`.

3. N. Ludwig†, T. Fehlmann†, **F. Kern**†, M. Gogol, W. Maetzler, S. Deutscher, S. Gurlit, C. Schulte, A. K. von Thaler, C. Deuschle, F. Metzger, D. Berg, U. Suenkel, V. Keller, C. Backes, H. P. Lenhof, E. Meese, and A. Keller. Machine Learning to Detect Alzheimer's Disease from Circulating Non-coding RNAs. *Genomics Proteomics Bioinformatics*, 17(4):430–440, 2019. ISSN 1672-0229 (Print) 1672-0229. doi: 10.1016/j.gpb.2019.09.004. URL `https://doi.org/10.1016/j.gpb.2019.09.004`.

4. **F. Kern**†, T. Fehlmann†, J. Solomon, L. Schwed, N. Grammes, C. Backes, K. Van Keuren-Jensen, D. W. Craig, E. Meese, and A. Keller. miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. *Nucleic Acids Res*, 48(W1):W521–W528, 2020. ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gkaa309. URL `https://doi.org/10.1093/nar/gkaa309`.

5. T. Fehlmann, B. Lehallier, N. Schaum, O. Hahn, M. Kahraman, Y. Li, N. Grammes, L. Geffers, C. Backes, R. Balling, **F. Kern**, R. Krüger, F. Lammert, N. Ludwig, B. Meder, B. Fromm, W. Maetzler, D. Berg, K. Brockmann, C. Deuschle, A. K. von Thaler, G. W. Eschweiler, S. Milman, N. Barziliai, M. Reichert, T. Wyss-Coray, E. Meese, and A. Keller. Common diseases alter the physiological age-related blood microRNA profile. *Nat Commun*, 11(1):5958, 2020.

ISSN 2041-1723. doi: 10.1038/s41467-020-19665-1. URL `https://doi.org/10.1038/s41467-020-19665-1`.

6. **F. Kern**[†], L. Krammes[†], K. Danz, C. Diener, T. Kehl, O. Küchler, T. Fehlmann, M. Kahraman, S. Rheinheimer, E. Aparicio-Puerta, S. Wagner, N. Ludwig, C. Backes, H. P. Lenhof, H. von Briesen, M. Hart, A. Keller, and E. Meese. Validation of human microRNA target pathways enables evaluation of target prediction tools. *Nucleic Acids Res*, 49(1):127–144, 2021. ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gkaa1161. URL `https://doi.org/10.1093/nar/gkaa1161`.

7. **F. Kern**[†], T. Fehlmann[†], and A. Keller. On the lifetime of bioinformatics web services. *Nucleic Acids Res*, 48(22):12523–12533, 2020. ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gkaa1125. URL `https://doi.org/10.1093/nar/gkaa1125`.

8. **F. Kern**, T. Fehlmann, I. Violich, E. Alsop, E. Hutchins, M. Kahraman, N. L. Grammes, P. Guimarães, C. Backes, K. L. Poston, B. Casey, R. Balling, L. Geffers, R. Krüger, D. Galasko, B. Mollenhauer, E. Meese, T. Wyss-Coray, D. W. Craig, K. Van Keuren-Jensen, and A. Keller. Deep sequencing of sncRNAs reveals hallmarks and regulatory modules of the transcriptome during Parkinson's disease progression. *Nature Aging*, 1(3):309–322, 2021. ISSN 2662-8465. doi: 10.1038/s43587-021-00042-6. URL `https://doi.org/10.1038/s43587-021-00042-6`.

Contents

List of Figures

Abbreviations

A

AD Alzheimer's disease 32–34, 36, 44, 47, 48, 156

AGO Argonaute 25–27, 29

ANOVA Analysis of Variance 40

ApoE Apolipoprotein E 33

B

BAM Binary Sequence Alignment/Map format 38

BED Browser Extensible Data 38

bps base pairs 25, 39

C

CNS central nervous system 32

COVID-19 coronavirus disease 2019 50

CPM counts per million 40

CSF cerebrospinal fluid 35

CWL Common Workflow Language 39

D

DAG directed acyclic graph 39

DNA deoxyribonucleic acid 21–23, 31

G

GBM gradient boosting machines 44

GFF General Feature Format 38

GO Gene Ontology 41

GSEA gene set enrichment analysis 41, 42

I

ILP integer linear program 43

K

KEGG Kyoto Encyclopedia of Genes and Genomes 41

L

lncRNA long non-coding RNA 25, 26, 31, 32, 155

M

m5C 5-methylcytosine 23

MCI mild cognitive impairment 33

miEAA miRNA Enrichment Analysis and Annotation tool 41, 42

miRNA microRNA

miRPathDB miRNA pathway dictionary database 41–43

ML machine learning 44

MMSE Mini–Mental State Examination 34

mRNA messenger-RNA 21–24, 26, 27, 29–32, 36, 43, 45, 48, 155, 156

N

NAR Nucleic Acids Research 36, 37

NCBI National Center for Biotechnology Information 36

NCER-PD National Centre for Excellence in Research on Parkinson's Disease 49

ncRNA non-coding RNA 23, 25, 31, 32, 35–37, 42, 44, 45, 155–157

NGS next-generation sequencing 39

nts nucleotides 23, 26, 27

O

ORA over-representation analysis 41

P

PBMCs peripheral blood mononuclear cells 30, 35

PCA principal component analysis 40

PD Parkinson's disease 32–34, 36, 49, 156, 157

PET positron emission tomography 35

piRNA piwi-associated RNA 25, 26

PPMI Parkinson's Progression Markers Initiative 34, 49

R

RISC RNA-induced silencing complex 27, 29

RNA ribonucleic acid

RPM reads per million 40

RPMMM reads per million mapped miRNA 40

rRNA ribosomal RNA 23, 25, 26

RT-qPCR / qRT-PCR Reverse transcription polymerase chain reaction combined with quantification real-time polymerase chain reaction 28

S

scaRNA small Cajal body-specific RNA 25, 26

siRNA short interfering RNA 25, 26

SNCA α-synuclein 33, 34

sncRNA small non-coding RNA 25, 39, 49, 156

snoRNA small nucleolar RNA 25, 26

snRNA small nuclear RNA 25, 26

SNV single-nucleotide variant 40

T

TFs transcription factors 23, 24, 47

tRNA transfer RNA 22, 23, 25, 26

U

UPDRS Unified Parkinson's Disease Rating Scale 34

UTR untranslated region 27

V

VST variance-stabilization transformation 40

W

WoS Web of Science 38

1
Altered gene activity as a latent proxy to molecular aging phenotypes

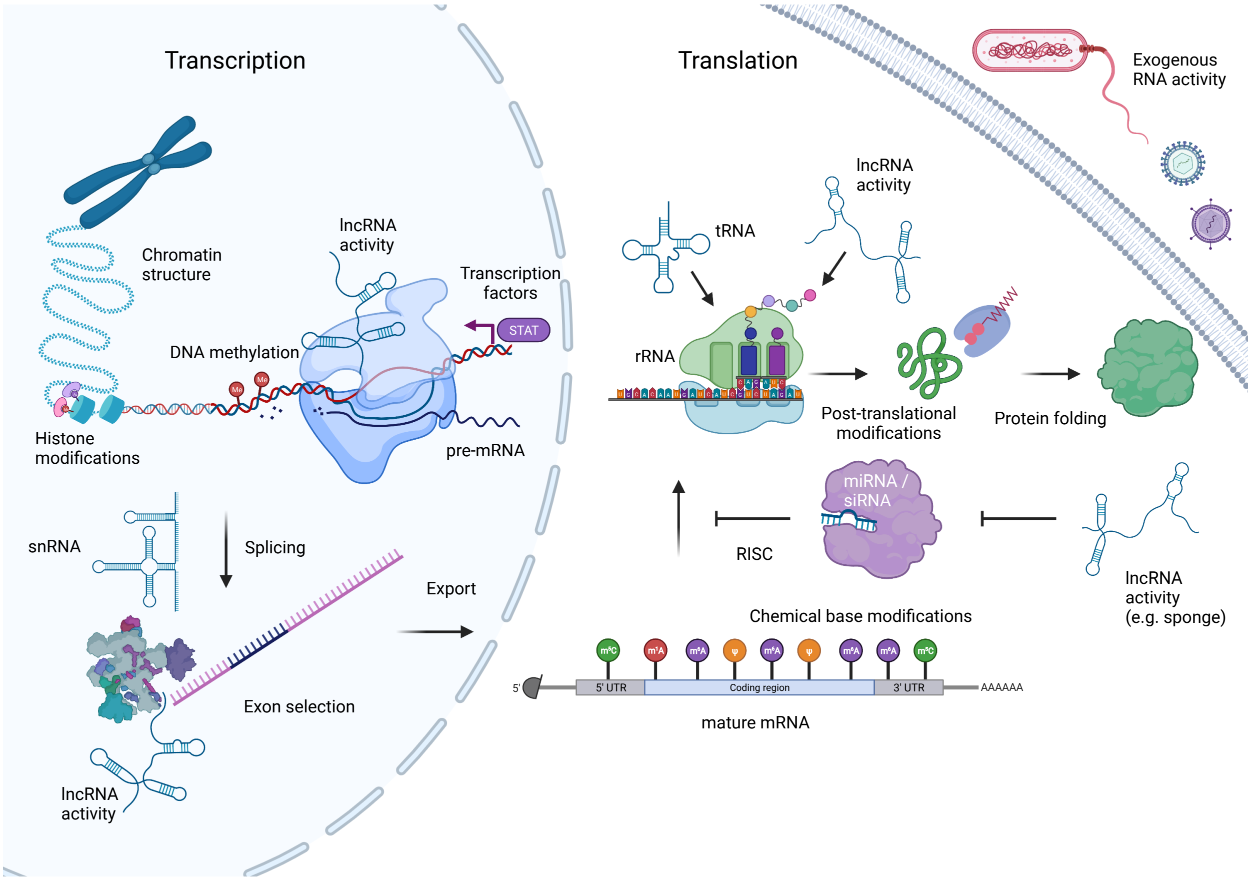
1.1 Gene expression

All organisms in the tree and domains of life share core architectural biological principles that emerged during 4, 6 billion years of evolution. Principles include compact organization into cells and directed flow of information through inheritance and transfer of genetic materials [9; 10]. Thereby, information essential for cell survival is naturally encoded within specific genetic segments termed genes, so-called blueprints of proteins, on deoxyribonucleic acid (DNA) molecules. So far, 2, 3 million eukaryote, prokaryote, and archaeal species were systematically characterized in the tree of life, of which approximately 15.000 genomes are partially or fully reconstructed [11]. While the DNA can fundamentally be separated into coding and non-coding regions, they are not functionally independent and constantly evolving through a tight coupling in evolutionary dynamics [12]. Gene expression depicts the series of events necessary to translate the parts of the genetic information into functional molecules governing every biological process in a cell.

As comparative genomics step-wisely unveiled the purpose of genes, it became clear that gene expression is affected by changes in either *cis* or *trans* direction. Referring to regulatory changes within a gene or through other genes and non-coding regions, these mechanisms govern the development of eukaryotic organisms and therefore shape complex phenotypes [13–15]. In case of *Homo sapiens*, for example, 19.955 protein-coding genes have been discovered so far [16]. Since not all genes are actively transcribed into their RNA counterpart, the messenger-RNA (mRNA), in every cell at the same time, researchers conducted tremendous efforts to unravel the mechanisms of gene regulation. Frequently, much simpler organized model organisms such as yeast and worms were investigated under evolutionary aspects[17; 18]. On the search for a rigorous theoretic understanding on how differential gene expression dynamics shapes the phenotype of an individual or entire populations, the field of epigenetics counts toward the most important breakthroughs in this matter [19; 20]. Broadly speaking, it unites our theoretic and biochemical understanding of how gene expression is dynamically altered, the processes underlying complex developmental and disease phenotypes [21].

In the context of this thesis, gene expression and regulation will be viewed and discussed as known for mammalian species, but primarily human, mouse and other model organisms. Starting from the canonical definition of gene expression, every protein-coding gene is transcribed from DNA into pre-mRNA by an RNA polymerase II. Since gene bodies contain both coding and non-coding segments, so-called exons and introns, in a splicing process introns are removed and selected exons combined. Then, a 5′ cap and a 3′ repetitive adenine sequence (poly-A tail) are ligated to the pre-mRNA, primarily for stabilizing and protecting the RNA from degradation by RNAses. The pre-mRNA is exported from the nucleus into the cytoplasm. Free mature mRNA is bound by the ribosomal complex and initiating from the start codon, a specific base triplet, the amino acid-chain synthesis and elongation is conducted until reaching a stop codon. During synthesis, each codon stands for one of the 20 amino acids, which are supplied to the ribosomal complex by bound transfer RNA (tRNA). The resulting peptide chain folds into secondary and tertiary structures, eventually producing a functional protein. The complete process of eukaryotic protein biosynthesis is depicted in Figure 1.1.

Figure 1.1: The process of protein biosynthesis is illustrated. Each step is annotated with different regulatory players that eventually modulate gene expression strength in eukaryotic cells. Created with BioRender.com.



1.1.1 Eukaryotic gene regulation

The cellular gene expression program is influenced and reprogrammed by a wide array of potentially inducing, inhibiting, or function modifying mechanisms [22] (Figure 1.1). On the transcriptional level, onset and strength of transcription are primarily relayed through transcription factors (TFs) and non-coding segments such as distally or intergenically located enhancers, which in turn can be bound by TFs [23–26]. A prominent role of TFs is to guide the gene expression program during cellular proliferation and differentiation, or to even induce and maintain pluripotency [27; 28]. Additionally, chemical modifications and organization of DNA through base methylation or histone modifications constitute both inductive and repressive milieus [29; 30]. Collectively, these features have been proven useful for effective prediction of available mRNA levels for multiple cell lines independently [31]. However, neither the rate of transcription nor the mRNA concentration are perfect predictors of protein abundance for multiple reasons, including additional regulatory steps of intervention, technical and biological noise and environmental context. For example, through alternative splicing, i.e. a differential selection of exons inserted in linear order into the final mRNA, protein isoforms with varying efficacy or function may arise [32; 33].

On the post-transcriptional level also several independent mechanisms come into play. First, specific chemical modifications of RNA nucleotides (nts) modulate mRNA concentration by altering secondary or tertiary structures and therefore influencing stability or speed of degradation [34; 35]. For example, dysregulated formation of 5-methylcytosine (m5C) on tRNAs has been linked to the development of human cancer [36]. In addition, several classes of non-coding RNA (ncRNA) such as microRNA (miRNA), tRNA or ribosomal RNA (rRNA) are directly involved in post-transcriptional regulation of mRNAs, eventually affecting protein abundance. The particular mechanisms of ncRNAs are discussed in more detail in section 1.2 and following. Finally, on the post-translational level, protein half-live is governed by enzymatic modifications, e.g. amino acid phosphorylation and targeted proteolysis primarily via the ubiquitin–proteasome system [22; 37–39].

1.1.2 Deviations from the canonical pathway and the problem of confounders

Given the basic framework of gene regulation mechanisms in eukaryotic cells, it is important to note that many exceptions from the canonical pathways exist, and depending on the environment are rather the norm than an exception [40; 41]. Especially upon stress response or in a disease context, non-canonical pathways seem to reroute gene expression, offering a potential escape of upregulated mRNA and protein synthesis from otherwise restrictive regulatory controls [42–44]. These important observations could explain why there are often crucial differences when comparing cellular systems *in* *vitro* and *in vivo* due to negligence of physiological conditions [45]. In addition, methods for profiling gene expression are often intrinsically confounded, making it difficult to distinguish between biological and technical noise or actual biological variation [46–48]. For example, personal (epi-)genetic background, demographic features, lifestyle or age may explain heterogeneity of gene expression [49]. Still, it is possible to considerably improve expression measurements by careful *in silico* correction using modeling techniques and applied statistics. As shown recently by Parsana *et al.* more accurate gene regulation networks can be obtained through the analysis and correction of latent confounders [50]. Discerning biological covariates from technical perturbations in differential abundance analysis is a long-withstanding issue in bioinformatics and biostatistics, and constantly changes shape due to the development of new experimental approaches, none of which is ultimately bias-free [51–54]. In conclusion, gene expression is an inherent stochastic process with numerous possibilities for fine-tuning, a process of which we can sample only to a limited extent, requiring advanced data analysis methods and knowledge.

1.1.3 *Biological context dependency*

In addition to the reasons for variations in gene expression described in the previous section, an overwhelming proportion of mammalian cells rapidly undergo differentiation. As a result, they restrict transcriptome diversity albeit to fulfill a highly specific function. That is the reason why cells are systematically characterized into cell types, cell lines, tissues, i.e. physically separated mesh of cell types, and organs using multi-scale models. Intercellular information exchange in between these compartments is known to be omnipresent via signaling pathways [55]. Even though any two cells of the same cell type will have slight variations in their transcriptome, cell type identity is primarily driven by gene expression programs, even enabling the estimation of cellular composition just based on mRNA levels [56–58]. From another angle, the interaction between TFs and their target genes is known to be highly tissue-specific [59]. It is assumed that gene expression is a multidimensional process, depending on the individual, genetic background, tissue of origin, cell type, cellular environment, as well as the temporal and spatial axis [60–63].

With over 30 trillion cells distributed over more than 70 organs composed of at least 200 cell types, a transcriptomic map of the human body is yet far from complete and comprises an ultra-high-dimensional search space at the extremes [60; 64–66] (Figure 1.2). To understand how genotype determines phenotype on the organism-level, comprehensive multi-scale models are in great demand. Nevertheless, previous efforts already revealed intriguing insights into some of the restrictions of this data space. The majority of genes seem to express one dominant form of a transcript, reducing variation due to isoforms, and for which transcriptome similarity across human tissues was found to be higher than within tissue similarity between

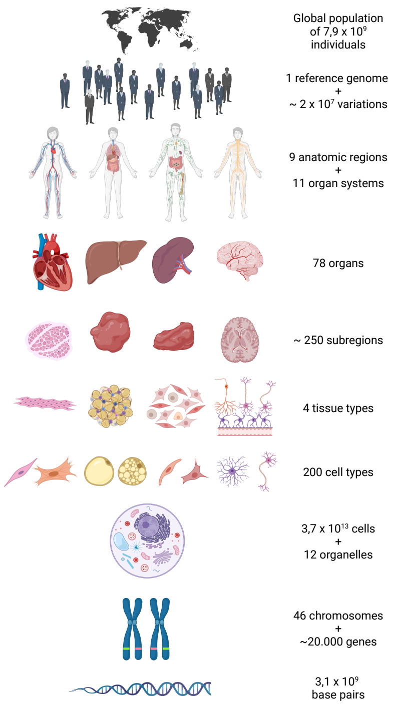

Global population of 7,9 x 10^9 individuals
1 reference genome + ~ 2 x 10^7 variations
9 anatomic regions + 11 organ systems
78 organs
~ 250 subregions
4 tissue types
200 cell types
3,7 x 10^{13} cells + 12 organelles
46 chromosomes + ~20.000 genes
3,1 x 10^9 base pairs

Figure 1.2: From modeling expression of a single gene in a single cell (bottom) to understanding the phenotype on a global level (top), data complexity increases numerous orders of magnitude. Feature dimensions for each level were stated exactly if known and estimated from the literature otherwise. Created with BioRender.com.

mouse and human, even though they are closely related [67; 68]. Previously underappreciated roles of exogenous RNA in host-pathogen interactions are now increasingly uncovered as bacterial origins may explain some of the observed differences [69–72]. Moreover, stress response pathways induced upon infection are partially conserved in bacteria and primarily regulated by ncRNAs, which may leak into the host system upon activation of the immune response [73]. In sum, eukaryotic gene expression regulation is a complex process with unlimited facets, requiring further comprehensive assessment under varying physiological and aberrant conditions to advance our understanding on how genotype determines phenotype.

1.2 Non-coding RNAs

During the two primary landmark projects of deciphering the human genome that ended in 2003, it became apparent that protein-coding segments constitute only a minor fraction of the approximately 3 billion base pairs (bps) [74; 75]. A further characterization concluded almost half of the transcriptome are nonpolyadenylated molecules with unknown function, raising questions about the role of transcribed ncRNA and non-transcribed genomic regions [76]. Debates about the actual validity of these observations and whether non-coding regions are simply biological junk and a leftover from evolution followed [77; 78]. As of today, we know that 2% of the human genome is covered by protein-coding genes, there are at least as many non-coding genes than coding, and these fulfill a broad range of regulatory functions [79–83].

Following the sheer diversity of ncRNAs discovered so far, new ontologies emerged to systematically characterize biogenesis, function and sequence relation [84; 85]. To this end, ncRNAs are typically classified by their size into long non-coding RNA (lncRNA) ($>$ 200 bps) or small non-coding RNA (sncRNA) and whether they occur inter- or intragenically, whereby some lncRNAs are precursors to sncRNAs [86; 87]. It was since discovered that lncRNAs exhibit unique features of biogenesis and function, are found in many cell types, and have diverse regulatory functions on both the transcriptional and post-transcriptional level [88–90]. SncRNAs on the other hand, partition clearly into distinct classes by sequential and structural properties, biogenesis and function. The most important ones are tRNA, rRNA, small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), small Cajal body-specific RNA (scaRNA), miRNA, short interfering RNA (siRNA), and piwi-associated RNA (piRNA) [16; 79; 91]. An overview of known coding and non-coding RNAs expressed in eukaryotic cells is illustrated in Figure 1.3. Several sncRNAs require protein interaction partners to function, e.g. miRNAs, siRNAs and piRNAs are selectively loaded by different members of the Argonaute (AGO) protein families [92]. Through advances in molecular biomedicine many ncRNAs have been shown to be implicated in human diseases [93]. The phenomenon is further examined in detail for miRNAs and

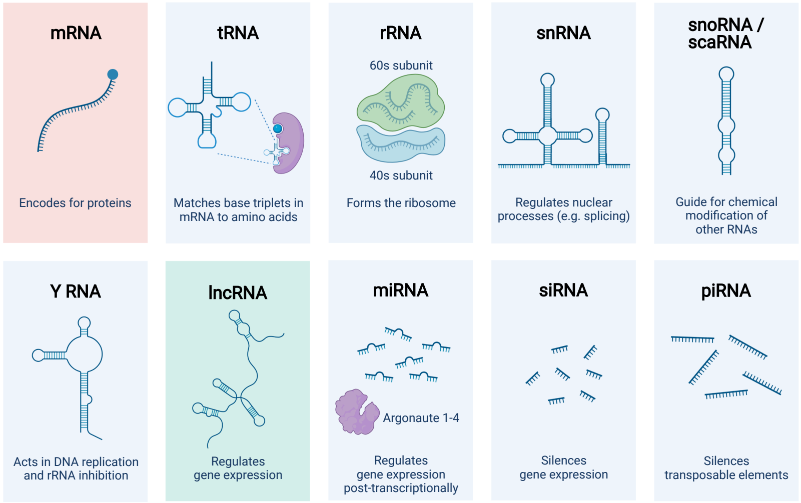

mRNA
Encodes for proteins
tRNA
Matches base triplets in mRNA to amino acids
rRNA
60s subunit
40s subunit
Forms the ribosome
snRNA
Regulates nuclear processes (e.g. splicing)
snoRNA / scaRNA
Guide for chemical modification of other RNAs
Y RNA
Acts in DNA replication and rRNA inhibition
lncRNA
Regulates gene expression
miRNA
Argonaute 1-4
Regulates gene expression post-transcriptionally
siRNA
Silences gene expression
piRNA
Silences transposable elements

Figure 1.3: Major classes of RNAs expressed in most eukaryotic cells. From left to right and top to bottom are shown mRNA, tRNA, rRNA, snRNA, snoRNA / scaRNA, Y RNA, lncRNA, miRNA, siRNA, and piRNA. Each panel contains exemplary two-dimensional RNA structure(s) characteristic for the class. Below each structure a short explanation on function is given. Coding RNAs are highlighted in red colored background, long non-coding RNAs in green and small non-coding RNAs in blue. Created with BioRender.com.

neurodegenerative diseases in this thesis.

1.2.1 *MicroRNAs*

The first metazoan miRNA was discovered in 1993 in nematodes and described as small RNA regulating the expression of the protein-coding gene LIN-14 through complementary binding [94]. In the early phase, miRNAs were investigated primarily in model organisms such as *C. elegans* and *D. melanogaster* and found to be highly conserved [95–99]. Due to a rapid increase in the number of validated miRNAs across dozens of species that nevertheless exhibit high sequence conservation, unified models of annotation and databases were developed [100–102]. However, these resources needed to undergo substantial revisions during another two decades of intense research. Reasons were technological challenges, like addressing sequencing bias, or emerging biological insights leading to a separation of siRNAs and snoRNAs that were confused with miRNAs due to their functional similarity [103–105]. Recently, the number of human miRNAs was reliably estimated to be approximately 2.300, many of which have been experimentally validated so far, but yet another couple of thousand potential candidates remain [54].

Since their discovery, biogenesis and molecular function of miRNAs have been extensively described. Briefly, in the canonical biogenesis pathway pri-miRNAs, which exhibit a characteristic stem-loop structure, arise from intronic or exonic sequences of host gene transcripts. These are processed by the Drosha, Pasha, and DGCR8 microprocessor complex yielding a shorter pre-miRNA hairpin [106–108]. Following export from the nucleus by the Exportin-5 complex, a second enzyme called Dicer cleaves the loop structure from the hairpin, resulting in a double-stranded RNA molecule of 18 to 22 nts length [108; 109]. The double-strand dissolves and preferentially one of the strands is loaded by a specific member of the AGO protein family. AGO protein loading is a selective process depending on the downstream mode of action of each miRNA [108; 110]. The common miRNA nomenclature thus specifies mature strands of each precursor by a $-3p$ or $-5p$ name suffix. AGO proteins with miRNA embedded are termed RNA-induced silencing complex (RISC) [111]. Competitive binding of complementary antisense target sites within the $3'$ or $5'$ untranslated region (UTR) typically results in either a prevailing loss of stability through depolyadenylation, mRNA cleavage, or a complete blockade of the ribosomal machinery [108; 112–114]. Therefore, miRNAs are primarily reported as negative gene regulators, although exceptions exist [115].

1.2.2 *MicroRNA target gene relation*

By closely studying potential target sequences up to the nucleotide level, further mechanistic insights on miRNA targeting were gained. Effective binding sites are at least 5 nts long segments in target mRNAs and antisense complementary matching to the seed region of the miRNA, i.e. position 2 to 8 counting from the 5′ end [116–118]. Additionally, the regulatory site can involve supplemental $3'$ pairing of the miRNA bases 13 to 16 and in rare cases involve no $5'$ seed pairing at all, most likely affecting regulation efficacy [116; 119]. The subsequent development of more than 100 computational approaches in combination with high-throughput experimental validation lead to the discovery of additional determinants of functional miRNA sites. Well-known examples are sequence conservation, sequence location, structure, and thermodynamic features [1; 120–122]. Therefore, several studies estimated a large proportion of the human protein-coding genome to be regulated by miRNAs with a pronounced evolutionary origin [123–126]. The observed abundance of binding sites in the genome raised many questions about how miRNAs specifically regulate developmental processes in a precise temporal and spatial manner [127–129].

1.2.3 *MicroRNA induced pathway modulation*

Further evidence and mechanistic insights suggested an inherent preference of single miRNAs or entire families to target multiple genes on specific pathways such as apoptosis [130–132]. These findings call for a systematic analysis and subsequent validation of entire miRNA targetomes. Functional target pathways were often studied in either a physiological setting, e.g. cardiovascular disease, cancer [133–139], or in the context of cell and organ development like proliferation and differentiation cascades, as for example in neurons [140; 141]. Intriguingly, several miRNAs seem to be involved in cellular immune system homeostasis and molecular aging, suggesting a potential role in age-related diseases [142; 143]. Ben-hamo and Efroni were the first to hypothesize in 2015 whether single miRNAs prefentially enrich their targets on cellular pathways and showed evidence for such in ten types of cancer [144]. In a second study by Kehl *et al.*, seed similarity was shown in principle to predict pathway similarity [145]. Although remarkable exceptions exist where apparent pathway similarity was not met by seed similarity, the findings suggest a complex relationship of regulatory dependencies between miRNAs. However, a considerable imbalance in the number of targets and functional pathways reported in the literature exists. The human miR-34a is exceptionally enriched in cancer research, making any objective comparison of miRNA targetomes a challenging task [146]. Therefore, several key resources have been developed to collect computational and experimental evidence for every known miRNA using statistical methods to judge significant enrichment [2; 147].

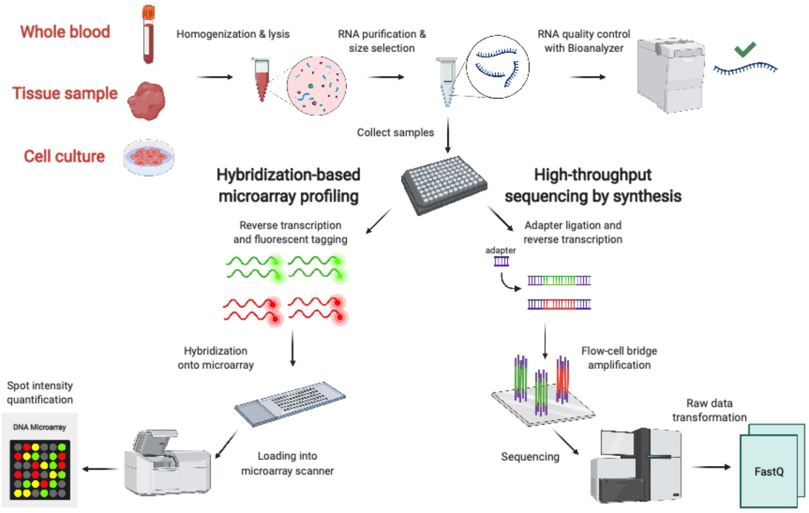

Whole blood
Tissue sample
Cell culture
Homogenization & lysis
RNA purification & size selection
RNA quality control with Bioanalyzer
Collect samples
Hybridization-based microarray profiling
High-throughput sequencing by synthesis
Reverse transcription and fluorescent tagging
Adapter ligation and reverse transcription
adapter
Hybridization onto microarray
Flow-cell bridge amplification
Spot intensity quantification
DNA Microarray
Loading into microarray scanner
Sequencing
Raw data transformation
FastQ

Figure 1.4: Comparison of microarray and high-throughput sequencing workflows for RNA profiling. Starting from tissue, fluid or cell culture samples, cells are homogenized and then lysed. Following purification and positive RNA quality analysis, resulting RNA suspension for each sample is either subject to microarray (left) or sequencing-based (right) profiling. Created with BioRender.com.

1.2.4 *Screening techniques*

A multitude of experimental methods have been used so far to quantify miRNA expression and each has unique features and application scopes [148]. Reverse transcription polymerase chain reaction combined with quantification real-time polymerase chain reaction (RT-qPCR / qRT-PCR), fluorescent labelling followed by antisense complementary hybridization on microarrays, and small RNA sequencing were established as gold-standard in the field [149–154]. A comparison of the lab workflows for the latter two is shown in Figure 1.4. Sample preparation, RNA extraction, quantification of RNA concentration and quality control belong to the standard workflow and precede any of the aforementioned methods [148]. In particular, this common procedure allows profiling of miRNAs across almost any type of biological aliquot like from solid tissue or body fluids. Sample handling is considered the most critical step, determining the amount of successfully recovered RNA [155; 156]. Still, variability due to technology persists, significantly affecting reproducibility in respective scientific literature [157–160]. While there is often a good correlation between all of the methods, crucial differences arise for miRNAs with a rather low expression, where microarrays were deemed to be more sensitive than sequencing-based profiling [157; 158]. In contrast, sequencing is necessary to discover new miRNAs and isoforms in a high-throughput manner [161; 162]. Several fundamentally different commercial protocols have been published, resulting in different capture efficiency, for example due to sequence composition or amplification bias [154; 163–166]. These aspects justify a considerate *in silico* analysis of miRNA expression and at best, results should always be confirmed with a second, independent technology and experiment.

1.2.5 Target validation techniques

MiRNA targets are often identified by computational means and subsequently validated *in vitro* using a broad toolset of experimental methods. The number of experimentally validated pairs per species contained in the most recent and all previous database releases of the miRTarBase is shown in Figure 1.5 [167]. In general, these methods can be classified into low- and high-throughput methods with varying levels of confidence. Starting with the low-throughput methods, dual-luciferase reporter assays and western blot are frequently used to validate in between a few up to a hundred mRNA targets per miRNA [168–172]. Both techniques act on the protein-level, making it difficult to judge the efficacy of individual binding sites and cooperative target sites. Nonetheless systematic target knockout through a number of experiments exponential in the site counts is also possible [118; 173]. In recent efforts one tried to improve the throughput of luciferase assays for target validation through cellular multiplexing and improved automation [6; 174–176]. Reporter-based techniques have been critiqued for being labor-intensive, overly artificial due to non-physiological overexpression of miRNA and target, and for ignoring the intrinsic nature of miRNA networks [6; 177].

High-throughput techniques instead are pivotal for miRNA-target relationship discovery in a genome-wide scope [178]. Unbiased profiling of miRNAs and mRNAs using paired microarray or sequencing from the same sample, often combined with overexpression or gene perturbation, is a popular approach. It provides rich information on potential regulatory networks at reasonable costs [139; 171]. However, in this setup a direct regulatory relationship between each miRNA and target remains difficult to prove and substantial computational analysis is required to find targets with high specificity [179]. For a more mechanistic high-throughput profiling, immunoprecipitation-based approaches such as CLIP-seq or CLASH were developed. AGO antibody-based purification helps to enrich for RISC complexes, followed by enzymatic release of bound miRNA-mRNA duplexes and sequencing [180–182]. Drawbacks are high noise levels in the obtained reads and restriction of regulatory pathways to the antibody-specific member of the AGO family. Similar to miRNA profiling, techniques

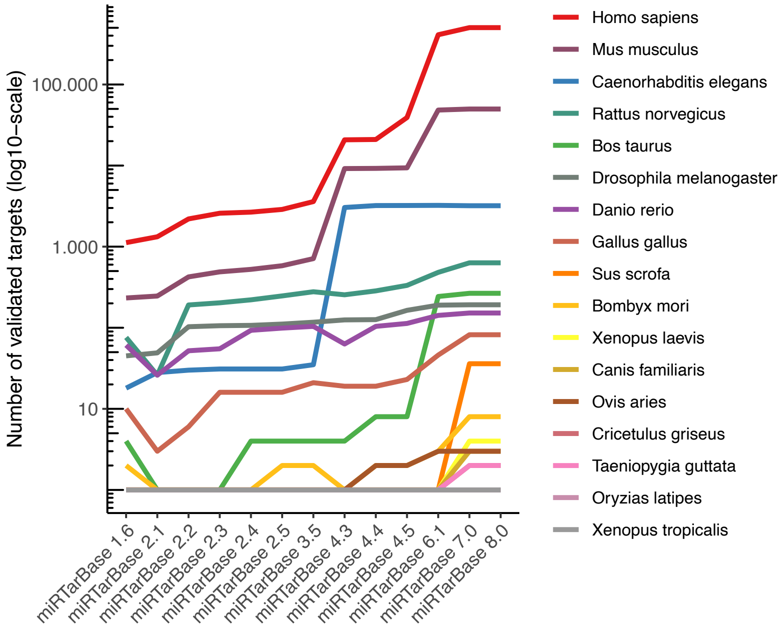


Figure 1.5: Number of validated miRNA-target pairs (y-axis) per species and miRTarBase release (x-axis) [167]. Lines are colored according to species.

should be selected ideally based on the application scope and complemented by independent validation.

1.2.6 *Bioavailability and physiology of microRNAs*

The in Section 1.2.3 mentioned phenomena motivate further studies to map miRNA expression in the entire human body and under varying physiological conditions; otherwise the conjectured relationships between miRNAs and target mRNAs remain theoretic. Enormous efforts for creating comprehensive molecular atlases of organs and biofluids have already been conducted using independent technologies such as sequencing and microarrays [183–186]. Albeit comparisons between such data sets should be performed carefully since miRNA expression measurements are extremely sensitive to tissue structure dissolution, sample quality and preparation, library protocols, and platform-dependent biases [148; 187]. Furthermore, comparing miRNAs across species with matched tissues and fluids is deemed important as sequence is often strongly conserved while expression is not necessarily [188; 189]. Whole-blood, plasma or serum specimens are often favored in biomarker studies due to the low-invasive sampling procedure and the high abundance of circulating miRNAs in the peripheral system [190; 191]. Physiological miRNA expression in blood is remarkably sensitive to environmental cues through personal lifestyle, like endurance and strength training, pointing at possible implications for biomarker research [192; 193]. It was further discovered that specific miRNAs vary along the age axis in human peripheral blood mononuclear cells (PBMCs) [194]. Furthermore, miRNAs tightly orchestrate organismal aging in *C. elegans*, raising more questions about a potentially conserved role in mediating longevity [195–197]. Another study found that *mmu-miR-204-5p* controls proliferation and differentiation cascades in preadipocytes in a mouse model of obesity [198]. A notable upsurge of discoveries on the contribution of lncRNAs to tissue aging and associated complications such as hypertension and cardiovascular incidents occurred in recent years [199; 200]. Taken together, the increasing body of evidence motivates a systematic description of physiological alterations caused by ncRNAs upon aging. In particular, distinguishing which of the conjectured effects are either correlative or causative towards aging phenotypes will be a continuous challenge.

1.3 Aging and age-related disease

Life on earth is limited by natural means with drastic differences in longevity between mammals and other vertebrates [201; 202]. Besides an individual's experience, genotype identity and epigenetic phenomena have been determined as major underlying factors driving the aging phenotype of accumulating deficits [203–205]. Aging *per se* is a key risk factor for aberrant developments, where a total of 92 age-related diseases have been described to date [206]. Intriguingly, individuals with similar age-of-onset showed higher genetic similarity in a comparative study [207]. Prominent examples for age-related diseases include metabolic disorders such as diabetes or hypertension, cardiovascular dysfunction, cancer, multiple chronic conditions and neurodegeneration, together posing a serious threat to global healthcare systems [208; 209]. Yet, aging is remarkably heterogeneous and proceeds non-linearly on the population level, motivating the development of a multi-factorial, cellular model to understand traits of human longevity [210–212]. With higher ages more and more cells of the human body fall into a mode of cellular senescence, failing to counteract the mutational burden [213; 214]. Additionally, a gradual loss of chromosome telomeres causes a deficiency in stem cells leading to increased apoptosis and insufficient self-renewal capacity [215]. Tremendous efforts have been directed to unravel the cellular mechanisms of aging, leading to the discovery of key apoptosis modulating pathways such as mTOR [216].

1.3.1 Gene expression trajectory with age

In order to better understand the development of age-related diseases, deciphering a solid picture on normal or healthy human aging at the gene-level is inevitable [217–219]. Previous work led to the first description of the gene expression hallmarks of cellular aging [220]. A crucial aspect of aging research is to perform deep longitudinal profiling in order to determine temporal on-set and specificity of alterations in DNA methylation, mRNA expression or protein abundance, that means from a multi-omics perspective [221–225]. Respective studies were able to predict the existence of human ageotypes, personal aging markers and lifespan limits [221; 226]. Non-linear changes in mRNA and protein levels, which were among others linked to inflammation, immune system activity and energy metabolism, were found to occur with a high tissue-specifity along the human and mouse lifespan [222; 227; 228]. In particular, a complicated tissue-dependent inflammatory signaling between the various cell types of the immune system as for instance between monocytes and T cells seems to play a key role in age-related diseases [229–233]. It is increasingly appreciated that many changes in gene expression observed upon aging and disease are controlled by ncRNAs, especially miRNAs and lncRNAs [5; 234; 235]. Human miR-34, for example, is a well-characterized miRNA exhibiting a strong positive correlation of expression with age and is implicated in cancer and neuron survival [6; 234]. It is therefore deemed promising to look more closely into the function of miRNAs in prevalent central nervous system (CNS)-associated diseases such as Alzheimer's disease (AD) and Parkinson's disease (PD) on the search for novel diagnostic and therapeutic approaches [236–243].

1.3.2 Neurodegenerative disorders

Severe nervous system diseases account for the second-most number of deaths worldwide with significant regional variation [244; 245]. Neurodegenerative disorders are a type of detrimental neurological aberrancy defined by varying protein pathologies and gradual loss of CNS capabilities such as motor and sleep function, cognitive abilities and progressive dementia [246; 247]. Both AD and PD belong to this type of disorder with the most and second-most number of cases, respectively [248; 249]. Conservative estimators report that several million people are already affected globally and the incidence continuous to rise (Figure 1.6). Many rational classification schemes have been proposed to set the diseases apart based on clinical and symptomatic criteria, physiological and anatomic features, or molecular neuropathology [246]. Early diagnosis is substantially hampered by the presence of long asymptomatic precursor stages [250]. Even though much work has already been accomplished to fully characterize the clinical and pathological features, only a limited number of medications and no single curative treatment are currently available. The overwhelming proportion of cases are classified as idiopathic but a small percentage is attributable to disease-specific genetic mutations and familial predisposition [251]. Strong neuroinflammation and activation of immune cells as well as the occurrence of other chronic conditions adopted during aging accompany neurodegenerative disorders [252; 253]. However, the cellular origins of AD and PD fundamentally differ by a selective vulnerability of neurons caused by distinct protein pathologies, as explained in the following.

Alzheimer's disease (in latin Morbus Alzheimer) is characterized by intracellular accumulation of tau protein and extracellular amyloid-β plaque formation pathology in the grey matter region of the cerebral

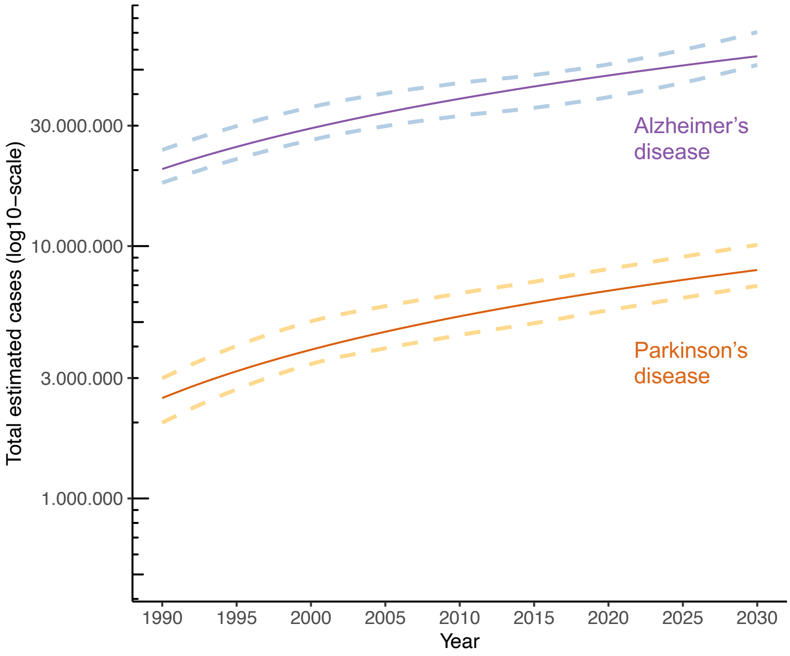


Figure 1.6: Observed and projected case numbers of Alzheimer's disease (purple line) and Parkinson's disease (orange line) in the global population between 1990 and 2030. Incidence rates and counts were obtained from [264; 265].

cortex [249]. Patients in the mild cognitive impairment (MCI) precursor stage or with terminal AD dementia show quantifiable cerebral atrophy, especially in the hippocampus region [254]. However, both plaque accumulation and atrophy are tissue-wide changes that occur during normal aging but are exacerbated and even accelerated in AD patients [254; 255]. The gradual accumulation of region specific neurofibrillary tangles was first described by Braak in 1991, allocating disease severity into the six so-called Braak stages [256].

Furthermore, genetic factors that relate to the protein pathologies such as the three alternative alleles of the gene Apolipoprotein E (ApoE), which modulate disease risk, have been identified [257]. Recently it was found that brain-resident, activated microglia both phagocytose and acummulate lipid-droplets and show a strong overlap of inflammatory markers with those previously described in AD [258]. The disease-associated microglia specifically express TREM2, which is necessary for proper stress response and accumulating protein clearance [259]. Intriguingly, AD is associated with a partial breakdown of the blood-brain barrier among others caused by cerebral amyloid angiopathy [260]. These findings suggest an altered substance and signaling exchange between these otherwise strictly separated biofluidial systems, questioning the role of the peripheral immune system in neurodegeneration [261–263].

Parkinson's disease (in latin Morbus Parkinson) in contrast to AD, was first characterized by the formation of Lewy bodies and a selective but severe loss of dopamin-producing (dopaminergic) neurons in the *substantia nigra*. The common PD pathology arises from toxic accumulation of misformed α-synuclein (SNCA), likely originating at the presynapse terminals and postsynaptic dendritic spines [266]. Braak also proposed a six-stage classification scheme for PD that correlates with symptomatic severity by tracking SNCA-immunopositive Lewy bodies, Lewy neurites and local lesions that ascend the brain stem [267]. However, in recent years emerging aspects caused a shift in this paradigm to prefer SNCA accumulation induced synaptic dysfunction as primary cause of the PD phenotype over the neuronal loss associated with Lewy body formation, which also occurs in normal aging [268–270].

The dysfunction of dopaminergic neurons causes the distinctive tremor and motor symptoms, but often dozens of other quality of life restricting symptoms such as sleep disruption, constipation, and anosmia are reported [248]. Some patients undergo cognitive decline and develop dementia, showing remarkable similarity to the AD pathology but with accumulation of lewy bodies in the cortex, also known as dementia with lewy-bodies [271].

Similar to AD, several genetic risk factors involving the genes *SNCA*, *LRRK2*, *GBA*, and *MAPT* were described, although most cases of PD are idiopathic [272; 273]. The cellular and molecular hallmarks of the disease include mitochondria and reactive-oxygen species dysregulation, calcium homeostasis, synaptic and lysosomal dysfunction, protein misfolding as well as apoptosis and neuroinflammation [274].

Basic research in AD and PD was naturally focused on the cells surrounding the affected brain regions, which is mirrored by the established clinical practice reaching diagnosis by comprehensive clinical and imaging features. Nevertheless, curative or preventive treatments are still lacking, motivating the search for new diagnostic and prognostic tools through massively extended biomarker discovery frameworks. To this end, large-scale multi-centered cohorts were established for each of the disease such as the Parkinson's Progression Markers Initiative (PPMI) by the Michael J. Fox Foundation. Low-invasive and cost-effective biomarkers for early diagnosis are of particular interest, where peripheral sampling of specific proteins and (extracellular) RNAs are currently believed to bear great potential [275]. At best, a good candidate biomarker would robustly correlate with some of the many objective disease rating scales such as the Unified Parkinson's Disease Rating Scale (UPDRS) or Mini–Mental State Examination (MMSE) [276–278]. Due to their high bioavailability and well understood targeting principles, miRNAs are profound candidates for low-invasive fluid biomarkers, for which an overview is provided in the next section.

1.3.3 RNA Biomarkers

Reliable biomarkers for neurodegenerative diseases are in great demand and several promising candidates are evaluated in pre-clinical studies for both AD and PD, although many have failed validation [279]. Most biomarker candidates are targeted around the specific

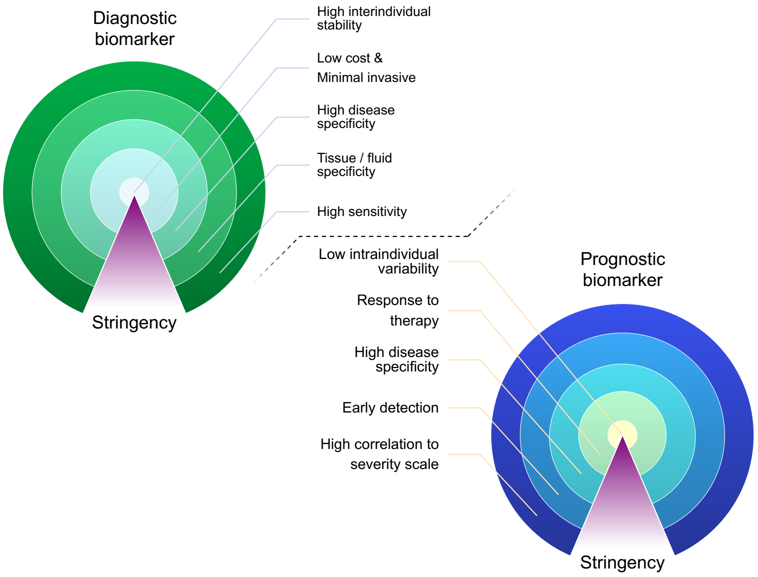


Figure 1.7: Selection criteria for reliable diagnostic (left) and prognostic (right) biomarkers. Criteria were derived from [299; 300] and sorted by increasing stringency from outer to inner circles.

protein pathologies observed in each disease, oftentimes using whole-blood, plasma, serum or cerebrospinal fluid (CSF)-based detection methods [280; 281]. In addition, expensive imaging techniques such as positron emission tomography (PET) provide diagnostic value [279; 282].

RNA biomarkers for human disease have been extensively studied, promising better cost-effectiveness and specificity but unfortunately showing lower stability in biofluids due to circulating RNAses [283]. Blood-borne miRNAs and other ncRNAs, however, were found to be remarkably stable and bioavailable, showing well-reproducible expression profiles [284–287]. Circulating miRNAs are commonly found in serum, plasma, PBMCs and extracellular vesicles [288]. Yet, confounding effects in blood-derived RNA expression profiles promoted an over-report of non-specific and non-reproducible miRNA biomarkers [160; 289–291]. Especially domain-specific meta-analysis studies highlighted an overly disconcordant picture of miRNA biomarkers in the literature [292–294]. Thus, to increase specificity, miRNA biomarkers are nowadays frequently reported as part of signatures or so-called panels using machine learning methods [3; 290; 295–298]. Different types of biomarkers exist, each with specific requirements (Figure 1.7).

Besides the traditional application in cancer research, potential miRNA biomarkers were proposed for infectious, cardiovascular, and neurodegenerative diseases [286; 301]. Intriguingly, miRNAs are able to cross tight cell boundaries such as the blood-brain barrier, presumably due to their tiny extent and encapsulation in exosomes [302; 303]. For example, *hsa-miR-9-5p* and *hsa-miR-124-3p* are enriched in the brain but were found in serum exosomes of patients suffering from acute ischemic stroke [304]. It was only during the last ten years that miRNA biomarkers gained traction in neurodegenerative disease research, and many challenges in the field remain to be solved [275; 305]. First, most studies involve only a limited number of patients or replicates [306]. Second, some miRNAs might bear more prognostic than diagnostic value, however longitudinal expression profiling is rarely performed. Third, a paired assessment of mRNAs and ncRNAs from the same biological sample should ideally be performed to capture the state of disrupted regulatory networks. As disease progression in AD and PD is often highly individual, research on age- or stage-specific and prognostic markers offers entirely new aspects to predict a patient's disease risk [8; 307]. Solving these challenges along the way using novel bioinformatics-driven and standardized approaches is important, with the current lack of such possibly explaining the very limited success in translating miRNA biomarkers to the clinics [308; 309].

1.4 Bioinformatics

Even though the gigantic increase in molecular data reaching exponential scales since the late 1990s and early 2000s is commonly attributed to the origins of bioinformatics, the actual history dates back to the 1950s [310]. Back then, computational biology arose at interface of biology and computer science, both of which faced tremendous technological advances in the last 70 years [311]. Today, bioinformatics is a diverse, multi-faceted and therefore hard to delineate field that is closely intertwined with molecular biology and clinical research. Among the grand challenges are the development of standard databases and computational models supporting interpretable data analysis, to provide reproducible and scalable standard-workflows as well as developing efficient software tools based on modern algorithms [312; 313]. The almost paralleled increase in computational power and amount of generated data required continuous overhauling of software, fostering a high level of innovation in the field. As for this profession replication of results is a major concern and following the reproducibility crisis, gets increasingly appreciated [314–316]. In addition, the massive upsurge in data-driven industrial research renders it even more difficult to distinguish between branches of data science and traditional bioinformatics, which show trends of mutual convergence [62; 317; 318]. Thus, several rather generic and wide-spread application platforms have emerged, each of which is introduced in turn.

1.4.1 Application types

Databases are systematic collections of almost any kind of scientific data — and at the very heart of many bioinformatics applications. The National Center for Biotechnology Information (NCBI) lists more than 40 resources in the core collection, together covering a broad range of -omics and clinical data sets. Moreover, the journal Nucleic Acids Research (NAR) dedicates one issue a year with 28 iterations so far for publishing peer-reviewed databases that are of broader interest for the community [319]. Therefore, a variety of standard databases exist in each of the specific domains. For example, Ensembl, RefSeq, UCSC Genome Browser, ENCODE, UniProt, and PDB are the largest reference collections for genome, transcriptome and proteome sequence and function annotation in vertebrates [320–325]. For ncRNAs in general, the resource RNAcentral whereas in particular for miRNAs, the databases miRBase, miRCarta, and MirGeneDB are most popular and together established a common nomenclature [101; 326–328].

Figure 1.8 shows the canonical steps covered in a standard miRNA analysis pipeline, and for each dozens of specialized database applications have been proposed. Typically, popular databases offer rich custom and interactive analysis functionality to support interpretation instead of just being pure data repositories. The scientific databases developed by the group of *Clinical Bioinformatics* at Saarland University collectively attracted 22.622 individual visitors from all around the globe within 1 year, causing a total of 53.822 unique page views (Figure 1.9).

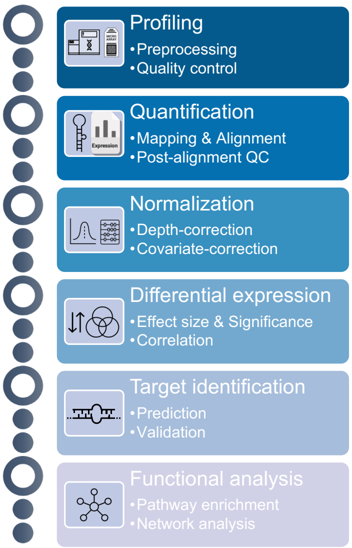


Figure 1.8: Workflow performed in a standard miRNA analysis pipeline of small RNA profiling data sets. The six major steps are profiling, quantification, normalization, differential expression, target identification and functional analysis, each comprising multiple subroutines.

Web servers are web-based tools performing repetitive analyses or complex workflows based on user-provided input and that complement static data collections. The need for bioinformatics-driven web servers is ubiquitous, so that NAR decided to dedicate another annual special-issue exclusively to web servers with 18 iterations completed so far [329]. Traditional sequence alignment tools such as BLAST rank among the most popular web server applications, generating millions of user-queries each year [7; 330]. However, a high availability of user-friendly development frameworks caused an inflation of available web servers showing low half-life times [7]. Therefore, new standards of good scientific practice are emerging to increase tool availability. Among others, common testing strategies, regular maintenance, comprehensive tutorials, and minimum security measures are required.

Computational models. Bioinformatics analyses inevitably rely on computational models combining different sources of data to predict a certain feature of interest. Also, they often incorporate prior statistical and biological knowledge. For instance, hidden markov models simulate stochastic and sequential processes with state transitions governing hidden factors [331]. Markov chains were used by ENCODE to assign function to regulatory elements in the genome based on epigenomic sequencing data. Thousands of models of varying flavors and scope have been published so far. Yet, specific assumptions and limitations are intrinsic to every such, and failing to know those is one of, if not the most frequent source of errors in the scientific literature. Since most public databases and web servers involve the application of computational models the potential impact is vast. Moreover, while use *a priori* information tends to improve prediction performance, models are at risk to pick up hidden bias, having sig-

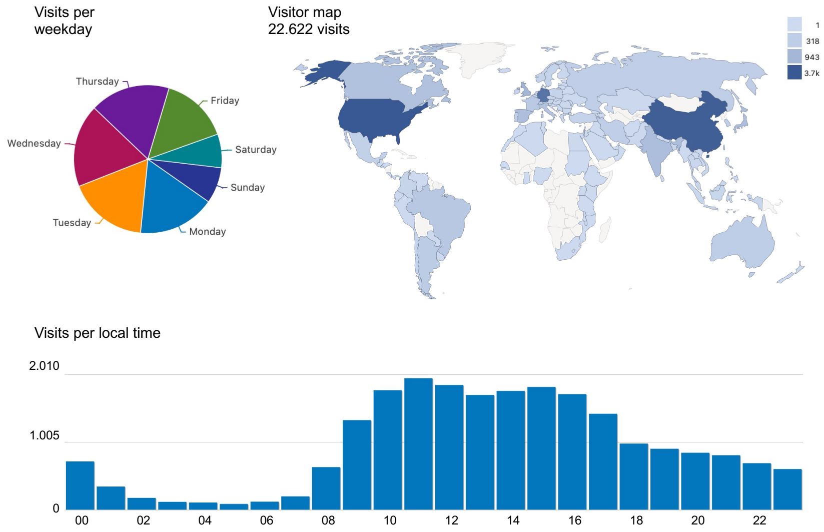


Figure 1.9: Combined usage statistics for seven peer-reviewed databases published by the *Clinical Bioinformatics* group at Saarland University. Presented data was collected for one year starting on the 31st of May in 2020.

nificant implications especially for machine learning [332–336]. Bias hidden in computational models contributes to a low success rate in translational research, where they fail to generalize to a larger and often more diverse set of observations. Therefore, it is crucial to assess as many confounding factors as possible during model development but also to explore new methods to detect such.

Algorithm, software and programming standards. As standardization is a common task in bioinformatics, rich libraries containing efficient implementations of algorithms exist in the community. Frequently they appear in combination with a set of standard data formats. For example, the Binary Sequence Alignment/Map format (BAM) accompanied by the Samtools software suite is completely technology independent and thus part of almost any sequencing pipeline, and has been cited more than 5.500 times according to the Web of Science (WoS) (accessed on June 19, 2021)[337]. Similarly, the Browser Extensible Data (BED) format and General Feature Format (GFF) storing genomic features, i.e. annotated locations and intervals, and the associated bedtools suite ($>$ 8.200 citations at WoS; accessed on June 19, 2021) are *de facto* standard in any genome annotation pipeline [338]. While downstream analysis and visualization is primarily performed using scripting languages such as *R* or *Python*, high-throughput scenarios require more efficient solutions. In the context of this thesis, most code requiring computational efficiency for sequence analysis was developed using the module-based Seqan C++ template library [339].

Pipelines. Following the successful development of key computational methods within domain-specific scopes, there is an unprecedented need for pipelines. A pipeline is a complex, non-linear workflow to transform given input data to output data under a set of transformation rules. Features inherent to each pipeline framework are definition of input format(s), output format(s), transformation command(s), necessary and optional parameters as well as resource management, e.g., space and time constraints. A pipeline is commonly interpreted as directed acyclic graph (DAG). Abstract definitions can be formulated in the Common Workflow Language (CWL) [340]. The most common workflow managers are Snakemake, Nextflow and Galaxy that interpret CWL or custom syntax in order to execute fully-defined pipelines [341–343]. Still, pipelines *per se* are deliberately kept abstract as to be independent from technical variables that would affect portability. Thus, to make results truly reproducible, they must be run in fully isolated environments. That is often achieved by combining any of the above frameworks with proper virtualization techniques such as Docker, Singularity, or software package management through Bioconda [344–346]. Each data transformation step is then performed on a selected compute node using automatic dependency deployment, enabling massive scaling through parallelization. Pipelines are becoming ever more popular and already have shown great success through commercial applications in biomedical industry. Together, the above introduced flexible toolbox of bioinformatics applications promoted fundamental advancements in molecular biology. Next, the actual role for some of these tools in sncRNA research and analysis is further elucidated.

1.4.2 Role in non-coding RNA research

Analyzing small RNA sequencing data using a high-throughput bioinformatics pipeline involves the application of dozens of methods and tools, which are configured using hundreds of parameters. Here, we focus on the standard workflow required for miRNA research, however, similar pipelines exist for other sncRNAs. The necessary steps can be basically grouped into *miRNA quantification*, *normalization and differential expression*, *target identification* and *in silico functional analysis* (Fig. 1.8). Each is introduced in more detail in the following sections.

MiRNA quantification The first step is to process raw high-throughput sequencing data eventually computing a miRNA to sample count matrix. Besides enforcing minimum base quality-thresholds, platform adapters are trimmed from sequencing reads and mapped against the target genome. Because of their small size, standard read lengths (75 bps) are often sufficient to cover entire miRNAs. They align against their precursor origins, forming characteristic read profiles with one major and one minor read stack per precursor. Then, known miRNAs are identified using reference genome annotation data from public standard databases. To date, hundreds of sequence and structure features of what makes a good precursor were described, and multiple tools for miRNA discovery and quantification from next-generation sequencing (NGS) data have been proposed [347]. Results presented in this thesis were primarily obtained using miRMaster for this particular job [348; 349]. Possible extensions to this analysis include isomiR and single-nucleotide variant (SNV) identification, miRNA arm shift quantification, contaminant analysis or novel candidate ranking [111; 347; 348; 350].

Normalization and differential expression Like for most sequencing protocols, raw miRNA quantification results should be normalized and scaled to account for technical variability, differences in sequencing depths and varying RNA quality. Several methods are commonly used for miRNAs including counts per million (CPM), reads per million (RPM), reads per million mapped miRNA (RPMMM), quantile-normalization or variance-stabilization transformation (VST) [351]. Most are suitable to compare miRNA counts between samples, although quantile-normalization is often favored for microarray-based profiling [352]. Inter-sample comparisons can be performed using statistical tests such as the Student's t assuming normal distribution, non-parametric Wilcoxon test or measures of effect size such as Cohen's d. The Shapiro-Wilk test can be used to judge the normality of miRNA expression before continuing with DE analysis. Good scientific practice involves a prefiltering of miRNAs with a very low expression as to yield a set of well detected features [160]. The significance of relationships between miRNAs and technical or biological covariates can be judged under the F test via an Analysis of Variance (ANOVA) or with correlation tests. Importantly, downstream applications as for example principal component analysis (PCA) tend to return better results for similarly distributed features, i.e. having equal mean and variance (z-scores), otherwise results are primarily driven by a successive hierarchy of expression strength.

Target prediction Following quantification and normalization, the next main step is to perform targetome prediction for each miRNA [347]. In high-throughput profiling studies several hundred miRNAs are typically of greater interest, thus *in silico* target prediction is often preferred over expensive target validation techniques in the early stages. In principle, every target prediction tools utilizes one or several of the four core features to varying extents; complementary sequence alignment, structural features, sequence homology and conservation or site accessibility [1]. It was previously shown that none of the many tools available consistently outperforms the others, although some seem to perform considerably better in certain setups. In addition, the required runtime varies on several orders of magnitude, precluding poorly scalable tools from any high-throughput pipeline. However, unbiased performance benchmarks for miRNA target prediction tools are notoriously hard to achieve in general. High quality validation sets for negative targets, i.e. non-functional targets that still contain a binding site are rare, even though many such non-functional sites are predicted to exist [353]. As a result of frequently underpowered validation, most tools predict a couple of hundred to thousand targets per miRNA, accumulating to several million miRNA-target pairs in human [354; 355]. It is therefore deemed necessary to develop new methods for prioritizing likely targets over low-quality ones. For instance, grouping predicted genes into enriched pathways or reaction networks, which are listed in popular databases such as Gene Ontology (GO) or the Kyoto Encyclopedia of Genes and Genomes (KEGG), has been shown to increase specificity of target predictions by filtering false positives [6; 356–360].

In silico functional analysis The last step is to combine findings about (dysregulated) miRNA expression and predicted / validated targets for a context-specific interpretation of function. The focus of such an analysis varies in several aspects, sometimes turning it difficult to discern the actual miRNA *function*, since different modes of action and regulatory dependencies are strongly interwoven [361; 362]. Substantial evidence suggests the existence of additional rules governing miRNA targeting for phenotype regulation *in vivo* [363–365]. These observations call for systematic methods and resources facilitating a comprehensive functional characterization of each mammalian miRNA [362]. To begin with, enrichment analysis is one of the most popular applications in miRNA research, borrowing statistical methods like gene set enrichment analysis (GSEA) from earlier approaches developed for whole-transcriptome studies [362; 366]. Furthermore, miRNA-pathway databases offer pre-computed annotations for miRNAs using validated or predicted target genes that are enriched on certain pathways from knowledge databases like GO or KEGG [2; 147; 362]. Also, more specialized databases such as miR2disease list specific associations for miRNAs being implicated in human disease [362; 367]. Finally, *in silico* reconstruction of miRNA-target coexpression networks as bipartite graph is frequently performed to generate further insights into regulatory dependencies and network motifs [362]. More explanations on methodological details are given in the upfollowing sections.

1.4.3 Computational methods

In the following paragraphs essential methodological and statistical background knowledge underlying the resources developed in this thesis (miRNA pathway dictionary database (miRPathDB) 2.0; miRNA Enrichment Analysis and Annotation tool (miEAA) 2.0) are introduced [2; 4].

Set statistics Two main techniques exist to judge whether a given set M of feature-annotated items is significantly covered in another sample set S. The abstract question is whether S contains more, less, or similar featured items than M than one would expect to observe in a random process. First and foremost, over-representation analysis (ORA) can be performed using the hypergeometric test, which is in fact used to judge whether a given statistical sample stems from a certain base population or not. There are four parameters involved; the background population size, the number of times a certain feature is contained within the population, the sample size and the number of times that same feature is contained within the sample. By enumerating all possible sampling combinations, a likelihood (p-value) for the null hypothesis that the data was randomly sampled from the population, is obtained. If this p-value is sufficiently small, the likelihood to observe the sample distribution or a more extreme one under the null hypothesis is small and therefore gets rejected.

The second main technique is not based on an unordered set but based on the rank of each item on a sorted list, also known as GSEA [366]. Thereby, a significant accumulation of a given feature at the beginning or the end of the item list should be tested. Briefly, the GSEA algorithm involves the enumeration of a running sum, which is equal to zero at the beginning of the list. The sum is then increased or decreased by a pre-defined amount depending on whether the item at the current rank position has the feature of interest or not. The significance and a therefore a p-value is determined from the maximal absolute deviation from zero, the so-called enrichment score. The p-value can be estimated through permutation or even calculated exactly using dynamic programming, and depending on the desired level of precision the one or the other is computationally more expensive [368]. Optional weights for each item in the list can be given, corresponding either to an unweighted or weighted Kolmogorov-Smirnov test with fixed or dynamic steps of the running sum, respectively [4].

Both aforementioned techniques were extensively applied in transcriptomics studies and during the last ten years adapted to other fields, including ncRNAs and miRNAs [369; 370]. While ontology-based methods were shown to improve detection of batch effects and interpretation of genomic or transcriptomic data sets, the translation of enrichment analysis to miRNA research yielded critical concerns about confounding bias [371–373]. Reasons are the significant imbalance of validated targets across all the known pathways and a general over-representation of cancer in the miRNA literature [372]. In addition, the level on which the comparisons using set statistic are performed (miRNA, gene, pathway), strongly influence the results. As a remedy, best practices have emerged that were also taken into account during the implementation of miEAA and miRPathDB [2; 4]. Finally, when testing more than one feature or category of interest among the same set of items, measures to correct for multiple hypothesis testing must be implemented, otherwise significant results occur simply by type-I error rate inflation. Prominent examples are the Bonferroni correction or the false-discovery rate controlling procedure by Benjamini and Hochberg [374].

Optimization Many practical programming tasks in Bioinformatics can be reformulated as optimization problem, allowing elegant and concise formulations of highly complex models [375; 376]. The technique was originally described in computer science and is broadly used in economics. A standard optimization problem basically involves a linear target function $f(x)$ depending on a fixed number of input variables and one or multiple side equations or inequalities, also known as constraints. Optimization problems with side constraints that require equation satisfying variables to be integer are called integer linear program (ILP). The feasibility of such a problem and the existence of an optimal solution is of primary interest, however, it has been shown to be a hard problem in general and therefore is of complexity *NP-complete*. Several independent techniques have been developed to solve ILPs exactly such as cutting planes or branch and bound methods, or approximately to a certain level of precision using heuristic approaches. Practical problems tend to have numerous optimal solutions, with important implications for the functional interpretation of gene selection models in the context of genomic or transcriptomic studies; disjoint sets may solve the same requirements equally well obfuscating their genuine importance. For instance, "what are the minimal number of miRNAs to cover a certain set of target genes", is a question that can be formulated as an ILP, and a solution to this problem was implemented in miRPathDB 2.0 [2].

Networks Graph theory has found broad applications in bioinformatics and was successfully used for creating genome assemblies, biological networks, or protein-complex simulations [377–379]. Every graph can be described as tuple $G = (V, E)$ with the set V being the nodes (vertices) and the set E denoting the undirected or directed links (edges) between any two vertices, with $E \subseteq \{(v, v')\}$ where $\forall v, v' \in V$ and $v \neq v'$. Via algorithms interesting properties about large graphs can be computed such as the shortest path between two nodes, the existence of fully connected subgraphs, so-called cliques, or biological network motifs such as feedback loops [380; 381]. Regulatory networks of miRNA-target interactions were often modeled as bipartite graphs, i.e., a graph with two types of nodes where only edges between nodes of different types exist, but not within one type [382]. In such networks, miRNA and gene nodes are often context-specifically selected based on their differential expression or association with a particular disease [383]. The edges, on the other hand, can be derived by means of multiple sources, for example, experimentally validated miRNA-target interactions or significant anti-correlation between miRNA and mRNA expression [384]. MiRNA-mediated networks are intrinsically redundant and exhibit specific degree distributions, where the topology can be indicative of disease-associations [385–389]. In addition, tools tailored for miRNA analysis that spawn networks based on established knowledge databases have been proposed, combining also different layout clusterings for comprehensive visualization [390].

Machine learning and data science The information content of certain features, e.g. gene or miRNA expression, as diagnostic or prognostic signatures predicting a certain clinical outcome can be assessed using machine learning (ML). The techniques combine advanced mathematics with unmatched performance of modern computers to uncover hidden relations from large data sets, a task that is impossible for humans to perform manually. Two primary purposes exist; prediction and inference. The former is used to determine robustness and accuracy, including sensitivity and specificity of individual feature sets. The latter refers to the interpretation of important features and generating insights about the relation between the dependent and independent variables of interest. For instance, promising results have been described for the detection of lung cancer with 91, 4% accuracy using ML-derived signatures of circulating miRNAs [295]. However, numerous pitfalls exist that frequently cause non-reproducible findings in biomedicine and frustrate translational research, and consequently predictive models should be evaluated from independent viewpoints [391; 392]. Errors might arise from undetected confounding factors or by mixing up correlation with causation [393]. In addition, improper handling of imbalanced data classes, e.g. case and control, and preselection of features before performing cross-validation introduce bias and obfuscate prediction performance.

Dozens of models have been evaluated for miRNA signature identification, ranging from linear regression to deep neural networks. Still, tree-based models such as gradient boosting machines (GBM) seem to frequently outperform others. Generating a GBM involves a sequential training of shallow decision trees (weak learners), eventually reaching an ensemble of weak but independent predictors, with a strong collective prediction performance. For instance, GBMs were used to accurately distinguish AD patients from controls based on expression levels of 21 circulating miRNAs [3].

1.4.4 *Role in clinical non-coding RNA research*

Clinical bioinformatics provides domain-specific expertise of computer science to streamline clinical research and applications. It is characterized by high-throughput data generation with quick turnarounds, high standards for data curation and robust modeling [394–396]. Many challenges have to be overcome to make precision medicine feasible [397]. Rigorous validation and reproducibility of sequencing-derived results is a major concern in the field, as for example clinical covariates like patient age or sex are known to influence -omics data. Taking into account the known biological or technical covariates is thus deemed essential [398]. Furthermore, it encompasses appropriate use of biostatistics to test differential data distribution together with pre-checking of statistical assumptions. For biomarker studies replication using an independent cohort along with the use of an independent technology is considered best practice. Standardization of data processing pipelines is thus another important challenge [399].

As compared to transcriptomics applications, ncRNA research is not yet as mature and specific challenges need to be overcome for more efficient clinical research. Most classes of ncRNAs are under-characterized, especially in non-primate species, leaving much room for improvement to reduce existing uncertainty in modelling ncRNA pathways [400]. Since many miRNAs are not yet functionally characterized and together with the existing literature bias towards cancer-associated miRs, interpretation of small non-coding assays is a challenging task [400; 401]. They also obey different levels of bioavailability and pathways to permeate physiological barriers in biofluids as compared to mRNAs [402]. Due to their broad versatility in mediating essential gene regulation, miRNA-based therapeutics have shown low success rate so far, motivating the development for new *in silco* methods and resources that take into account all relevant aspects.

2
Goals of the PhD thesis

The aim of this thesis was to advance our understanding of miRNA-induced gene regulation in a differentiated, step-wise fashion. It was hypothesized earlier that these gene regulatory mechanisms differentially govern cellular pathways at either homeostatic conditions, during aging or in age-releated diseases of *Homo sapiens*. The eight main publications included herein systematically approach this goal. Successive interleaving of basic research on miRNA-target gene relationships, methodological advances, and applied research on neurodegenerative diseases, facilitated a comprehensive analysis of most human miRNAs in the aging context. In the following, each publication is thus briefly introduced in chronological order and the most important accomplishments are mentioned. An overview of the covered projects is given in Figure 2.1 and respective publications are presented in full in Chapter 3.

To begin with, by comparing almost 100 tools for *in silico* miRNA-target prediction, a comprehensive overview of the most effective targeting features was generated as well as assumptions and intrinsic biases of these models were unveiled [1]. Even though four frequently cited mechanisms are utilized for target prediction, output and parameters of most published tools are difficult to interpret. In addition, individual prediction features show considerably varying importance based on the biological context. Subsequently, a novel version of miRPathDB was developed, including a revised selection of predicted and experimentally validated miRNA-target pairs, generating the so far largest database of statistically enriched miRNA pathways in human and mouse [2]. New interactive analysis functionality enables straightforward dissection of large panels of interesting miRNA and gene candidates as often encountered in high-throughput sequencing studies. We then compiled a well-performing blood-based miRNA-signature for AD detection using machine learning methods together with a functional interpretation of the selected miRNAs through miRPathDB [3].

Since miRNAs are increasingly described in terms of more specialized features, as for instance cellular localization, tissue specificity or interaction with other regulatory players such as TFs, a new major release of miEAA was developed [4]. It now allows to annotate a set of miRNAs from ten different species by a variety of new, potentially

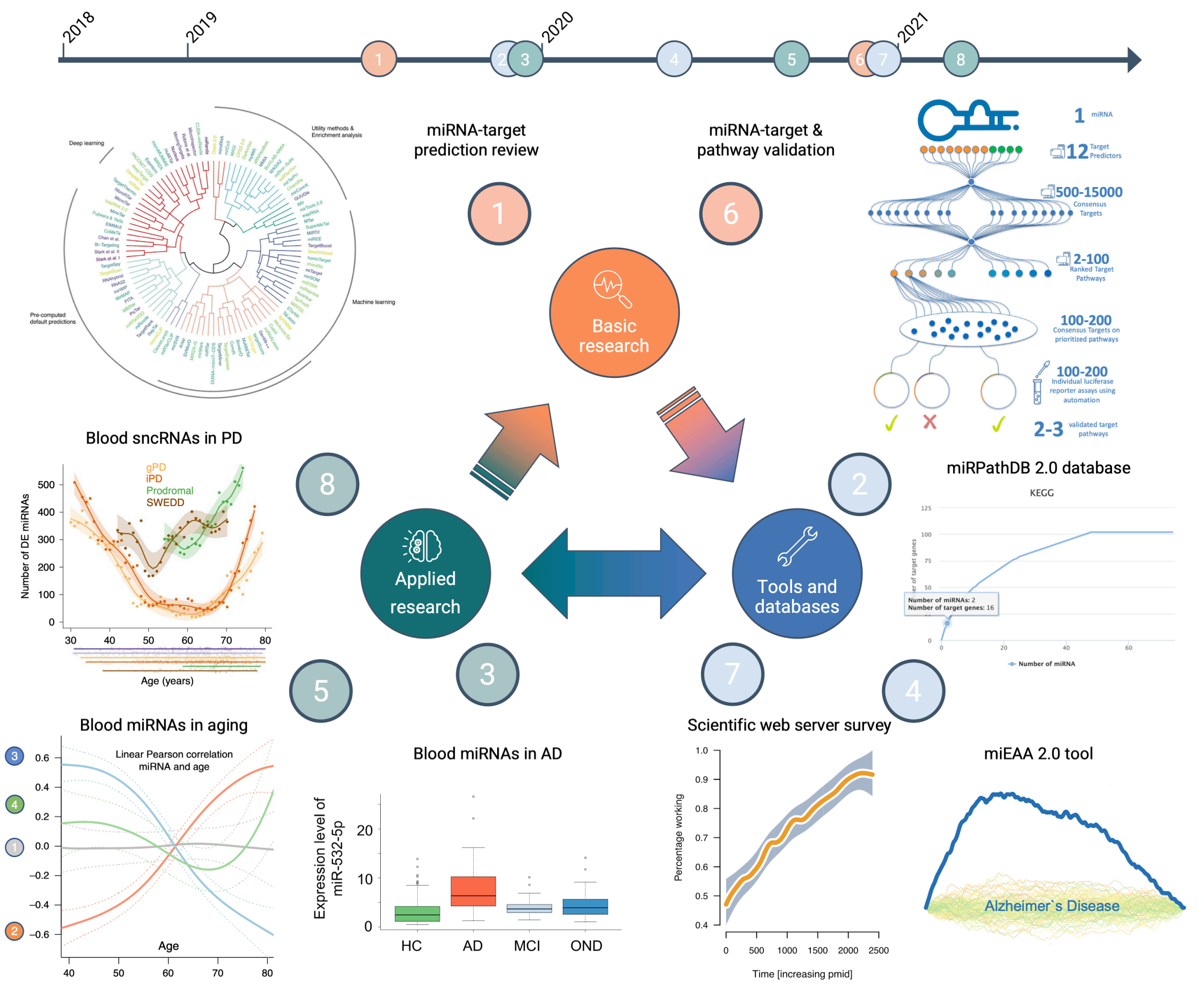

2018
2019
2020
2021
1
2
3
4
5
6
7
8
Deep learning
Utility methods & Enrichment analysis
Machine learning
Pre-computed default predictions
miRNA-target prediction review
miRNA-target & pathway validation
1 miRNA
12 Target Predictors
500-15000 Consensus Targets
2-100 Ranked Target Pathways
100-200 Consensus Targets on prioritized pathways
100-200 Individual luciferase reporter assays using automation
2-3 validated target pathways
Basic research
Applied research
Tools and databases
Blood sncRNAs in PD
gPD
iPD
Prodromal
SWEDD
Number of DE miRNAs
Age (years)
miRPathDB 2.0 database
KEGG
Number of miRNAs: 2
Number of target genes: 16
Number of miRNA
Blood miRNAs in aging
Linear Pearson correlation miRNA and age
Age
Blood miRNAs in AD
Expression level of miR-532-5p
HC
AD
MCI
OND
Scientific web server survey
Percentage working
Time [increasing pmid]
miEAA 2.0 tool
Alzheimer`s Disease

Figure 2.1: The main projects presented in this thesis were classified either as *basic research*, *tools and databases* or *applied research*. Arrows in between the three groups indicate flow of knowledge gained during development. Projects were numbered according to increasing chronological order of publication, which is also depicted at the top.

enriched aspects derived from the literature, further broadening the application scope. From a technical viewpoint, miEAA was also completely redesigned allowing users automated access to the algorithm and facilitating inclusion in custom bioinformatics pipelines. Both miEAA 2.0 and miRPathDB 2.0 were used extensively to interpret large sequencing and microarray data sets as shown hereafter.

Seeking to understand how blood-borne miRNAs are affected by human aging, a large-scale analysis of 4.393 microarray samples revealed strong disease- and age-specific but less gender-specific changes in physiological patterns of expression [5]. Age-declining miRNAs significantly overlapped with those described being downregulated in AD. Most importantly, miRNA expression was observed to change in a non-linear way along human lifespan and those changes were partially reflected in anti-correlated protein levels of affected target mRNAs.

Yet, experimental validation of entire candidate target pathways for one or several miRNAs was still considered impractical, especially with high-confidence but low-throughput methods like reporter assays. Therefore, a novel workflow combining multiple computational and experimental approaches was developed to specifically narrow down the most likely set of target genes or target pathway(s) for a preselected miRNA [6]. Because most target prediction tools suffer from Type-I error inflation, experimental validation previously yielded low rates of success and specificity. By executing this new workflow for *hsa-miR-7-5p* and *hsa-miR-34a-5p*, both of which are implicated in PD, unprecedented target validation success rates were obtained, both on the per-gene and per-pathway level. In order to support users in performing the workflow, a new web-based tool for target plasmid design and a database containing validated miRNA-target pathways were published alongside.

Having gained fundamental experience from developing multiple web-based tools and databases, each accessed by thousands of users annually, an invited survey and summary paper was published in Nucleic Acids Research [7]. After continuously tracking the availability of a semi-automatically curated set of several thousand scientific web servers, a 50% availability ratio just ten years following initial publication was observed. High citation rates correlated with high availability and good maintenance, and therefore best practices for web server development were derived and discussed.

Finally, the most complete, longitudinal description of circulating sncRNAs in PD was performed [8]. In total, 5.450 small RNA sequencing samples were analyzed, yielding a comprehensive atlas of miRNA expression in human blood cells and pivotal candidates for diagnostic and prognostic biomarkers of PD. Similar to our previous observations, miRNAs were affected non-linearly across the lifespan in PD and recapitulate all molecular hallmarks of the disease. Using paired RNA-seq, the impact of disease progression-associated miRNA modules on the transcriptome was modeled *in silico*, motivating a further experimental characterization of these pathways in the future. The main results from the larger PPMI cohort were then successfully replicated in the independent National Centre for Excellence in Research on Parkinson's Disease (NCER-PD) cohort from Luxembourg using microarray technology instead of sequencing. This study was selected as cover story for the 2021 March issue of Nature Aging. Figure 2.2 shows the original cover artwork of Volume 1 Issue 3 (March 2021) as it appeared in the journal
(https://www.nature.com/nataging/volumes/1/issues/3).

Despite the main collection of papers presented above, a total of 24 ancillary manuscripts were published in related fields. In brief, topics comprise basic research and reviews on miRNAs and targets [105; 118; 145; 350], miRNA physiology [193; 403], circulating miRNA dysregulation in cancer [404–406], properties of small RNA sequencing technologies and data [91; 154], new miRNA databases and repositories [188; 327; 407] as well as novel web servers for miRNA analysis [111; 349; 390; 408]. In translating our observations from the summary and survey report we created a monitoring web service, which continuously creates long-term availability reports for thousands of peer-reviewed web tools [7; 409]. In the broader context of gene regulation, two collaborative projects on integrative models for transcription factor binding and gene expression prediction were successfully completed [31; 410]. Last but not least, following an unprecedented global pandemic caused by the human SARS-CoV-2, a scientometric literature analysis and two transcriptome studies on the emerging coronavirus disease 2019 (COVID-19) were published [411–413].

www.nature.com/nataging / March 2021 Vol. 1 No. 3
nature
aging
Small noncoding RNA dysregulation in Parkinson's

Figure 2.2: Cover artwork of Volume 1 Issue 3 (March 2021) in Nature Aging. Original source caption: **Image: Dr. Valentina Galata, University of Luxembourg, 2021. Cover design: Lauren Heslop.** All required permissions were obtained.

3
Results

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

3.1 What's the target: understanding two decades of in silico microRNA-target prediction

Briefings in Bioinformatics
doi.10.1093/bib/bbz111
Advance Access Publication Date: 3 December 2019
Review article
This is a pre-copyedited, author-produced version of an article accepted for publication in Briefings in Bioinformatics following peer review.
The version of record Briefings in Bioinformatics, Volume 21, Issue 6, November 2020, Pages 1999-2010 is available online at:
https://doi.org/10.1093/bib/bbz111.

What's the target: understanding two decades of *in silico* microRNA-target prediction

Fabian Kern[1], Christina Backes[1], Pascal Hirsch[1], Tobias Fehlmann[1], Martin Hart[2], Eckart Meese[2] and Andreas Keller [1,3,4,5]*

[1]Chair for Clinical Bioinformatics, Saarland University, Saarbrücken, Germany
[2]Department of Human Genetics, Saarland University Hospital, Homburg Germany
[3]Center for Bioinformatics, Saarland University, Saarbrücken, Germany
[4]School of Medicine Office, Stanford University, Stanford, CA, USA
[5]Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA, USA

*To whom correspondence should be addressed.

Received on 18.07.2019; revised on 01.08.2019; accepted on 02.08.2019

Abstract

Motivation: Since the initial discovery of microRNAs as post-transcriptional, regulatory key-players in the 1990s, a total number of $2,656$ mature microRNAs have been publicly described for *Homo sapiens*. As discovery of new miRNAs is still on-going, target identification remains to be an essential and challenging step preceding functional annotation analysis. One key challenge for researchers seems to be the selection of the most appropriate tool out of the larger multiverse of published solutions for a given research study set up.
Results: In this review we collectively describe the field of *in silico* target prediction in the course of time and point out long withstanding principles as well as recent developments. By compiling a catalogue of characteristics about the 98 prediction methods and identifying common and exclusive traits, we signpost a simplified mechanism to address the problem of application selection. Going further we devised interpretation strategies for common types of output as generated by frequently used computational methods. To this end, our work specifically aims to make prospective users aware of common mistakes and practical questions that arise during the application of target prediction tools.
Availability: An interactive implementation of our recommendations including materials shown in the manuscript is freely available at https://www.ccb.uni-saarland.de/mtguide.
Contact: andreas.keller@ccb.uni-saarland.de, Ph. +49 (174) 1684638
Supplementary information: Supplementary data are available at *Briefings in Bioinformatics* online.
Keywords: microRNAs; target prediction; model interpretation; post-transcriptional gene regulation; model selection guide

Introduction

Almost twenty years ago Lee and Ambros initiated the era of small, non-coding RNAs by describing the existence of gene regulatory RNA sequences in *C. elegans* [1]. They found that short RNA sequences of about 22 nts in length, thereafter termed microRNAs, confer post-transcriptional regulation of target messenger-RNAs (mRNAs) by interrupting translation. In the mean time microRNAs have been described across a wide range of mammalian species [2, 3, 4]. From microRNA genes stem-loop like structured, ncRNA precursors (pri-miRNAs) are typically transcribed [5]. These are subsequently processed by the enzyme complexes Drosha-DGCR8 and Dicer to trim precursor ends and to cut away the central hairpin, respectively [5]. This process leaves behind a double-stranded stretch of RNA from which mostly one single-stranded form, either the $5'$ major (-5p) or the $3'$ major (-3p) mature miRNA, is selected as guide strand and to accumulate in the cytoplasm [6]. To confer gene regulation, mature microRNA transcripts are loaded into proteins of the AGO(1-4) family to bind sequence stretches in target molecules through extensive Watson-Crick base pairing [7]. As a result targets are either endonucleolytically cleaved, destabilized through exonucleolytic decay, or the ribosomal machinery is blocked [8]. While in plant species miRNAs are prefentially conserved and known to form hairpin precursors similar to metazoan miRNAs, enzymes associated with biogenesis and underlying modes of action are different [9, 10]. For example, AGO1 was shown to be the most effective member of the AGO family in *Arabidopsis* [10]. More prominently, miRNAs hybridize almost perfectly to their target site along the entire mature sequence, allowing for minor mismatches and bulges more often at the $3'$ end of the mature strand [11]. In turn, this property not only simplifies the search for target sites but also suggests a higher specificity of gene regulation in plants, resulting in a lower number of potential targets that can be predicted computationally. Also, the RISC binds mostly to coding regions, i.e. the open reading frame (ORF) of mRNAs in plants, whereas the $3'$ UTR preferentially harbors the target sites of RISCs in animals [11, 12]. Following these observations, target prediction in plants obeys different rules where exact alignments can be more strict with respect to mismatches and hence are faster to compute. Due to an asymmetric affinity of target binding in metazoan species, the seed region became the most important notion in the field; it defines the nucleotide sequence from position 2 of the $5'$ end to 7 in the microRNA that preferentially determines the targetome, i.e. a set of mRNAs targeted by a microRNA [13]. Since the true binding motifs may vary in length and involve mismatches or bulges, searching for potential targets in the human genome turned out to be a classical needle in a haystack problem. Therefore, comprehensive and efficient computational methods for target prediction are in great demand. Choosing one or several from the existing tools may not be simple at all; one should at least know the underlying assumptions as well as how to interpret the output. Here, we systematically organize existing programs into methodological categories and provide practical hints for the aforementioned challenges to simplify the entrance into the field. Notably, this review refrains from providing a detailed technical benchmark as there are several reviews on this topic that can be recommended [14, 15, 16].

Literature overview

The rapid development of next-generation sequencing (NGS) techniques catalyzed an explosion in the number of annotated microRNAs [17]. In turn this promoted the development of target identification methods [18]. In Figure 1 the number of publications describing computational methods for target-prediction from 2003 to 2019, split into articles for novel tools and update notes, is shown. Listing approximately 8 articles per year (Mean≈ 7.6, SD≈ 4.0), *in silico* miRNA-target prediction remains an attractive field within bioinformatics and life sciences. Intriguingly, the number of publications has an apparent peak at and around the year 2013, which we contribute towards the increasing application of cross-linking immunoprecipitation-high-throughput sequencing (CLIP-seq) techniques that allow to map microRNA interaction sites in a genome-wide manner [19, 20]. Even though the publishing activity started to decline from 2017 onwards, recent advancements in graphics processing unit (GPU) computing, in particular its applications such as deep learning in genomics lead to an upsurge in publications, which is expected to further increase within the next years [21]. Lastly, the update cycle is fairly frequent for some tools e.g. TargetScan, receiving updates more than 10 years after their initial publication (Supplementary Figure 1) [22].

Principles of target prediction

A catalogue comprising 32 features and meta-characteristics about 98 tools was compiled from the literature and analyzed (Supplementary Table 1). To provide a simple measure to categorize and classify tools according to their underlying methodology, five key-principles used as features for miRNA-target identification by any tool under consideration were identified. Those are sequence or seed complementary (SC), structural and energetic properties (ST), site accessibility (SA), species conservation (C), and expression analysis (E). While the former four are typically derived from sequence information only, the latter one can be performed in combination with Microarrays or RT-qPCR experiments. In Figure 2 the four sequence feature categories are illustrated. Beginning with the sites-level descriptors a fundamental feature of target prediction programs consists of searching for any sequential match between a short query, i.e. a microRNA, and a large target sequence, i.e. $3'$ UTR, to find a seed match. Optionally, several tools scan for additional $3'$ binding of nucleotides at positions 13 to 16 of the microRNA. Sequence matches can also be characterized from a thermodynamics point of view; substantial base pairing provides a negative binding free energy to favor a duplex structure between microRNA and target site. miRanda was the first tool that successfully combined these two notions as it seeks for site matches and filters any hits according to their computed minimum free energy (MFE), where lower values of MFE indicate more stable bindings as measured in kcal/mol [23]. Subsequent publications mostly adopted this mechanism or were solely dependant on structural properties without requiring a seed match at all, e.g. rna22 and RNAhybrid [24, 25, 26]. Later it was hypothesized that the structure of the target site before getting occupied by the RNA-induced silencing complex (RISC) plays an informative role as well leading to the discovery of the site accessibility features, which were first indirectly and directly measured in MicroTar and PITA, respectively [27, 28]. In parallel, tools were developed that take conservation levels of microRNA and target into account, presumably to reduce the number of weak site mappings and with TargetScan and DIANA-microT leading these findings [29, 30]. Over the years more features have been described such as the flanking AU nucleotide content as replacement for site accessibility, synergistic regulatory effects by closely located binding sites, or preferential locations of sites within $3'$ UTRs that were predominantly combined with machine learning [31, 22, 32]. Tools incorporating the last of the five key-principles namely the expression analysis pursue an idea that is orthogonal to the aforementioned ones. Using any experiment that is capable of reporting the

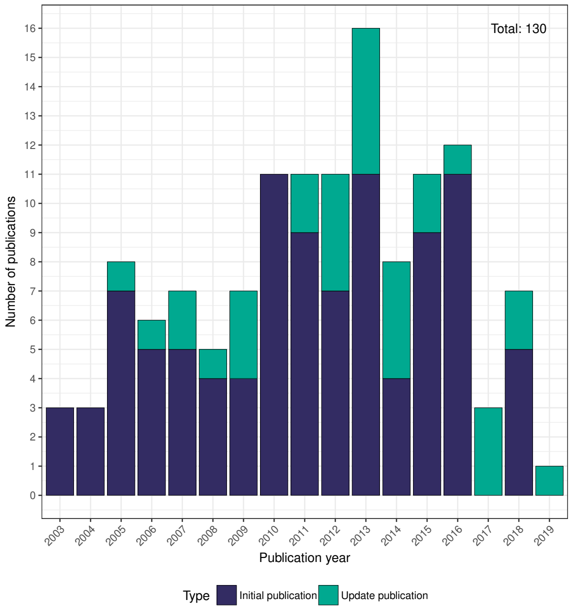


Fig. 1: Number of publications per year describing mammalian microRNA-target prediction methods, split by the number of original in dark blue color and updates including web server applications in green color, respectively. In total 130 publications have been collected during an initial literature study and were subsequently organized into a tool catalogue (Supplementary table 1).

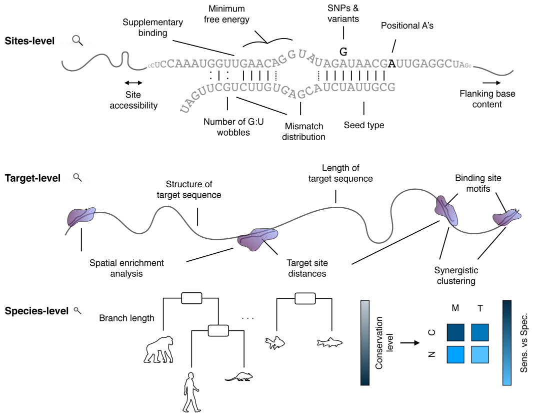


Fig. 2: Schematic overview of the key-principles and features of microRNA-gen targeting as implemented by methods collected for this review. On top, the sites-level features are depicted with the most prominent features being the seed-type match, the minimum free energy of the microRNA-target duplex Δ_{duplex}, the flanking AU content, and the site accessbility $\Delta\Delta G = \Delta G_{duplex} - \Delta G_{open}$. On the whole-target level, i.e. $3'$ UTR or mRNA, more complex features such as the binding site multiplicity, the location bias of target sites and the secondary structure folding of the target sequence constitute important descriptors. Finally, sequence conservation analysis of microRNAs and / or targets between closely located species, for example measured through the branch-length, is commonly used to restrict sensitive methods towards highly conserved patterns. Hence, conservation can be used to balance specificity against sensitivity, whereby forcing a conservation of microRNAs tends to be more effective than it is for the targets.

expression levels of both miRNA and mRNA simultaneously, one searches for anti-correlated expression patterns. Pearson and Spearman correlation or mutual information are popular measures to assess these patterns, for example as implemented by MMIA, CoMeTa, and Cupid [33, 34, 35].

Systematic categorization according to model features

We labelled each tool with the corresponding key-principle or multiple of it to identify $N = 17$ groups, i.e. unique combinations. The methodological overlaps are shown as Venn diagram in Figure 3. As expected the largest group ($N = 17$) is defined by tools implementing the full palette of sequence-based features, because one would expect the full-feature models to have higher predictive power. Taking only the sequence match and structure features into account the second largest group is formed, showing that these two criteria seem to go well together. Intuitively this makes sense since one can obtain a paired RNA sequence alignment by minimizing the duplex free energy over two nucleotide strings. Still, 13 tools also take the site accessibility into account, which is a simple and natural extension of the minimum free energy based models. Another noticeable overlap concerns the category about expression analysis as it primarily appears together with several or all of the sequence-based features, suggesting that expression data can effectively be used to complement or refine classical sequence-based prediction models. Moreover, 49 programs do not include information on conservation, supporting earlier findings that conservation is not an appropriate measure for functionality of target sites but rather just acts as an efficient false-positive site filter for very sensitive methods [36].

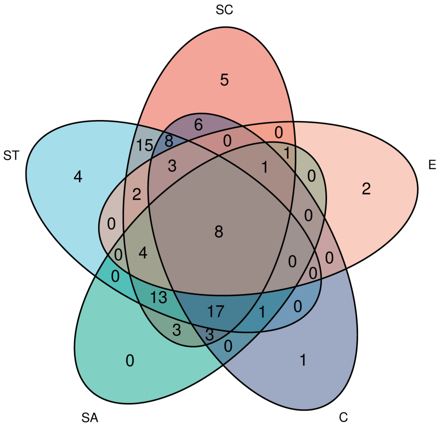


Fig. 3: Venn-diagram illustrating the categorization of tools from the catalogue with respect to their methodological key-principles as described in the main text. Categories and abbreviations are seed or sequence match (SC), structural properties (ST), site accessibility (SA), conservation filter (C), and expression analysis (E). Each uniquely colored region highlights a distinct overlap of the five key-principles along with the number of tools that implement the corresponding features.

Interpreting tool outputs

Correctly understanding assumptions, parameters and output of computational methods promotes effective research as recent work by Shah *et al* on misunderstood parameters of BLAST, a popular sequence alignment tool, showed [37]. In case of microRNA-target prediction the most canonical workflow comprises two FASTA files provided by the user that are turned into a list of paired predictions along with one or several kind of output scores [38]. Depending on the tool each pair on the list can either be a microRNA and a binding site within some target sequence, a microRNA and an entire target mRNA, or both. While the former provides a more fine-grained, molecular description of how the miRNA is predicted to interact with its target, the latter is easier to understand and even mandatory to perform functional enrichment analysis. More importantly, not all available models are capable of generating predictions that are ready to use for downstream applications and thus require manual postprocessing. Therefore, we sought to provide a more generalized description of the most common types of output from the models in our catalogue and sketch possible interpretations in Table 1.

Starting with the most simple type of result a binary class label may be reported, which can be useful to generate a simple separation of prospective candidates from unlikely ones but on the other hand might be too generous. To generate a ranking of pairs residing in the same output class one can consult the mapped seed types or the number of extracted binding sites to reassess the predictions. Next, the computed MFE is a popular measure to rank different binding sites among each other. However, low values may occur by chance, a likelihood that increases with the length of the target sequence if not corrected properly [49]. As argued earlier, relying on conservation scores to rank predictions may effectively be used to filter false-positives and boost specificity [31]. Nevertheless, conservation does not necessarily imply functionality nor does it the other way around [50, 51]. In particular, if one seeks to perform *de novo* predictions we recommend to go for non-conservation models. Context scores provide an independent way of judging the binding quality of distinct target sites as they capture the direct and surrounding nucleotide content and thus

Table 1. Common types of output generated by microRNA-target prediction tools. In the second column the numeric range(s) for each type are defined. In the third column the possibility of using the output to rank predictions among each other, in particular with respect to functionality are listed. Each type is assigned a granularity level (fourth column), depending on its usage to characterize either binding-sites, entire target sequences, or both. In the literature, the term binding site is also commonly referred to as microRNA recognition element (MRE), marking stretches of target RNA that exhibit partial Watson-Crick pairing to a miRNA. The last two columns state a possible interpretation for each type of output along with an associated example method.

Output type	**Range of values**	**Suitable for ranking?**	**Granularity**	**Interpretation**	**Example(s)**
Binary classes	Discrete (e.g. $b = \{0, 1\}$)	N	Target, Binding site	0=Interaction unlikely, 1=Interaction likely *	homoTarget [39]
Minimum free energy	$\mathbb{R}$	P	Binding site	The lower the more stable the duplex indicating an efficient binding. Standard thresholds are defined between -17 and -12 kcal/mol [40].	IntaRNA 2.0 [41]
Conservation score	$\mathbb{R}$	P	Target, Binding site	Higher scores indicate greater conservation across a larger number or very closely related species.	phastCons score [42], Probability of conserved targeting (P_{CT}) [43]
Context score	$\mathbb{R}$	P	Binding site	Indicates favorable nucleotide compositions and position of a site.	TargetScan (Context++ score) [22]
Fold-change	$\mathbb{R}$	Y	Target	Expected fold-change of miRNA x on target y assuming that both are expressed above background.	miRepress [44]
Correlation / Mutual information	$[-1, 1], \mathbb{R}_{>0}$	Y	Target	Anti-correlations suggest regulatory dependencies, e.g. microRNA is upregulated and respective targets are downregulated	CoMeTa [34]
P-value / Z-score / Signal-to-noise ratio	$[0, 1]$, $[-3\text{SD}, 3\text{SD}]$	P	Target, Binding site	Extreme values or significant p-values mark divergence from a norm, and in turn increased regulatory potential.	RNAhybrid [26]
Probability	$[0, 1]$	P	Target, Binding site	Likelihood of miRNA x and target y to interact.	mirMark [45]
Seed type / Match length	$\mathbb{R}_{>0}$	P	Binding site	Rate sites based on the accepted seed type hierarchy: 8mer >7mer-m8 >7mer-A1 >6mer >Offset 6mer	GUUGle [46]
Custom score	$\mathbb{R}$	NA	Target, Binding site	Prioritize targets according to several, convoluted criteria. Note that respective score ranges might be specifically distributed or normalized for one microRNA or target sequence, or across the entire input dataset [47].**	DIANA-microT-CDS: Linear combination of summed MRE scores from CDS- and 3′ UTR-derived sites [48].

SD = Standard deviation, N = No, P = Partial, Y = Yes, NA = Not available, * Neglecting a possible switch of classes, ** Always conduct the corresponding manuscript.

other hidden features such as the site accessibility. The most popular implementation is the context++ model as provided by TargetScan, which ranks predictions by combining sequence descriptors with estimated seed type contributions that can correlate with experimentally measured target repression [22]. Predicted fold-changes suggest a very convenient way of assessing efficacy of regulatory interactions, however only 5 out of 98 tools report such. We reason that fold-changes may be highly microRNA specific making it difficult to provide models that generalize well for previously unseen microRNA sequences. Furthermore, fold-changes might be overly optimistic since side-effects such as sponge behavior or cross-targeting events have to be taken into account as well [52]. Correlation and mutual information are classical scoring functions to assess microRNA and mRNA co-expression data. While significant anti-correlations in their expression patterns are indicative of a regulatory dependency, combining expression analysis models with sequence-based predictors could be necessary to resolve ambiguous multi-correlations and cooperative targetings of several microRNAs targeting the same mRNA. In particular early approaches were designed to report p-values or signal-to-noise ratios based on the estimation of a background distribution, e.g. the number of seed matchings for randomly shuffled microRNA sequences, to infer enrichments above this noise level. Not only are these methods sensitive towards outliers but must be applied with caution since a recalibration with newer datasets might be required in order to get a better estimate of the statistical parameters. In addition, growing sample sizes can cause significant p-values to occur by chance, an effect that can be counteracted by correcting the p-values or considering statistical effect sizes instead. On the other hand probabilities provide an intuitive way of ranking and classifying predictions. Most of the selected tools directly report the probability of a pre-trained machine learning model, e.g. a support-vector machine, which can be used in combination with some reasonable decision threshold to distribute the pairs into two or more regulatory classes. Conceptually, probabilities located close to the decision threshold are understood to be less confidently assigned than those being extremely distributed, i.e. close to 0 or 1.

As a special case for per-binding site predictors, considering the seed type matches of one microRNA along a target can be used to judge the regulatory potential. For example, targets harboring only one or two minor site types such as 6-mers tend to be less effective and thus have a higher chance of being false-positives [47]. Consequently, binding site multiplicity in combination with seed type distributions can in principle be used to point out target candidates. Lastly, tools that compute specifically designed scores for a weighted combination of multiple features constitute a special category for which we recommend to check the individual definition carefully. Understanding the hypothetical range of values a score can take on and verifying whether simple confounding factors exist that influence the distribution makes up an acceptable starting point. Note that in Table 1 partial rankings refer to types typically accompanied by some pre- or user-defined threshold, e.g. keeping binding-sites with an MFE of -12 kcal/mol or less, and discarding otherwise. This implies that all pairs passing a reasonable threshold should be treated as being significant. In contrast, ranking-compatible scores such as the fold-change are not dependent on an associated cut-off, so that one can directly traverse predictions in a prioritized manner.

Resolution of common usage pitfalls

Besides methodological considerations about individual tools, we asked whether common usage issues exist that may arise while working with computational models for target prediction. Our respective findings that describe both biologically and technically motivated challenges are summarized in Table 2. First and foremost, a single gene locus can give rise to different transcripts and taken together with alternative polyadenylation mechanisms, $3'$ UTR shortening or lengthening can occur *in vivo*, effectively altering the landscape of regulatory binding sites [53, 54]. Most target prediction resources solve this problem by selecting either any or the longest $3'$ UTR from available databases such as Ensembl or NCBI's Refseq. However, this may be too simplistic as a counter example shows. The human gene MDM2 exhibits 19 protein-coding transcripts (GRCh38) with partly non-overlapping $3'$ UTRs (Supplementary Figure 2). Moreover, due to their nature of being non-coding, UTR sequences tend to exhibit a lower nucleotide complexity as compared to sequences in the open reading frame, i.e. increasing risk of spurious binding site matchings because of repetitive sequences. As a case example, we analyzed the $3'$ UTR of the human gene ARHGEF10L using the latest version of RepeatMasker resulting in 104 bases (21.27%) to be masked due to simple repeats (Supplementary Notes 1, Supplementary Figure 3) [55]. We then applied rna22 v2 on the unmasked sequence and searched for targets of all human microRNAs from miRBase v22, resulting in 924 distinct predicted binding sites, 269 of which ($\approx$ 29%) are located within the extracted repeats [2]. Recent findings also support the hypothesis that functional microRNA binding sites reside within coding sequences and $5'$ UTRs as well, motivating the question whether existing models can be applied seamlessly to the full-length mRNA [56]. Interestingly, binding principles and sequence features described earlier are only partly transferable and specific models for these novel target regions are in favor of classical $3'$ UTR algorithms [57]. Due to the inherent complexity of microRNA gene targeting and the fact that no tool consistently outperforms all others, intersecting or taking the union of the output from several tools established to be common practice among end-users. Indeed, earlier studies report that union approaches provide a good trade-off between specificity and sensitivity [58]. Albeit this approach bears some caveats; first one must make sure to use the same reference databases, i.e. sources for microRNA and gene annotation to prevent mapping issues, second there may be instances where prediction programs disagree with each other, and finally the amount of methodological overlap between tools under consideration confounds the bias-variance trade-off [59].

Turning to the experimental side, predicted interactions are often validated by either Microarrays, qRT-PCR, Western Blots, Luciferase assays, pSILAC, or NGS-based methods including CLIP-seq and CLASH [69]. While each method has its own strengths and weaknesses the principal difference between transcriptomics-based and proteomics-based evidence is crucial. While methods that rely on the presence of the mRNA of interest like Microarrays and qRT-PCR can only detect cleaved but not repressed targets, methods that capture the assembled protein like Luciferase assays or Western Blots fail to distinguish cleavage from repression events. Even though the latter techniques usually come at a higher cost than the former, they are guaranteed not to miss any regulatory effect, contrasting the accuracy of their transcriptomic counterparts. Also, popular databases like miRTarBase and TarBase report not only positively confirmed interactions but also negative ones, whereby negative results should be understood as having no evidence for a regulation instead of showing the absence of any regulatory interaction at all [74, 75, 16]. Finally, a popular downstream application concerns the enrichment of target genes within functional categories or biological pathways as provided by Gene Ontology or Kyoto Encyclopedia of Genes and Genomes (KEGG) among the predicted targetome of microRNAs [76, 77, 78, 79, 80]. As pointed out earlier, enrichment analyses should be performed using highly stringent p-value cut-offs to avoid likely false-positive associations [59].

Challenges associated with prediction benchmarks

Since it might be desirable to perform customized prediction benchmarks for a set of candidate tools that suit a given research setup, several probable issues are required to be overcome for a reliable benchmark. Reconsidering the key-principles introduced in Figure 3 it is apparent that not all of these categories are directly comparable with each other. Therefore, we propose that a respective benchmark should at least distinguish between classical sequence-based, expression analysis, and NGS-based methods, the latter of which include mapping reads from CLIP-seq or CLASH experiments, to ensure a certain base-level of commensurability. As shown in Supplementary Table 1, published tools were either in a minority of cases not validated, tested with simulated data, or in most cases verified using experimental data sets, adding yet another level of complexity into benchmarks. To this end, it is crucial to know which interactions have been used for evaluating the implementations, especially in case of machine-learning driven applications, otherwise one is at risk to test a model with a sub-set from the training data. Additionally, validated interactions as publicly available through miRTarBase and TarBase are categorized into distinct levels of confidence, depending whether a low-throughput method, e.g. Western Blot or a high-throughput method, e.g. Microarray was used. In the respective databases the number of weak and strong interactions differs more than ten-fold raising a class-imbalance problem that should be taken into account during the design of test sets as well. To increase the complexity even more, benchmarks can get biased towards already established methods because a selection of prospective candidates subject to wet-lab validation was likely accomplished with one or several prediction methods beforehand. Further, interactions labelled with a negative outcome require special care. A non-positive outcome can be interpreted as missing evidence of a regulatory dependency, however in some manuscripts, negative interactions are simulated by re-shuffling the features of positively tested pairs. Finally, not all approaches were developed to function *ab initio* where the input from one or several tools might be required to apply tools falling in this particular category in a correct fashion. It is conceivable that depending on the parameters of the tool(s) used upstream, the outcome of any down-stream application can be altered considerably.

Method selection guides

To simplify the task of selecting appropriate and use-case specific tools from the larger set of available solutions as listed in Supplementary Table 1 we devised an ordered set of questions to guide the selection procedure:

1. For which organism(s) should the targets be predicted?

Table 2. Common questions, problems, and pitfalls researchers may face while working with in silico microRNA-target prediction methods. For each case example, the pitfall is described in the second column, followed by an example or source for each pitfall. Further, at least two solutions are given in the remaining columns of the table.

Problem / Question / Usecase	Pitfall(s)	Example / Citation / Source	Possible solution	Alternative solution
Multiple transcripts, 3’ UTRs per gene	Often only the longest 3’ UTR isoform is selected	MDM2 gene showing 19 3′ UTR isoforms.*	Ensembl: Take Havana-Ensembl overlap, TSL: 1-2, NCBI: Choose the principal isoform	Take overlap between refseq and Ensembl
Low-complexity target sequences	Spurious binding site mappings	Repeats in 3′ UTR of gene ARHGEF10L.*	Use pre-masked genome	Apply repeat masker on custom sequences
Predict targets for 5’ UTR or CDS	Using a tool developed for 3’ UTR prediction on CDS sequences	MinoTar [60], See also review by Chipman and Pasquinelli [61]	Use a tool specifically crafted for CDS	Search for the supplementary bindings
Intersect default predictions from different tools	Non-matching reference databases, e.g. Ensembl, NCBI	See study by Ritchie et al [59]	Convert identfiers / genome coordinates using utility tools, e.g. UCSC’s LiftOver [62]	Re-run the tools on updated datasets and then merge
Compute overlap of tool predictions	Bias-Variance trade-off, Sensitivity vs Specificity trade-off	See study by Oliveira et al [58]	Search for tools using distinct features on distinct datasets	Stick with one tool and search for other criteria to re-rank the predictions
Applying sequence conservation filter	Conservation != Functionality, Conservation in CDS is implied	hsa-miR-15a targets from mirSVR study [31]	If possible run with and without conservation filter	Run other tool that makes no assumptions about conservation
Getting large number of predictions	Detecting targets with very low specificity	Default prediction counts vary up to 4 orders of magnitude [16]	Use conservation filter or expression data to filter false positives	Re-rank or filter predictions using some third-party ranking methods, e.g. MirAncesTar or SeedVicious [63, 64].
Comparing tool output with measured fold-change	Output score is not a measure for fold-repression	mirWIP: aggregates context-specific scorings unrelated to fold-change [65]. miRNALasso: explicitly quantifies down-regulation from expression data [66].	Define two discrete regulatory classes and iterate all possible thresholds to compute a ROC-AUC	Use dedicated ranking methods developed for this purpose, e.g. myMIR [67]
Predicting targets for reference sequences only	Missing the sequence variation, e.g. isomiRs, SNPs / SNVs in UTRs	Variants in COPD application: SubmiRine [68]	Rerun selected tools on input sequence with variants applied	Use specific tools for this purpose, e.g. SubmiRine
Working with "negative" interactions from experimental databases	"Negative" means absence of evidence not evidence of absence, Class imbalance	mirDIP prediction database [16]	Correct for class imbalance using standard methods, e.g. down- and upsampling, imputing	Perform binding site knockout experiments, e.g. Luciferase assays to label each binding site
Validating predictions with Luciferase or Immunoblotting	Cleaving and translational repression of ribosome cannot be distinguished	hsa-miR-15a-5p regulates TP53 (MIRT005763), See review [69]	Combine with transcriptomics-based experiments (see below)	Confirm with tool that predicts mRNA change, for example targetScore [70]
Validating predictions with Microarray and qRT-PCR	Contrary to previous one; Can detect cleaved but not repressed targets	hsa-miR-197-3p regulates TSPAN3 (MIRT000215), See review [69]	Combine with protein-based experiments (see above)	Confirm with tool that predicts protein fold change, for example miSTAR or DIANA-microT-ANN [71, 72]
Extracting significant targets / enriched pathways using p-values	Hitting a significance by chance	See study by Ritchie et al [59]	Correct for the p-values, e.g. using Benjamini-Hochberg procedure [73]	Take only highly significant p-values

TSL = Transcript Support Level, * Genome browser screenshot can be found in Supplementary Data,
* Source: http://mirtarbase.mbc.nctu.edu.tw/php/detail.php?mirtid=MIRT005763,*** Source: http://mirtarbase.mbc.nctu.edu.tw/php/detail.php?mirtid=MIRT000215

2. Which target region should be scanned, e.g. 3′ UTR, coding sequences or entire mRNAs?
3. What type of input is available, i.e. sequences, expression levels or both?
4. Should predictions comprise a list of target sites, a list of target transcripts or both?
5. Are NGS reads (e.g. from CLIP-seq experiments) available or not?
6. Should *ab initio* methods be preferred over ensemble / deductive methods?

If the remaining set of tools remains to be unspecific one can consult additional attributes to prioritize the candidates. For example, one might prefer full sequence-feature (SC, ST, SA, C) methods over simpler models. An alternative requirement might be whether targets should only be computed using known or novel microRNAs or even custom target sequences as well. To enhance the flexibility of the provided data set with respect to a broader range of applications we computed a hierarchical clustering of all collected methods using 16 categorical variables depicted as a tree in Figure 4. In its essence the tree combines the information of $1,568$ informative cells to group similar tools together, i.e. a lower vertical distance in the tree reflects a higher similarity between any two given methods. We motivate this procedure by two case examples;

1. Suppose the tool miRanda was selected using the above outlined questionnaire and one wants to check similar tools to intersect the output with. The tree reveals that the lowest branch occurs between miRanda and CUDA-miRanda, a GPU based high-throughput re-implementation that allows to analyze extremely large sets of miRNAs and mRNAs [81]. Moreover, MicroInspector, which is located at the next upper branch, not only implements the same key-principles as miRanda but provides a secondary-structure filter for miRNA-target site hairpins exhibiting self-folding artifacts that cannot be detected solely by MFE-based scorings (Supplementary Figure 4)[82].
2. Same assumption as in (1) but a combination of the output with a tool having a distinct methodology is desired. For example, targetScore makes up a good candidate as it is located in the tree several branches away from miRanda. A closer look into the catalogue affirms the selection, because targetScore not only takes distinct sequence features into account but also makes use of expression data [70].

We have implemented an interactive webpage that supports user requested queries based on any criteria available in our literature catalogue to perform tool selection. The implementation is freely available at https://www.ccb.uni-saarland.de/mtguide.

Future perspectives

MicroRNA target prediction tools constitute a crucial step in sncRNA analyses seeking to understand the transition from geno- to phenotype. Given the yet limited availability of experimentally verified interactions, computational methods remain the method of choice for researchers to make a pre-selection of likely functional targets. However, as shown by a recent study of Fridrich *et al.*, prediction tools tend to exhibit a substantial level of noise, especially in non-model organisms or previously unexplored lineages, residing between 65% and 85% [84]. Even though experimentally validated tools for non-model organisms remain to be explored, several current advancements in genomics motivate the need for novel methods of model organisms as well. First, the development of single-cell sequencing techniques, in particular co-sequencing of microRNA- and mRNA molecules from the same cell offer new possibilities in understanding microRNA mediated gene regulation by modelling sequence-family specific regulatory networks [85]. For example, such a co-sequencing offers a great potential in understanding the heterogeneity of cancer tissues and differentially regulated RISC targetomes. Moreover, augmenting existing data sets with time-series measurements could provide an entirely new dimension of how we view development-specific regulatory interactions. To this end, recent findings suggest that the dominantly expressed mature arm of a microRNA can alternate either from the $5'$ dominant to the $3'$ dominant form or the other way around, a mechanism also known as microRNA arm switching [86]. Furthermore, flexible models taking into account the natural sequence variation typically in form of isomiRs, and SNPs in respective target mRNAs are superior to static reference models, especially in clinical contexts [68]. Here, deep learning delineates itself as an excellent candidate to drive the next-generation of target prediction methods since it scales well with large data sets while offering the potential to unveil previously undetected coherences, mitigating the need for handcrafted features [87, 88].

Methods

Literature study

For the initial literature catalogue we collected relevant publications describing either original work or updates including applications such as web servers in the field of microRNA-gene targeting using NCBI's PubMed. From each publication 32 features described in Supplementary Table 1 were extracted from either the main text or the supplementary documents. Citation counts were collected on 4th of February 2019 using Web of Science (https://apps.webofknowledge.com/).

Figure design

Figure 1 was rendered using the R package *ggplot2* and Figure 3 using *VennDiagram* and *ggsci* [89, 90, 91]. The initial distance matrix underlying Figure 4 was computed using the gower distance implemented in the R package *cluster* and subsequently visualized with *circlize*, *dendextend*, and *viridis* [83, 92, 93, 94]. The column indices for the 16 features selected for clustering the tools are 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 18, 23, 32. Respective columns are also highlighted with italic column names in Supplementary Table 1.

Case example analysis

Sequences for human microRNAs were obtained in FASTA format from miRBase release v22 [2]. Genomic locations for the tested $3'$ UTR sequences of the human genes MDM2 and ARHGEF10L were acquired using Ensembl's BioMart Release 96 [95]. Respective sequences were cut from the human reference genome assembly *GRCh38.p12* as obtained through GENCODE release 30 [96]. Subsequently, RepeatMasker and rna22 configured with default parameters were run on the selected $3'$ UTR of ARHGEF10L followed by manual analysis of the output [55, 25].

Interactive website guide

The webservice was implemented using Django v2.1.7, SQLite3 and bootstrap v4.3.1 for the back- and front-end, respectively. To support interactive and flexible user-queries the content of Supplementary Table 1 was stored across several query-optimized tables in a SQLite database. Acting on top of this dataset an ordered set of questions was implemented, once supporting exactly the recommended procedure as presented in the manuscript and once in an arbitrary fashion for advanced usage. A more natural selection of domains and taxonomic rankings up to single species, as supported by individual tools, was implemented using a phylogenetic mapping from the organisms annotation file of miRBase v22.

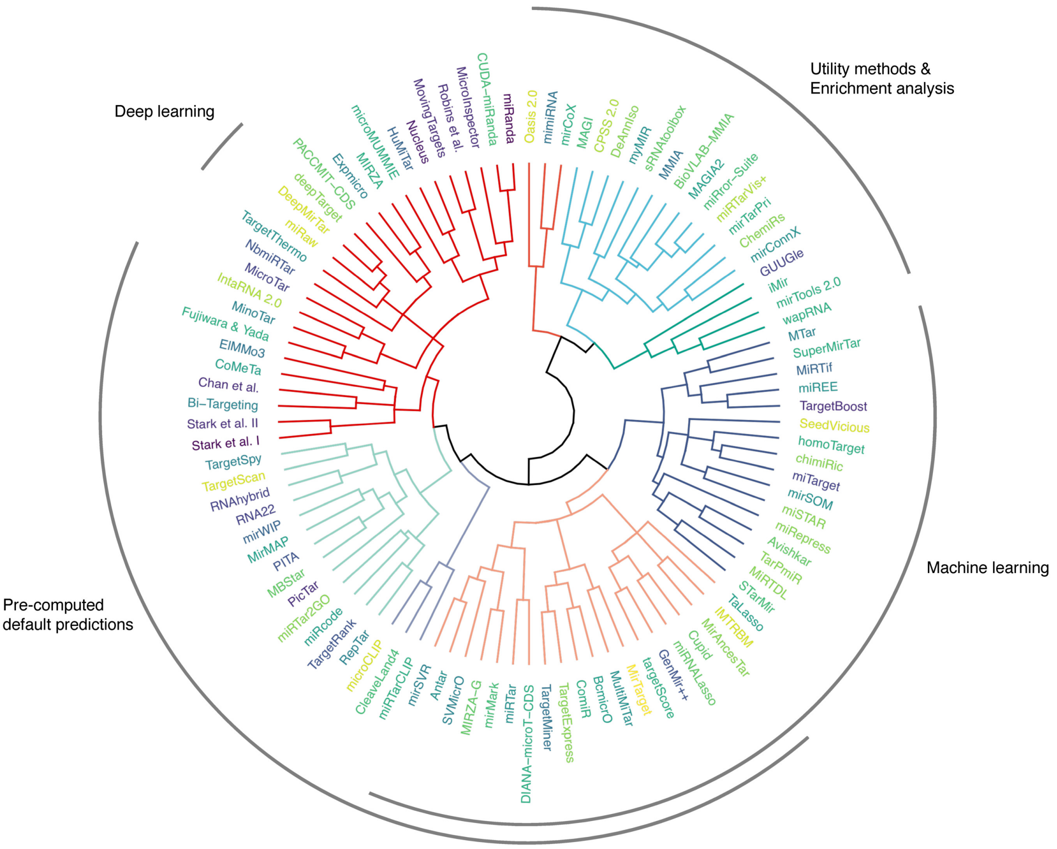


Fig. 4: Hierarchical clustering of 98 target prediction methods based on 16 categorical variables, a subset from Supplementary Table 1, illustrated as polar tree dendrogram. Pairwise dissimilarities were calculated using the gower coefficient as distance metric from the CRAN R-package cluster [83]. Each leaf in the tree is labelled with a corresponding tool name or the first author of the publication if no name is given. Tool labels are colored according to the last year of publication, i.e. original or latest update publication where yellow color indicates very recent and dark blue colors older publications. Partial sub-trees were colored at the cut $k = 8$ to highlight groups of similar methods. Circular annotations around the tree provide additional information about categories being enriched among even larger groups of tools that possibly span multiple sub-trees.

Key points

- *In silico* microRNA target-prediction tools incorporate a broad collection of molecular aspects but are mostly based on five key-features.
- Serving all possible research questions is difficult to accomplish using a standalone implementation. Intersection approaches that combine tools with distinct methodology can provide a possible remedy.
- Knowing the underlying assumptions and input requirements of a given tool is crucial for a successful interpretation of model output.
- The application of microRNA-target prediction tools bears several pitfalls that can influence down-stream analysis considerably.

Funding

This work was supported by internal funds of Saarland University; as well as by the state government of Saarland; and the Michael J. Fox Foundation for Parkinson's Research (US) [14446].

Author descriptions

Fabian Kern is a PhD student at the Chair for Clinical Bioinformatics, Saarland Informatics Campus, Saarbrücken Germany. His research is focused on gene regulation, in particular post-transcriptional mechanisms including microRNAs and their functional roles. **Christina Backes** is a Postdoctoral fellow at the Chair for Clinical Bioinformatics, Saarland Informatics Campus, Saarbrücken Germany. **Pascal Hirsch** is an assistant researcher at the Chair for Clinical Bioinformatics and pursues a Master's degree in Bioinformatics at Saarland University, Saarbrücken Germany. **Tobias Fehlmann** is a PhD student at the Chair for Clinical Bioinformatics, Saarland Informatics Campus, Saarbrücken Germany, studying small non-coding RNAs with next-generation-sequencing methods. **Martin Hart** is a Postdoctoral fellow at the Medical school and the Department for Human Genetics at Saarland University Hospital, Homburg Germany.

His research focuses on experimental validation of microRNA targets in several cell-types that are associated to a disease phenotype. **Eckart Meese** is a full-professor for human genetics and is the leader of the Institute for Human Genetics at the Medical School of Saarland University Hospital, Homburg Germany. **Andreas Keller** is a full-professor at the Medical School of Saarland University Hospital, Homburg Germany, leads the Chair for Clinical Bioinformatics, and an associate professor of the Saarland Informatics Campus, Saabrücken Germany.

Author contributions statement

F.K. conducted the initial literature review and performed data analysis, **C.B.** and **T.F.** advised the implementation of data science procedures and the design of figures and tables, **P.H.** implemented the associated webpage, **M.H.** helped to define usage pitfalls, profiled the guidance statements, and provided support for experimental techniques mentioned in the main text. **F.K.**, **E.M**, and **A.K.** wrote the manuscript. All authors reviewed and agreed upon the manuscript.

Competing interests

The authors have declared no competing interests.

References

[1]Lee RC, Ambros V. An extensive class of small RNAs in Caenorhabditis elegans. *Science* 2001;**294**(5543):862–864.

[2]Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Research* 2018; **47**(D1):D155–D162.

[3]Fehlmann T, Backes C, Pirritano M, *et al.* The sncRNA Zoo: a repository for circulating small noncoding RNAs in animals. *Nucleic acids research* 2019;**47**(9):4431–4441.

[4]Fromm B, Billipp T, Peck LE, *et al.* A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. *Annual Review of Genetics* 2015; **49**(1):213–242.

[5]Han J, Lee Y, Yeom KH, *et al.* The Drosha-DGCR8 complex in primary microRNA processing. *Genes & development* 2004; **18**(24):3016–3027.

[6]Zinovyeva AY, Veksler-Lublinsky I, Vashisht AA, *et al.* Caenorhabditis elegans ALG-1 antimorphic mutations uncover functions for Argonaute in microRNA guide strand selection and passenger strand disposal. *Proceedings of the National Academy of Sciences* 2015;**112**(38):E5271 LP – E5280.

[7]Bartel DP. Metazoan MicroRNAs. *Cell* 2018;**173**(1):20–51.

[8]Brodersen P, Voinnet O. Revisiting the principles of microRNA target recognition and mode of action. *Nature Reviews Molecular Cell Biology* 2009;**10**(2):141–148.

[9]Reinhart BJ, Weinstein EG, Rhoades MW, *et al.* MicroRNAs in plants. *Genes & development* 2002;**16**(13):1616–1626.

[10]Wang J, Mei J, Ren G. Plant microRNAs: Biogenesis, Homeostasis, and Degradation. *Frontiers in Plant Science* 2019;**10**:360.

[11]Kidner CA, Martienssen RA. Macro effects of microRNAs in plants. *Trends in Genetics* 2003;**19**(1):13–16.

[12]Voinnet O. Origin, Biogenesis, and Activity of Plant MicroRNAs. *Cell* 2009;**136**(4):669–687.

[13]Brennecke J, Stark A, Russell RB, *et al.* Principles of microRNA-target recognition. *PLoS Biology* 2005;**3**(3):0404–0418.

[14]Fan X, Kurgan L. Comprehensive overview and assessment of computational prediction of microRNA targets in animals. *Briefings in Bioinformatics* 2014;.

[15]Li Y, Jin X, Wang Z, *et al.* Systematic review of computational methods for identifying miRNA-mediated RNA-RNA crosstalk. *Briefings in Bioinformatics* 2017;.

[16]Tokar T, Pastrello C, Rossos AE, *et al.* MirDIP 4.1 - Integrative database of human microRNA target predictions. *Nucleic Acids Research* 2018;.

[17]Alles J, Fehlmann T, Fischer U, *et al.* An estimate of the total number of true human miRNAs. *Nucleic acids research* 2019;**47**(7):3353–3364.

[18]Reyes-Herrera PH, Ficarra E. One Decade of Development and Evolution of MicroRNA Target Prediction Algorithms. *Genomics, Proteomics & Bioinformatics* 2012;**10**(5):254–263.

[19]Cook KB, Hughes TR, Morris QD. High-throughput characterization of protein-RNA interactions. *Briefings in Functional Genomics* 2015; **14**(1):74–89.

[20]Spengler RM, Zhang X, Cheng C, *et al.* Elucidation of transcriptome-wide microRNA binding sites in human cardiac tissues by Ago2 HITS-CLIP. *Nucleic Acids Research* 2016;**44**(15):7120–7131.

[21]Eraslan G, Avsec Ž, Gagneur J, *et al.* Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics* 2019;**20**(7):389–403.

[22]Agarwal V, Bell GW, Nam JW, *et al.* Predicting effective microRNA target sites in mammalian mRNAs. *eLife* 2015;**4**(AUGUST2015).

[23]Enright AJ, John B, Gaul U, *et al.* MicroRNA targets in Drosophila. *Genome biology* 2003;**5**(1):R1.

[24]Krek A, Grün D, Poy MN, *et al.* Combinatorial microRNA target predictions. *Nature Genetics* 2005;**37**(5):495–500.

[25]Miranda KC, Huynh T, Tay Y, *et al.* A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes. *Cell* 2006;**126**(6):1203–1217.

[26]Krueger J, Rehmsmeier M. RNAhybrid: microRNA target prediction easy, fast and flexible. *NUCLEIC ACIDS RESEARCH* 2006; **34**(SI):W451–W454.

[27]Thadani R, Tammi MT. MicroTar: predicting microRNA targets from RNA duplexes. *BMC BIOINFORMATICS* 2006;**7**(5):20.

[28]Kertesz M, Iovino N, Unnerstall U, *et al.* The role of site accessibility in microRNA target recognition. *Nature Genetics* 2007;**39**(10):1278–1284.

[29]Lewis BP, Shih IH, Jones-Rhoades MW, *et al.* Prediction of Mammalian MicroRNA Targets. *Cell* 2003;**115**(7):787–798.

[30]Kiriakidou M, Nelson PT, Kouranov A, *et al.* A combined computational-experimental approach predicts human microRNA targets. *GENES & DEVELOPMENT* 2004;**18**(10):1165–1178.

[31]Betel D, Koppal A, Agius P, *et al.* Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biology* 2010;**11**(8):R90.

[32]Ding J, Li X, Hu H. TarPmiR: A new approach for microRNA target site prediction. *Bioinformatics* 2016;.

[33]Nam S, Li M, Choi K, *et al.* MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *NUCLEIC ACIDS RESEARCH* 2009; **37**(S):W356–W362.

[34]Gennarino VA, D'Angelo G, Dharmalingam G, *et al.* Identification of microRNA-regulated gene networks by expression analysis of target genes. *GENOME RESEARCH* 2012;**22**(6):1163–1172.

[35]Chiu HS, Llobet-Navas D, Yang X, *et al.* Cupid: simultaneous reconstruction of microRNA-target and ceRNA networks. *GENOME RESEARCH* 2015;**25**(2):257–267.

[36]Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA

targets. *Cell* 2005;**120**(1):15–20.

[37]Shah N, Nute MG, Warnow T, *et al.* Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows. *Bioinformatics* 2019;**35**(9):1613–1614.

[38]Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America* 1988;**85**(8):2444–2448.

[39]Ahmadi H, Ahmadi A, Azimzadeh-Jamalkandi S, *et al.* HomoTarget: A new algorithm for prediction of microRNA targets in Homo sapiens. *GENOMICS* 2013;**101**(2):94–100.

[40]Hsu JBK, Chiu CM, Hsu SD, *et al.* miRTar: an integrated system for identifying miRNA-target interactions in human. *BMC BIOINFORMATICS* 2011;**12**.

[41]Mann M, Wright PR, Backofen R. IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. *NUCLEIC ACIDS RESEARCH* 2017;**45**(W1):W435–W439.

[42]Siepel A, Bejerano G, Pedersen JS, *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* 2005;**15**(8):1034–1050.

[43]Friedman RC, Farh KKH, Burge CB, *et al.* Most mammalian mRNAs are conserved targets of microRNAs. *GENOME RESEARCH* 2009; **19**(1):92–105.

[44]Ghosal S, Saha S, Das S, *et al.* miRepress: modelling gene expression regulation by microRNA with non-conventional binding sites. *SCIENTIFIC REPORTS* 2016;**6**:22334.

[45]Menor M, Ching T, Zhu X, *et al.* mirMark: a site-level and UTR-level classifier for miRNA target prediction. *Genome biology* 2014; **15**(10):500.

[46]Gerlach W, Giegerich R. GUUGle: a utility for fast exact matching under RNA complementary rules including G-U base pairing. *BIOINFORMATICS* 2006;**22**(6):762–764.

[47]Nielsen CB, Shomron N, Sandberg R, *et al.* Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA* 2007; **13**(11):1894–1910.

[48]Reczko M, Maragkakis M, Alexiou P, *et al.* Functional microRNA targets in protein coding sequences. *Bioinformatics* 2012;**28**(6):771–776.

[49]Rehmsmeier M, Steffen P, Hochsmann M, *et al.* Fast and effective prediction of microRNA/target duplexes. *RNA (New York, NY)* 2004; **10**(10):1507–17.

[50]Peterson SM, Thompson JA, Ufkin ML, *et al.* Common features of microRNA target prediction tools. *Frontiers in Genetics* 2014; **5**(FEB):23.

[51]Pinzón N, Li B, Martinez L, *et al.* microRNA target prediction programs predict many false positives. *Genome Research* 2017; **27**(2):234–245.

[52]Ling H, Fabbri M, Calin GA. MicroRNAs and other non-coding RNAs as targets for anticancer drug development. *Nature Reviews Drug Discovery* 2013;**12**(11):847–865.

[53]Mayr C, Bartel DP. Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell* 2009;**138**(4):673–684.

[54]Di Giammartino DC, Nishida K, Manley JL. Mechanisms and Consequences of Alternative Polyadenylation. *Molecular Cell* 2011; **43**(6):853–866.

[55]Smit, AFA, Hubley, R & Green P. RepeatMasker Open-4.0, 2015.

[56]Lee I, Ajay SS, Jong IY, *et al.* New class of microRNA targets containing simultaneous 5âŁ²-UTR and 3âŁ²-UTR interaction sites. *Genome Research* 2009;**19**(7):1175–1183.

[57]Marin RM, Sulc M, Vanicek J. Searching the coding region for microRNA targets. *RNA-A PUBLICATION OF THE RNA SOCIETY* 2013;**19**(4):467–474.

[58]Oliveira AC, Bovolenta LA, Nachtigall PG, *et al.* Combining results from distinct microRNA target prediction tools enhances the performance of analyses. *Frontiers in Genetics* 2017;**8**(MAY):59.

[59]Ritchie W, Flamant S, Rasko JEJ. Predicting microRNA targets and functions: traps for the unwary. *Nature Methods* 2009;**6**:397.

[60]Schnall-Levin M, Zhao Y, Perrimon N, *et al.* Conserved microRNA targeting in Drosophila is as widespread in coding regions as in 3 ‘ UTRs. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA* 2010; **107**(36):15751–15756.

[61]Chipman LB, Pasquinelli AE. miRNA Targeting: Growing beyond the Seed. *Trends in Genetics* 2019;**35**(3):215–222.

[62]Hinrichs AS, Karolchik D, Baertsch R, *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic acids research* 2006; **34**(Database issue):D590–8.

[63]Leclercq M, Diallo AB, Blanchette M. Prediction of human miRNA target genes using computationally reconstructed ancestral mammalian sequences. *NUCLEIC ACIDS RESEARCH* 2017; **45**(2):556–566.

[64]Marco A. SeedVicious: Analysis of microRNA target and near-target sites. *PLoS ONE* 2018;**13**(4):1–9.

[65]Hammell M, Long D, Zhang L, *et al.* mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. *NATURE METHODS* 2008;**5**(9):813–819.

[66]Wang Z, Xu W, Liu Y. Integrating full spectrum of sequence features into predicting functional microRNA-mRNA interactions. *BIOINFORMATICS* 2015;**31**(21):3529–3536.

[67]Corrada D, Viti F, Merelli I, *et al.* myMIR: a genome-wide microRNA targets identification and annotation tool. *BRIEFINGS IN BIOINFORMATICS* 2011;**12**(6, SI):588–600.

[68]Baxevanis AD, Maxwell EK, Spira A, *et al.* SubmiRine: assessing variants in microRNA targets using clinical genomic data sets. *Nucleic Acids Research* 2015;**43**(8):3886–3898.

[69]Thomson DW, Bracken CP, Goodall GJ. Experimental strategies for microRNA target identification. *Nucleic Acids Research* 2011; **39**(16):6845–6853.

[70]Li Y, Goldenberg A, Wong KC, *et al.* A probabilistic approach to explore human miRNA targetome by integrating miRNA-overexpression data and sequence information. *BIOINFORMATICS* 2014;**30**(5):621–628.

[71]Van Peer G, De Paepe A, Stock M, *et al.* MiSTAR: MiRNA target prediction through modeling quantitative and qualitative miRNA binding site information in a stacked model structure. *Nucleic Acids Research* 2017;**45**(7):e51–e51.

[72]Reczko M, Maragkakis M, Alexiou P, *et al.* Accurate microRNA target prediction using detailed binding site accessibility and machine learning on proteomics data. *Frontiers in Genetics* 2012;**2**:103.

[73]Benjamini, Yoav ; Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological 1995.pdf. *Journal of the Royal Statistical Society Series B (Methological)* 1995; **57**(1):289–300.

[74]Chou CH, Shrestha S, Yang CD, *et al.* MiRTarBase update 2018: A resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research* 2018;**46**(D1):D296–D302.

[75]Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, *et al.* DIANA-TarBase v8: A decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic Acids Research* 2018; **46**(D1):D239–D245.

[76]Ashburner M, Ball CA, Blake JA, *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 2000;**25**(1):25–29.

[77]The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research* 2018; **47**(D1):D330–D338.

[78]Kanehisa M, Sato Y, Furumichi M, *et al.* New approach for understanding genome variations in KEGG. *Nucleic acids research* 2019;**47**(D1):D590–D595.

[79]Vlachos IS, Zagganas K, Paraskevopoulou MD, *et al.* DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic acids research* 2015;**43**(W1):W460–6.

[80]Backes C, Kehl T, Stöckel D, *et al.* MiRPathDB: A new dictionary on microRNAs and target pathways. *Nucleic Acids Research* 2017; **45**(D1):D90–D96.

[81]Wang S, Kim J, Jiang X, *et al.* GAMUT: GPU accelerated microRNA analysis to uncover target genes through CUDA-miRanda. *BMC MEDICAL GENOMICS* 2014;**7**(1).

[82]Rusinov V, Baev V, Minkov IN, *et al.* MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. *NUCLEIC ACIDS RESEARCH* 2005;**33**(2):W696–W700.

[83]Maechler M, Rousseeuw P, Struyf A, *et al.* cluster: Cluster Analysis Basics and Extensions, 2018.

[84]Fridrich A, Hazan Y, Moran Y. Too Many False Targets for MicroRNAs: Challenges and Pitfalls in Prediction of miRNA Targets and Their Gene Ontology in Model and Non-model Organisms. *BioEssays* 2019;**41**(4):1800169.

[85]Wang N, Zheng J, Chen Z, *et al.* Single-cell microRNA-mRNA co-sequencing reveals non-genetic heterogeneity and mechanisms of microRNA regulation. *Nature Communications* 2019;**10**(1):95.

[86]Pinhal D, Bovolenta LA, Moxon S, *et al.* Genome-wide microRNA screening in Nile tilapia reveals pervasive isomiRs' transcription, sex-biased arm switching and increasing complexity of expression throughout development. *Scientific Reports* 2018;**8**(1):8248.

[87]Pla A, Zhong X, Rayner S. miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts. *PLOS Computational Biology* 2018;**14**(7):e1006185.

[88]Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* 2019;**25**(1):44–56.

[89]Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

[90]Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* 2011;**12**(1):35.

[91]Nan Xiao ML. ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for 'ggplot2', 2018.

[92]Brors B, Gu L, Schlesner M, *et al.* circlize implements and enhances circular visualization in R . *Bioinformatics* 2014;**30**(19):2811–2812.

[93]Galili T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 2015; **31**(22):3718–3720.

[94]Simon Garnier Noam Ross BRMSCS. viridis: Default Color Maps from 'matplotlib', 2018.

[95]Zerbino DR, Achuthan P, Akanni W, *et al.* Ensembl 2018. *Nucleic Acids Research* 2018;**46**(D1):D754–D761.

[96]Frankish A, Diekhans M, Ferreira AM, *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research* 2019;**47**(D1):D766–D773.

3.2 miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database

D142–D147 Nucleic Acids Research, 2020, Vol. 48, Database issue *Published online 6 November 2019*
doi: 10.1093/nar/gkz1022

miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database

Tim Kehl[1,†], Fabian Kern[2,†], Christina Backes[2], Tobias Fehlmann[2], Daniel Stöckel[1,3], Eckart Meese[4], Hans-Peter Lenhof[1] and Andreas Keller[2,5,6,*]

[1]Chair for Bioinformatics, Saarland Informatics Campus, 66123 Saarbrücken, Germany, [2]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [3]EMD Digital, Merck KGaA, Darmstadt, Germany, [4]Department of Human Genetics, Saarland University, 66421 Homburg, Germany, [5]School of Medicine Office, Stanford University, Stanford, CA, USA and [6]Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA, USA

Received September 15, 2019; Revised October 17, 2019; Editorial Decision October 17, 2019; Accepted October 19, 2019

ABSTRACT

Since the initial release of *miRPathDB*, tremendous progress has been made in the field of microRNA (miRNA) research. New miRNA reference databases have emerged, a vast amount of new miRNA candidates has been discovered and the number of experimentally validated target genes has increased considerably. Hence, the demand for a major upgrade of *miRPathDB*, including extended analysis functionality and intuitive visualizations of query results has emerged. Here, we present the novel release 2.0 of the miRNA Pathway Dictionary Database (*miRPathDB*) that is freely accessible at https://mpd.bioinf.uni-sb.de/. *miRPathDB* 2.0 comes with a tenfold increase of pre-processed data. In total, the updated database provides putative associations between 27 452 (candidate) miRNAs, 28 352 targets and 16 833 pathways for *Homo sapiens*, as well as interactions of 1978 miRNAs, 24 898 targets and 6511 functional categories for *Mus musculus*. Additionally, we analyzed publications citing *miRPathDB* to identify common use-cases and further extensions. Based on this evaluation, we added new functionality for interactive visualizations and down-stream analyses of bulk queries. In summary, the updated version of *miRPathDB*, with its new custom-tailored features, is one of the most comprehensive and advanced resources for miRNAs and their target pathways.

INTRODUCTION

Understanding the mechanisms of gene regulation is one of the major challenges in molecular biology and bioinformatics. In order to get the big picture, diverse sub-fields emerged to study the underlying principles of transcriptional, post-transcriptional, translational and post-translational levels of gene regulation. Short, conserved and non-coding RNA families, so-called microRNAs (miRNAs), were shown to orchestrate major pathways in a post-transcriptional manner by targeting 3′ untranslated regions (UTRs) of mRNAs in mammals and plants (1,2). While early studies focused on the validation of human microRNAs and those found in important model organisms such as mouse and rat, the focus has been broadly expanded to characterize miRNAs in a larger set of metazoan species (3). To this end, several reference databases such as miRBase, miRCarta and miRGeneDB and different nomenclatures were established (4–6). Since the number of miRNAs discovered is steadily rising (7), a remarkable amount of studies already validated microRNA target genes and their function in a multitude of cell-types, tissues and disease phenotypes (8,9). These global research efforts have led to an accumulation of novel data. To scale up with these developments and to gain deeper insights into miRNA functionality, robust statistical methods and curated databases are in great demand, especially to integrate all the important findings from miRNA discovery, target validation and target gene function (10,11).

One of the key questions of functional miRNA analysis is which pathways or cellular functions are regulated by a given miRNA (miRNA-centric view), or conversely, which miRNAs regulate a given gene set or pathway (pathway-centric view) (12,13). To solve these problems, several tools and databases have been proposed so far. From a miRNA-centric view, the miRTar database, which links individual miRNAs to metabolic pathways (14) and miRSystem, providing pre-computed enrichments of target genes in pathways (15), should be noted. Moreover, pure enrichment-based tools like miEAA, the bioconductor package miRNApath, or BUFET that are based on many-to-many rela-

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

tionships can process lists of miRNA identifiers to compute pathway associations (16–18). More specialized miRNA-centric tools include miRNet (19), which is a networks-based approach and miTALOS v2 (20) that annotates miRNA functions in a tissue-specific manner. PolymiRTS (21) is a pathway-centric database that maps SNPs in target sites to gene categories and phenotypes, i.e. disease traits.

Only a minor fraction of the tools and databases support both miRNA- and pathway-centric applications. These include the online database miRNApath (22), the R package CORNA (23), DIANA-miRPath v3.0 (24) incorporating GO and KEGG enrichments derived from predicted and validated miRNA-target interactions, and finally *miRPathDB* v1 (25), which in turn is based on our very first dictionary on miRNAs and target pathways (26).

After the initial release of *miRPathDB*, miRNA research has made notable progress. Novel miRNAs have been discovered, the number of experimentally validated target genes has increased tremendously. Most importantly, new reference databases emerged that either catalog validated miRNAs with high confidence (6), or that contain thousands of novel miRNA candidates (5). Additionally, we evaluated publications, citing our database, to identify common application scenarios, new visualizations and useful downstream applications (27–29). An overview of these publications can be found in Supplementary Table S1.

The new version of *miRPathDB*, provides access to target genes and regulated pathways not only for miRNAs from miRBase (Version 22.1), but also from miRCarta (Version 1.1). This increases the provided information by more than a factor of ten compared to the original version. Second, our database now also provides similarity information for all miRNAs based on their sequence, genomic position, target genes and target pathways. This information not only allows to query miRNAs with similar properties and to cluster miRNAs based on their similarity, but also to assess the regulatory potential of new candidate miRNAs. On top of the new data compilation, *miRPathDB* provides several interactive tools for user-specific analyses. From a miRNA perspective, we developed an appealing miRNA-to-pathway heatmap visualization that intuitively shows which pathways are regulated by a given set of miRNAs. To serve the pathway-centric use-case as well, we have formulated and implemented an Integer Linear Program (ILP) to automatically extract a set of miRNAs whose targetome covers a user-provided pathway or set of genes. Taken together, the new version of *miRPathDB* is a comprehensive resource to study the function of miRNAs in human and mouse.

MATERIALS AND METHODS

Our database integrates information of miRNAs, miRNA–target interactions (MTIs), and signaling pathways from several third-party resources. In the following sections, we describe the respective data sources and all processing steps performed to create the underlying data collection. Additionally, we describe the methodology of new downstream analysis features.

miRNA resources

The database stores information on all human and mouse miRNAs from miRBase (Version 22.1) and from miRCarta (Version 1.1), including miRNA candidates. Validated MTIs were acquired from miRTarBase (Version 7) (30) and pre-processed to create two subsets for each miRNA: all MTIs independent of their type of experimental evidence and only those with a strong level of evidence. On top of this, we predicted target genes for each miRNA sequence using TargetScan (Version 7.1) (31) and MiRanda (Version 3.3a) (32). Based on the prediction output, we also created two further list of MTIs: the intersection and the union of all predictions, which is a common strategy to account for putative sources of bias from target prediction tools and to balance sensitivity versus specificity (25,33). As $3'$ UTR input target set for the two algorithms, we used the curated annotations from *targetscan.org* for both human and mouse runs. Each program was executed using its default set of parameters.

Pathway databases and enrichment analysis

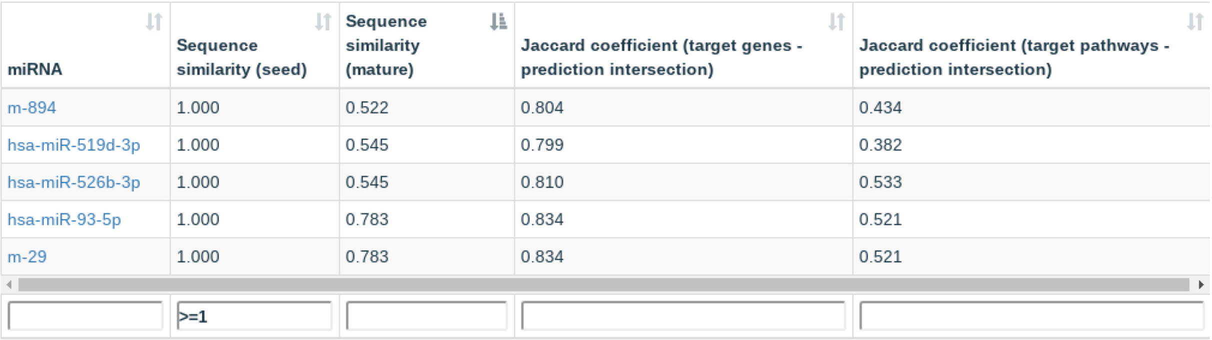
In order to determine whether a specific miRNA is associated with a particular biological process or signaling pathway, we used the enrichment analysis functionality of the GeneTrail2 C++ library (34). To this end, we analyzed functional categories from the Gene Ontology (35), as well as signaling pathways from KEGG (36), Reactome (37) and WikiPathways (38). For each pair of miRNA and functional category, we applied a hypergeometric test to check if the pathway contains significantly more target genes than expected by chance. Resulting p-values were FDR-adjusted (39) and a significance level of 0.05 was selected.

miRNA similarities

We also calculated similarities between all miRNAs and miRNA candidates based on their seed sequence, mature sequence, target genes and target pathways. For the string comparison, we calculated the Hamming distance between the sequences of all miRNA pairs, once using the full mature sequences and once the 7-nt substrings starting at position 2 from the $5'$ end of the mature sequence. Given the hamming distance H_d between two sequences of length l, we defined the pairwise sequence similarity as $1-(\frac{H_d}{l})$. The similarity of two sets containing either target genes or pathways was calculated using the Jaccard coefficient. Moreover, we compared miRNAs according to the positions of their genomic loci by computing the minimal distance between miRNAs annotated to the same chromosome.

Customized pathway heatmaps

The custom heatmap depicts which pathways are regulated by a user defined set of miRNAs. To create a heatmap, we first select all pathways that are significantly enriched for the targets of at least one of the specified miRNAs. The obtained p-values are used to construct a matrix that contains the $-\log_{10}$-transformed and discretized P-values for



miRNA	Sequence similarity (seed)	Sequence similarity (mature)	Jaccard coefficient (target genes - prediction intersection)	Jaccard coefficient (target pathways - prediction intersection)
m-894	1.000	0.522	0.804	0.434
hsa-miR-519d-3p	1.000	0.545	0.799	0.382
hsa-miR-526b-3p	1.000	0.545	0.810	0.533
hsa-miR-93-5p	1.000	0.783	0.834	0.521
m-29	1.000	0.783	0.834	0.521
	>=1			

Figure 1. Example of the new pairwise miRNA similarity table. The figure shows the pre-computed similarities for hsa-miR-106a-5p sorted by sequence similarity (mature) in increasing order. Furthermore, the table is filtered to show only miRNAs and miRNA candidates having 100% seed similarity. The Jaccard index provides additional information about the functional similarity of each miRNA and miR-106a-5p for predicted targets and target pathways.

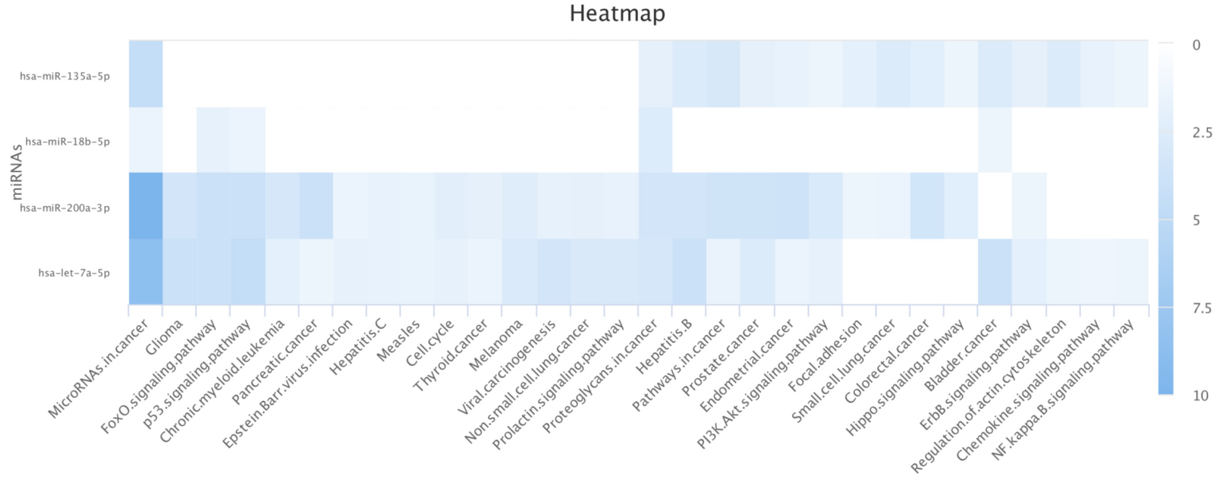


Figure 2. Example of the custom heatmap visualization. The figure depicts the enrichment results of hsa-miR-18b-5p, hsa-miR-135a-5p, hsa-let-7a-5p and hsa-miR-200a-3p for the categories of the KEGG database and strongly experimentally validated MTIs. Rows represent the enrichment results for the targets of the four miRNAs. Columns represent all KEGG pathways that are significant for the different miRNAs. For demonstration purposes, the heatmap was filtered to only show pathways with at least two associated miRNAs. The color of individual fields represent the $-\log_{10}$-transformed P-value of the respective enrichment results. Darker colors indicate more significant associations between miRNA and target pathway.

the set of miRNAs and all enriched pathways. Finally, similar miRNAs and pathways are clustered together by applying an hierarchical approach (Ward's method with Euclidian distance) to both rows and columns of the matrix. The clustered matrix is subsequently displayed as an interactive heatmap, implemented using the Highcharts JavaScript library.

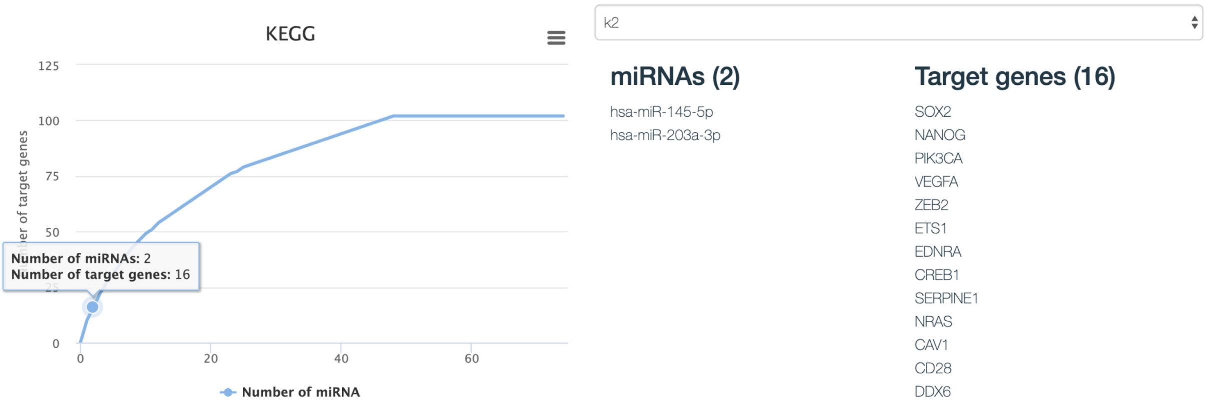
Maximum targetome coverage analysis

A noteworthy issue in functional miRNA research is to find a small number of miRNAs that are sufficient to regulate a given gene set, e.g. a particular signaling cascade or pathway. To solve this problem, we first search for the 'best' miRNA ($k = 1$) that regulates the maximal number of genes of the given target set. Next, we increase the considered number of miRNAs step-by-step ($k := k + 1$) until all target genes are covered or a predefined $k_{\max}$ is reached. For each k, we report an optimal set of miRNAs and the regulated target genes.

The problem to find the optimal set of miRNAs for one particular k is closely related to the maximum coverage problem, which can be solved using Integer Linear Programming (ILP). A formal definition of this problem can be found in the online documentation and Supplement S2. The ILP was implemented in C++ using the CPLEX optimization framework. Finally, results of an analysis are visualized by an interactive plot using the Highcharts JavaScript library.

OVERVIEW OF MIRPATHDB 2.0

miRPathDB stores information on (candidate) miRNAs, their target genes and their target pathways. To access this information, our database offers users two distinct representations: a miRNA-centric and a pathway-centric view. An overview table and a detailed description of each



miRNAs (2)	Target genes (16)
hsa-miR-145-5p	SOX2
hsa-miR-203a-3p	NANOG
	PIK3CA
	VEGFA
	ZEB2
	ETS1
	EDNRA
	CREB1
	SERPINE1
	NRAS
	CAV1
	CD28
	DDX6

Figure 3. Example of the interactive visualization for a user-specific maximum-coverage analysis. The curve on the left indicates how many of the specified target genes can be targeted by an increasing number of miRNAs. Here the x-axis shows the increasing number of miRNAs and the Y-axis the number of covered target genes. Users are able to click on every point of the curve to inspect the corresponding miRNAs and targeted genes. An example for $k = 2$ is depicted on the right-hand side.

miRNA or pathway are available. The representations can either be accessed through the overview tables or a query in the quick-search bar. A general description of these representations has already been presented in the original manuscript (26). Hence, we describe the extensive changes of the miRNA-centric view, the interactive analysis tools, and the new export functionality in the following sections.

NEW MIRNA-CENTRIC VIEW

Here, we explain the different levels of information *miRPathDB* offers for each miRNA or miRNA candidate.

General information

On the top of each miRNA page, we provide general information about the respective miRNA: the precursor mapping, the sequence of the mature miRNA, seed and corresponding parent stem loops, and all annotated genomic loci. Additionally, specific links to external reference database (miRBase and miRCarta) entries for all miRNAs and for corresponding precursors and family assignments are deposited. On top of this, each miRNA entry is linked to other third-party databases, like the TissueAtlas (40) or miRTargetLink (41), not only to improve the usability, but also to complement the features of *miRPathDB* with other essential tools for miRNA analysis.

miRNA-target interactions (MTIs)

Below the general information section, the website renders a responsive, sortable, and fully searchable table containing all target genes of an examined miRNA. For each gene, we also highlight in which of the four evidence sets it is contained. Table rows can be filtered using the text boxes below each column. Users can export both filtered and unfiltered tables in different file formats (CSV, Excel and PDF).

Targeted pathways

One major focus of our database is to provide information on associations between miRNAs and their putative target pathways. Likewise to the table for target genes, the pathways are shown in another fully responsive table. It contains, for the different evidence sets, all pathways that are significantly enriched with targets of the examined miRNAs. For each pathway, the number of contained target genes, the number of target genes that are expected by chance, and a FDR-adjusted *P*-value are listed. Since users might be interested in a specific subset of results, the table can be filtered with respect to all fields. For example, users can select significant pathways for a certain MTI evidence level, or only pathways that contain a specific gene of interest. Each pathway cell is linked to the corresponding external database entry, which often displays additional information like a description of the pathway or the underlying gene network.

miRNA similarities

At the bottom of each miRNA page, a novel table containing similarity information of the selected miRNA with respect to all other miRNAs from the same organism, including miRNA candidates, is displayed (Figure 1). The table lists the seed and full sequence similarities, the chromosomal distance, in case miRNAs are annotated on the same chromosome and eight Jaccard coefficients, measuring the similarity of target genes and target pathways for the different evidence sets of MTIs. Analogously to information about target genes and pathways, this table can be filtered, searched, sorted, resized, and exported for further usage.

NEW INTERACTIVE DATABASE FUNCTIONALITY

In addition to a new data compilation, *miRPathDB* features several new interactive tools for advanced user-specific database queries and analyses.

Custom pathway heatmaps

A common question in miRNA research is, whether targets of deregulated miRNAs are similarly enriched in certain biological processes or are associated with distinct molecular functions (27,28). In order to help users to tackle this question, we developed an interactive heatmap visualization. To create this plot, a user needs to specify a list of miRNAs as well as the evidence level for the MTIs. *miRPathDB* automatically selects all functional categories that are significantly enriched for the targets of at least one of the specified miRNAs. Results are represented as a heatmap, where each row depicts enrichment results for the respective functional categories. The color of individual entries corresponds to the p-value of the associated enrichment result. Darker colors indicate more significant enrichments of miRNA target genes in the corresponding biological processes. On top of this, users may specify the resolution of the resulting heatmap and download the image in different file formats (PNG, JPEG, PDF and SVG). An example heatmap is shown in Figure 2. Our customized heatmap feature provides a rapid overview of molecular functions and signaling pathways that are potentially regulated by a specific miRNA set. This analysis might even be helpful to assess possible downstream effects of deregulated miRNAs in high-throughput studies.

Maximum targetome coverage analysis

While the previous feature allows downstream analysis from a miRNA-centric view, by mapping a given miRNA set to enriched target pathways, *miRPathDB* also provides functionality for the reverse direction, i.e. given a set of target genes G, find a minimal set of miRNAs that target all genes in G. To this end, we provide a tool that iteratively computes k miRNAs (for all $k \in \{1, 2, ..., k_{\max}\}$) with a maximal number of targets in G (see Materials and Methods). To start the maximum coverage analysis, a user must upload a list of genes, select the desired level of evidence that should be used to lookup the MTIs and set the largest $k = k_{\max}$ where the algorithm should stop. The results of such an analysis are displayed in an interactive line-graph that plots k against the number of covered target genes (Figure 3, left). For each k, a node is inserted in the graph that can be selected. Upon selection of a node, the website displays an optimal set of miRNAs of the corresponding size k along with the list of overlapping target genes (Figure 3, right).

DATA EXPORT

Most of the views in *miRPathDB* offer dedicated export functionality. All tables in the miRNA-centric and the pathway-centric view can be filtered and downloaded in different formats (CSV, Excel and PDF). Additionally, we host downloads for all processing steps of the enrichment analyses. Users are able to acquire the unprocessed enrichment results, i.e. a table containing detailed information for each functional category. Furthermore, a table containing all pairs of miRNA and pathways and their $-log_{10}$-transformed p-values is available. *miRPathDB* also supplies all functional categories in Gene Matrix Transposed (GMT) format (cf. Online documentation).

CONCLUSION

Recent advancements in miRNA research yielded huge numbers of novel miRNAs, miRNA candidates, and experimentally validated MTIs. This circumstance motivated a novel release of *miRPathDB*. Besides miRNAs and their targets, our database also provides information about associations between pathways and miRNAs. Beyond the tenfold increase of data, our database now offers powerful tools for the visualization and downstream analysis of database queries. In particular, users are able to search similar miRNAs, create interactive clustered heatmaps and to determine a minimal set of candidate regulators that are sufficient to target a specified gene list. In summary, *miRPathDB* 2.0 is the most comprehensive publicly available resource to assess the relationship between microRNAs, their targets and cellular functions for human and mouse.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: Internal funds of Saarland University.

Conflict of interest statement. None declared.

REFERENCES

1. Jonas,S. and Izaurralde,E. (2015) Towards a molecular understanding of microRNA-mediated gene silencing. *Nat. Rev. Genet.*, **16**, 421–433.
2. Bartel,D.P. (2018) Metazoan microRNAs. *Cell*, **173**, 20–51.
3. Fehlmann,T., Backes,C., Pirritano,M., Laufer,T., Galata,V., Kern,F., Kahraman,M., Gasparoni,G., Ludwig,N., Lenhof,H.P. *et al.* (2019) The sncRNA Zoo: a repository for circulating small noncoding RNAs in animals. *Nucleic Acids Res.*, **47**, 4431–4441.
4. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2018) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
5. 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.
6. Fromm,B., Domanska,D., Høye,E., Ovchinnikov,V., Kang,W., Aparicio-Puerta,E., Johansen,M., Flatmark,K., Mathelier,A., Hovig,E. *et al.* (2019) MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res.*, doi:10.1093/nar/gkz885.
7. 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.*, **47**, 3353–3364.
8. Rupaimoole,R. and Slack,F.J. (2017) MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat. Rev. Drug Discov.*, **16**, 203–221.
9. Karagkouni,D., Paraskevopoulou,M.D., Chatzopoulos,S., Vlachos,I.S., Tastsoglou,S., Kanellos,I., Papadimitriou,D., Kavakiotis,I., Maniou,S., Skoufos,G. *et al.* (2018) DIANA-TarBase v8: A decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic Acids Res.*, **46**, D239–D245.
10. Liu,B., Li,J. and Cairns,M.J. (2012) Identifying miRNAs, targets and functions. *Brief. Bioinformatics*, **15**, 1–19.
11. Fehlmann,T., Laufer,T., Backes,C., Kahramann,M., Alles,J., Fischer,U., Minet,M., Ludwig,N., Kern,F., Kehl,T. *et al.* (2019) Large-scale validation of miRNAs by disease association, evolutionary conservation and pathway activity. *RNA Biol.*, **16**, 93–103.
12. Davis,J.A., Saunders,S.J., Mann,M. and Backofen,R. (2017) Combinatorial ensemble miRNA target prediction of co-regulation

networks with non-prediction data. *Nucleic Acids Res.*, **45**, 8745–8757.

13. Sticht,C., De La Torre,C., Parveen,A. and Gretz,N. (2018) miRWalk: an online resource for prediction of microRNA binding sites. *PLoS ONE*, **13**, 1–6.
14. Hsu,J.B., Chiu,C.M., Hsu,S.D., Huang,W.Y., Chien,C.H., Lee,T.Y. and Huang,H.D. (2011) MiRTar: an integrated system for identifying miRNA-target interactions in human. *BMC Bioinformatics*, **12**. 300.
15. Lu,T.P., Lee,C.Y., Tsai,M.H., Chiu,Y.C., Hsiao,C.K., Lai,L.C. and Chuang,E.Y. (2012) miRSystem: An Integrated System for Characterizing Enriched Functions and Pathways of MicroRNA Targets. *PLoS ONE*, **7**, 1–10.
16. Backes,C., Khaleeq,Q.T., Meese,E. and Keller,A. (2016) MiEAA: MicroRNA enrichment analysis and annotation. *Nucleic Acids Res.*, **44**, W110–W116.
17. Cogswell,J.P., Ward,J., Taylor,I.A., Waters,M., Shi,Y., Cannon,B., Kelnar,K., Kemppainen,J., Brown,D., Chen,C. *et al.* (2008) Identification of miRNA changes in Alzheimer's disease brain and CSF yields putative biomarkers and insights into disease pathways. *J. Alzheimer's Dis.: JAD*, **14**, 27–41.
18. Zagganas,K., Vergoulis,T., Paraskevopoulou,M.D., Vlachos,I.S., Skiadopoulos,S. and Dalamagas,T. (2017) BUFET: boosting the unbiased miRNA functional enrichment analysis using bitsets. *BMC Bioinformatics*, **18**, 399.
19. Fan,Y., Siklenka,K., Arora,S.K., Ribeiro,P., Kimmins,S. and Xia,J. (2016) miRNet - dissecting miRNA-target interactions and functional associations through network-based visual analysis. *Nucleic Acids Res.*, **44**, W135–W141.
20. Preusse,M., Theis,F.J. and Mueller,N.S. (2016) miTALOS v2: analyzing tissue specific microRNA function. *PLoS ONE*, **11**, 1–15.
21. Bhattacharya,A., Ziebarth,J.D. and Cui,Y. (2013) PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. *Nucleic Acids Res.*, **42**, D86–D91.
22. Chiromatzo,A.O., Oliveira,T.Y.K., Pereira,G., Costa,A.Y., Montesco,C.A.E., Gras,D.E., Yosetake,F., Vilar,J.B., Cervato,M., Prado,P.R.R. *et al.* (2007) miRNApath: a database of miRNAs, target genes and metabolic pathways. *Genet. Mol. Res.: GMR*, **6**, 859–865.
23. Wu,X. and Watson,M. (2009) CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics*, **25**, 832–833.
24. 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.
25. Backes,C., Kehl,T., Stöckel,D., Fehlmann,T., Schneider,L., Meese,E., Lenhof,H.P. and Keller,A. (2017) MiRPathDB: A new dictionary on microRNAs and target pathways. *Nucleic Acids Res.*, **45**, D90–D96.
26. Backes,C., Meese,E., Lenhof,H.P. and Keller,A. (2010) A dictionary on microRNAs and their putative target pathways. *Nucleic Acids Res.*, **38**, 4476–4486.
27. Denham,J., Gray,A.J., Scott-Hamilton,J., Hagstrom,A.D. and Murphy,A.J. (2018) Small non-coding RNAs are altered by short-term sprint interval training in men. *Physiol. Rep.*, **6**, e13653.
28. Ragni,E., De Luca,P., Perucca Orfei,C., Colombini,A., Viganò,M., Lugano,G., Bollati,V. and de Girolamo,L. (2019) Insights into inflammatory Priming of adipose-derived mesenchymal stem cells: validation of extracellular vesicles-embedded miRNA reference genesas a crucial step for donor selection. *Cells*, **8**, E369.
29. Kehl,T., Backes,C., Kern,F., Fehlmann,T., Ludwig,N., Meese,E., Lenhof,H.P. and Keller,A. (2017) About miRNAs, miRNA seeds, target genes and target pathways. *Oncotarget*, **8**, 107167–107175.
30. Chou,C.H., Shrestha,S., Yang,C.D., Chang,N.W., Lin,Y.L., Liao,K.W., Huang,W.C., Sun,T.H., Tu,S.J., Lee,W.H. *et al.* (2018) MiRTarBase update 2018: A resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **46**, D296–D302.
31. 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.
32. Enright,A.J., John,B., Gaul,U., Tuschl,T., Sander,C. and Marks,D.S. (2003) MicroRNA targets in Drosophila. *Genome Biol.*, **5**, R1.
33. Bhattacharya,A. and Cui,Y. (2015) MiR2GO: comparative functional analysis for microRNAs. *Bioinformatics*, **31**, 2403–2405.
34. Stöckel,D., Kehl,T., Trampert,P., Schneider,L., Backes,C., Ludwig,N., Gerasch,A., Kaufmann,M., Gessler,M., Graf,N. *et al.* (2016) Multi-omics enrichment analysis using the GeneTrail2 web service. *Bioinformatics*, **32**, 1502–1508.
35. The Gene Ontology Consortium (2018) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
36. Kanehisa,M., Sato,Y., Furumichi,M., Morishima,K. and Tanabe,M. (2018) New approach for understanding genome variations in KEGG. *Nucleic Acids Res.*, **47**, D590–D595.
37. Fabregat,A., Jupe,S., Matthews,L., Sidiropoulos,K., Gillespie,M., Garapati,P., Haw,R., Jassal,B., Korninger,F., May,B. *et al.* (2017) The Reactome Pathway Knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
38. Slenter,D.N., Kutmon,M., Hanspers,K., Riutta,A., Windsor,J., Nunes,N., Mélius,J., Cirillo,E., Coort,S.L., Digles,D. *et al.* (2017) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.*, **46**, D661–D667.
39. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodological)*, **57**, 289–300.
40. 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.
41. Hamberg,M., Backes,C., Fehlmann,T., Hart,M., Meder,B., Meese,E. and Keller,A. (2016) MiRTargetLink-miRNAs, genes and interaction networks. *Int. J.f Mol. Sci.*, **17**, 564.

3.3 *Machine Learning to Detect Alzheimer's Disease from Circulating Non-coding RNAs*

Genomics Proteomics Bioinformatics 17 (2019) 430–440

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Machine Learning to Detect Alzheimer's Disease from Circulating Non-coding RNAs

Nicole Ludwig [1,#,a], Tobias Fehlmann [2,#,b], Fabian Kern [2,#,c], Manfred Gogol [3,d], Walter Maetzler [4,5,6,e], Stephanie Deutscher [1,f], Simone Gurlit [7,g], Claudia Schulte [5,6,h], Anna-Katharina von Thaler [5,6,i], Christian Deuschle [5,6,j], Florian Metzger [8,k], Daniela Berg [4,5,6,l], Ulrike Suenkel [5,6,m], Verena Keller [9,n], Christina Backes [2,o], Hans-Peter Lenhof [10,p], Eckart Meese [1,q], Andreas Keller [2,10,*,r]

[1] *Department of Human Genetics, Saarland University, 66421 Homburg/Saar, Germany*
[2] *Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany*
[3] *Institut für Gerontologie, Universität Heidelberg, 69047 Heidelberg, Germany*
[4] *Department of Neurology, Christian-Albrechts-University of Kiel, 24118 Kiel, Germany*
[5] *Center for Neurology and Hertie Institute for Clinical Brain Research, Department of Neurodegeneration, University of Tuebingen, 72074 Tuebingen, Germany*
[6] *German Center for Neurodegenerative Diseases (DZNE), 72076 Tuebingen, Germany*
[7] *Department of Anesthesiology and Intensive Care, St. Franziskus Hospital Muenster, 48145 Muenster, Germany*
[8] *Department of Psychiatry and Psychotherapy, University Hospital Tuebingen, 72016 Tuebingen, Germany*
[9] *Department of Medicine II, Saarland University Medical Center, 66421 Homburg/Saar, Germany*
[10] *Center for Bioinformatics, Saarland Informatics Campus, 66123 Saarbrücken, Germany*

[*] Corresponding author.
E-mail: andreas.keller@ccb.uni-saarland.de (Keller A).

[#] Equal contribution.

[a] ORCID: 0000-0003-4703-7567.

[b] ORCID: 0000-0003-1967-2918.

[c] ORCID: 0000-0002-8223-3750.

[d] ORCID: 0000-0001-7169-2970.

[e] ORCID: 0000-0002-5945-4694.

[f] ORCID: 0000-0003-4080-708X.

[g] ORCID: 0000-0002-3944-7841.

[h] ORCID: 0000-0003-4006-1265.

[i] ORCID: 0000-0001-8161-5813.

[j] ORCID: 0000-0001-5571-7293.

[k] ORCID: 0000-0002-8236-1170.

[l] ORCID: 0000-0003-1187-219X.

[m] ORCID: 0000-0002-5348-3996.

[n] ORCID: 0000-0003-3240-1397.

[o] ORCID: 0000-0001-9330-9290.

[p] ORCID: 0000-0002-5820-9961.

[q] ORCID: 0000-0001-7569-819X.

[r] ORCID: 0000-0002-5361-0895.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2019.09.004

Received 14 May 2019; revised 26 August 2019; accepted 7 October 2019
Available online 4 December 2019

Handled by Hongxing Lei

KEYWORDS

miRNAs;
Neurodegeneration;
Alzheimer's disease;
Biomarker;
Non-coding RNAs;
Gene regulation

Abstract Blood-borne small non-coding (sncRNAs) are among the prominent candidates for blood-based diagnostic tests. Often, high-throughput approaches are applied to discover **biomarker** signatures. These have to be validated in larger cohorts and evaluated by adequate statistical learning approaches. Previously, we published high-throughput sequencing based microRNA (miRNA) signatures in **Alzheimer's disease** (AD) patients in the United States (US) and Germany. Here, we determined abundance levels of 21 known circulating **miRNAs** in 465 individuals encompassing AD patients and controls by RT-qPCR. We computed models to assess the relation between miRNA expression and phenotypes, gender, age, or disease severity (Mini-Mental State Examination; MMSE). Of the 21 miRNAs, expression levels of 20 miRNAs were consistently de-regulated in the US and German cohorts. 18 miRNAs were significantly correlated with **neurodegeneration** (Benjamini-Hochberg adjusted $P < 0.05$) with highest significance for miR-532-5p (Benjamini-Hochberg adjusted $P = 4.8 \times 10^{-30}$). Machine learning models reached an area under the curve (AUC) value of 87.6% in differentiating AD patients from controls. Further, ten miRNAs were significantly correlated with MMSE, in particular miR-26a/26b-5p (adjusted $P = 0.0002$). Interestingly, the miRNAs with lower abundance in AD were enriched in monocytes and T-helper cells, while those up-regulated in AD were enriched in serum, exosomes, cytotoxic t-cells, and B-cells. Our study represents the next important step in translational research for a miRNA-based AD test.

Introduction

Alzheimer's disease (AD) represents one of the most demanding challenges in healthcare [1,2]. In light of demographic changes and failures in drug development [3], early detection of the disease offers itself as one of the most promising approaches to improve patients' outcome in the mid- to long term. Especially minimally invasive molecular markers seem to have a significant potential to facilitate a diagnosis of AD, even in early stages.

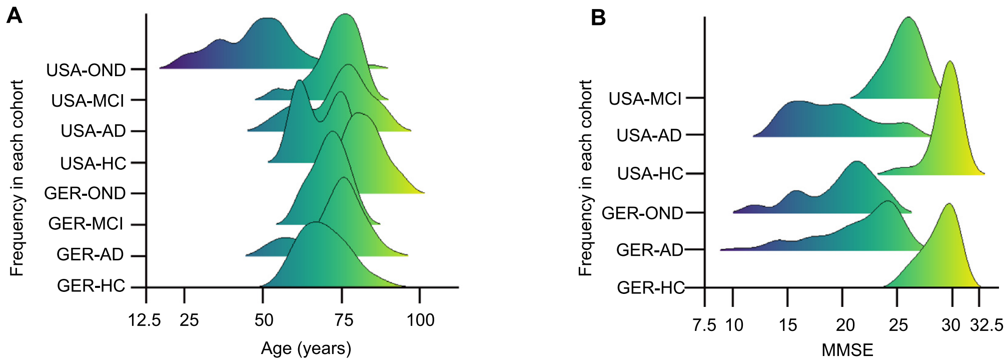
The importance of minimally invasive molecular markers for AD is reflected by over 3000 original articles and reviews related to AD diagnosis from blood, serum, or plasma samples published and indexed in PubMed. Among the promising approaches are plasma proteomic markers measured by mass spectrometry [4], metabolic patterns [5], gene expression profiles [6], DNA methylation [7], and small non-coding RNAs (sncRNAs) [8]. However, cohort sizes of such studies are often limited and larger validation cohorts frequently did not always match the original results [9]. One of the major challenges is the complexity of signatures that is often required to reach high specificity and sensitivity.

For AD, many miRNA-related studies from tissue [10], blood [11], serum [12], exosomes, [13] or cerebrospinal fluid (CSF) [12] have been performed. In one of the most comprehensive reviews [14], Hu and co-workers investigated 236 papers and reviewed the de-regulated miRNA abundance in different parts of AD patients. In another comprehensive recent review, Nagaraj and co-workers show that out of 137 miRNAs found to exhibit altered expression in AD blood, 36 have been replicated in at least one independent study. Moreover, out of 166 miRNAs being differentially abundant in AD CSF, 13 have been repeatedly found [15].

In previous studies, we performed deep sequencing to measure blood-borne AD miRNA signatures in a cohort of 54 AD patients and 22 controls from the United States (USA) that have been partially validated on a larger cohort of 202 samples by RT-qPCR [8]. In a second study using the same technique, we aimed to validate the results in a patient cohort collected in Germany (GER) that included 49 AD cases, 55 controls and 110 disease controls [16]. The results of both studies were largely consistent with a correlation between both studies of 0.93 (95% confidence interval 0.89–0.96; $P < 10^{-16}$).

Although deep-sequencing applications are increasingly introduced into clinical care, they are mostly performed for the analysis of DNA or RNAs coding for genes. Small non-coding RNA profiling, however, is mostly achieved by microarray and RT-qPCR based approaches. In the present study, we provide further evidence that blood-borne miRNA signatures can be measured by standard RT-qPCR, becoming valuable tools for the minimally-invasive detection of AD. From our above-mentioned studies and the literature, we selected a set of 21 miRNAs and determined the abundance of these miRNAs in the blood of 465 individuals. The 465 individuals consist of 169 individuals from our initial study (36%) [8], 107 individuals from the second study (23%) [16] as well as 189 newly collected individuals (41%). An overview and summary on the German and US samples is provided in **Figure 1**A–C, the full details for each individual samples, including age gender, diagnosis, Mini-Mental State Examination (MMSE), and the miRNA measurements, are provided in Table S1.

With the present study we pursue the five main goals to demonstrate that (1) miRNAs from NGS studies can be well reproduced by RT-qPCR experiments; (2) given a reasonable heterogeneity in samples still reproducible measurements in



C

	AD		MCI		HC		OND	
	USA	GER	USA	GER	USA	GER	USA	GER
No of subjects	79	66	17	21	23	191	50	18
Gender (M/F)	40/39	36/30	8/9	10/11	11/12	92/99	16/34	7/11
Age (mean ± SD)	73.2 ± 10.3	72.5 ± 9.3	72.8 ±7.2	70.6 ± 5.3	67.4 ±6.7	69.1 ±8	48.2 ± 13.5	81.7 ± 6.4
MMSE (mean ± SD)	19 ± 3.5	21 ±4.7	25.6 ±1.4	NA	29.4 ± 1.2	28.8 ±1.5	NA	18.8 ± 5.3

Figure 1 Distribution of age, gender, diseases, and MMSE

A. Histogram for the age distribution in the different cohorts. The diagram shows for each cohort/disease the age distribution. Only the OND group from the US shows a deviation towards younger patients, while all other groups have similar age ranges. **B.** Histogram for the MMSE values. HCs and MCI patients show significantly larger MMSE values as compared to AD and OND patients. **C.** Metrics. For each of the cohorts and diseases, the number of patients in the US and Germany, the mean and SD for age and MMSE as well as the gender distribution are provided. GER, Germany; MMSE, Mini-Mental State Examination; AD, Alzheimer's disease; OND, other neurological diseases; HC, healthy control; MCI, mild cognitive impairment.

larger cohorts are possible; (3) miRNAs are also correlated to clinical features such as the MMSE value; (4) statistical learning approaches with as few as possible features lead to accurate diagnostic results; (5) the miRNAs likely have functionality in AD via targeting genes.

Results

Two endogenous control RNAs show concordant results

Because the selection of the most appropriate endogenous control RNAs for RT-qPCR experiments can be challenging, we previously evaluated systematically whether different endogenous controls lead to differences in miRNA measurements [17]. Especially, most miRNAs seem to be affected by development stages, tissues [18], or diseases [19], limiting their ability as controls and calling for endogenous controls other than miRNAs. Our results suggested that differences can be observed that are however moderate. In the present study we nonetheless evaluated and compared the performance of two commonly used endogenous controls RNU48 and RNU6. Both endogenous controls have been measured in duplicates. In comparing the results, we verified the generally high concordance between the two endogenous controls with a Pearson correlation of 0.854 (95% CI: 0.828–0.877; $P < 10^{-16}$). We thus report the result in the current study based on our standard endogenous control RNU48.

In the same direction we also investigated the general stability of RT-qPCR based miRNA measurement. One control sample has been measured 12 times over the study for all miRNAs. The median Pearson correlation coefficient (PCC) exceeded 0.99 as the heatmap and the box plot in Figure S1 show.

miRNAs are highly significantly correlated with neurodegeneration

In total, 465 participants have been analyzed by RT-qPCR. The abundance levels of 18 of the 21 miRNAs were significantly different between the four groups considered, *i.e.*, AD, mild cognitive impairment (MCI), other neurological diseases (OND), and healthy controls (HC). With an Benjamini-Hochberg (BH) adjusted P value of 4.8×10^{-30}, the most significant miRNA was miR-532-5p, which showed markedly decreased levels in AD patients, and slightly decreased levels in patients with OND and MCI (**Figure 2**A). The abundance levels of miR-17-3p, the miRNA with the second lowest P value ($P = 8.8 \times 10^{-28}$), showed a similar pattern as miR-532-5p (PCC > 0.9). The overall correlation matrix between the 21 miRNAs showed three large clusters of miRNAs with similar expression in the following referred to as Clusters A, B, and C (Figure 2B). The third and fourth most significant miRNAs in ANOVA, *i.e.*, miR-103a-3p and miR-107 ($P = 2.4 \times 10^{-18}$ and $P = 3.6 \times 10^{-15}$, respectively), came from Cluster C, like miR-532-5p, and miR-17-3p. MiR-1468-5p (Cluster A, $P = 6.2 \times 10^{-12}$; Figure 2C) shows an opposite expression pattern, i.e. a higher abundance in AD patients as compared to HC. The boxplots in Figure 2A/2C also underline that the deregulation of these miRNAs is strongest in AD compared to the HC. There is, however, a deregulation in MCI or OND, but to a lesser extent, such that the altered abundance is at least partially specific for AD. This result is consistent with our previous work based on high-throughput sequencing.

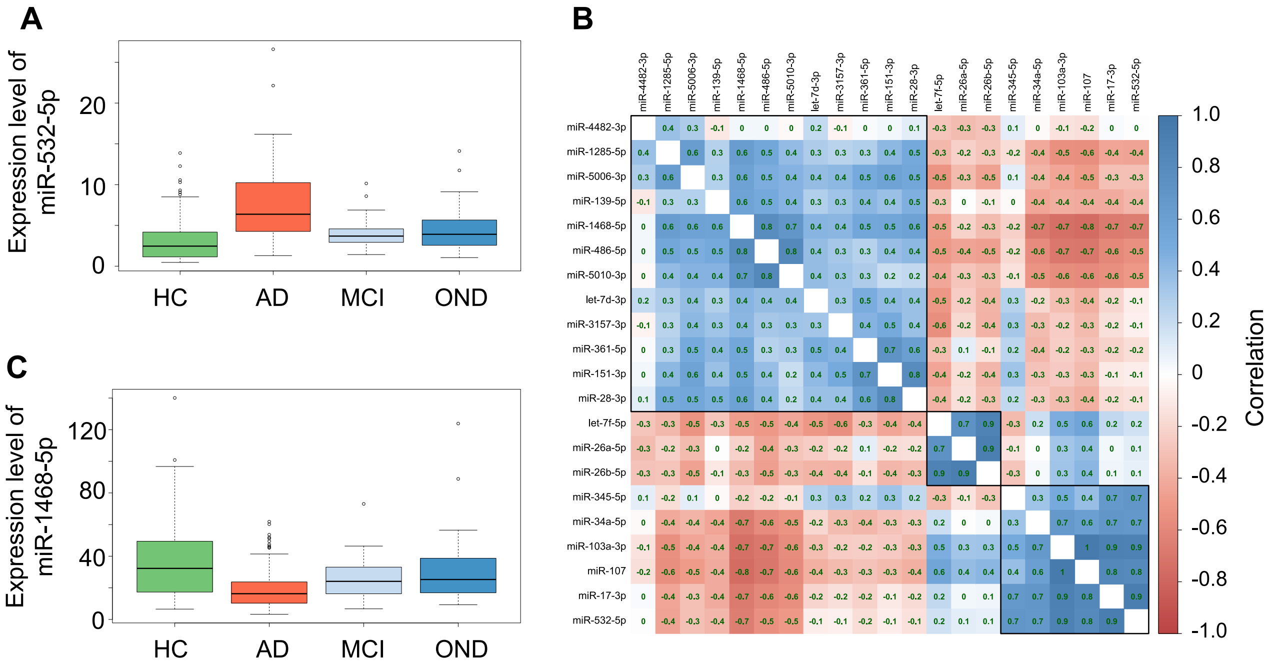


Figure 2 miRNAs are specifically dysregulated in the four cohorts and are partially co-expressed
A. Expression of miR-532-3p. The boxes display the 2nd and 3rd quartile of expression values for miR-532-3p in HC, patients with AD, MCI, or OND. The range of expression values in the four groups is indicated by the error bars with outliers represented by unfilled dots. Median expression of miR-532-3p is indicated as thick black line. **B.** Correlation of miRNA expression. This correlation matrix graphically represents the pair-wise correlation coefficient for all miRNAs tested. According to the color scale on the right side of the matrix, positive and negative correlations are indicated in shades of blue and red, respectively. PCC is given for each pair-wise correlation. Three clusters of miRNAs with highly similar expression patterns are indicated as Clusters A, B, and C on the left side. **C.** Expression of miR-1468-5p. The boxes display the 2nd and 3rd quartile of expression values for miR-1468-5p in HC, patients with AD, MCI, or OND. The range of expression values in the four groups is indicated by the error bars with outliers represented by unfilled dots. Median expression of miR-1468-5p is indicated as thick black line. PCC, Pearson correlation coefficient.

For a more detailed understanding of the miRNAs and their correlation to AD and other factors, we next assessed whether the abundance levels were correlated to age or gender, or, in case of AD and MCI with the MMSE results (**Table 1**). As Table 1 highlights, none of the miRNAs was associated with gender and five miRNAs were weakly associated with age of patients. Following adjustment for multiple testing, 14 miRNAs showed a significant differential expression in AD patients compared to controls (*i.e.*, HC, MCI, and OND combined). The above mentioned miR-532-5p and miR-17-3p were again the most significant markers for AD. Furthermore, ten miRNAs were significantly correlated with the MMSE value. Interestingly, all three miRNAs of Cluster B (Figure 1B), *i.e.*, miR-26a, 26b-5p, and let-7f-5p, showed the highest significance for the correlation to MMSE ($P < 0.005$). Since neither all miRNAs nor the MMSE values were normally distributed we repeated the analyses with non-parametric and ranked based Spearman correlation coefficient (SCC), overall leading to comparable results (see Table S2).

Besides the comparison of healthy controls to AD we also asked whether MCI patients can be separated from AD patients using miRNAs. Indeed, eleven miRNAs had significant differential expression in MCI versus AD following adjustment for multiple testing: miR-17-3p ($P = 10^{-12}$; down in AD), miR-532-5p ($P = 8 \times 10^{-10}$; down in AD), miR-103a-3p ($P = 10^{-8}$; down in AD), miR-107 ($P = 4 \times 10^{-7}$; down in AD), let-7d-3p ($P = 9 \times 10^{-7}$; up in AD), let-7f-5p ($P = 3 \times 10^{-5}$; down in AD), miR-345-5p ($P = 0.0002$; down in AD), miR-26a-5p ($P = 0.002$; down in AD), miR-26b-5p ($P = 0.009$; down in AD), miR-1468-5p ($P = 0.02$; up in AD), and miR-139-5p ($P = 0.03$; up in AD).

miRNA profiles from the US and German cohort show consistent results

It is essential to understand whether biomarkers can be concordantly determined in different cohorts. Although a direct comparison of ethnic groups was not in the scope of our analysis we nonetheless asked whether miRNA profiles for one disease measured on two different continents are concordant to each other. We thus compared the profiles measured from GER and USA cohorts. As the GER cohort was about twice as large as the USA cohort and P values depend on the number of individuals in each cohort, a comparison based only on P values is potentially biased. Therefore, we computed the fold changes (on a logarithmic scale) between AD and controls (**Figure 3**A). In this plot miRNAs in the upper right quadrant are down-regulated and miRNAs in the lower left quadrant

Table 1 Raw and adjusted *P* values of miRNAs for age, gender, AD, and MMSE

miRNA	Gender		Age		AD		MMSE	
	Raw	Adjusted	Raw	Adjusted	Raw	Adjusted	Raw	Adjusted
miR-**532-5**p	0.3466	0.4917	0.3089	0.4634	5.99E–22	**1.26E–20**	0.5048	0.5890
miR-**17-3**p	0.4885	0.5811	0.0639	0.2238	6.24E–18	**6.55E–17**	0.5004	0.5890
miR-**1468-5**p	0.0568	0.1491	0.5645	0.6973	5.98E–16	**4.19E–15**	0.0016	0.0057
miR-**5010-3**p	0.0176	0.0971	0.4787	0.6283	4.81E–12	**2.52E–11**	0.0131	0.0345
miR-**103**a-**3**p	0.3596	0.4917	0.2791	0.4508	1.56E–11	**6.56E–11**	0.5805	0.6416
miR-**1285-5**p	0.5195	0.5811	0.8097	0.8502	4.85E–11	**1.70E–10**	0.2269	0.3725
miR-**345-5**p	0.2217	0.4041	0.0008	0.0081	8.85E–11	**2.65E–10**	0.0174	0.0364
miR-**107**	0.2174	0.4041	0.3568	0.4995	6.94E–10	**1.82E–09**	0.7545	0.7545
miR-**486-5**p	0.5535	0.5811	0.9667	0.9667	2.79E–06	**6.51E–06**	0.2306	0.3725
miR-**139-5**p	0.0031	0.0656	0.7862	0.8502	8.12E–05	**0.0002**	0.3384	0.4441
miR-**361-5**p	0.3747	0.4917	0.0032	0.0180	0.0004	**0.0008**	0.0896	0.1710
miR-**5006-3**p	0.2271	0.4041	0.1433	0.3352	0.0006	**0.0011**	0.0043	0.0130
miR-**28-3**p	0.2309	0.4041	0.6780	0.7909	0.0057	0.0093	0.6572	0.6900
miR-**34**a-**5**p	0.0185	0.0971	0.2526	0.4508	0.0067	0.0100	0.2486	0.3730
miR-**4482-3**p	0.0378	0.1325	0.0004	0.0078	0.0365	0.0511	0.0015	0.0057
let-**7**f-**5**p	0.0454	0.1362	0.1596	0.3352	0.0702	0.0921	0.0002	**0.0016**
miR-**3157-3**p	0.3504	0.4917	0.2626	0.4508	0.0833	0.1029	0.0013	0.0057
miR-**151-3**p	0.0131	0.0971	0.1057	0.3170	0.0902	0.1052	0.3031	0.4244
miR-**26**b-**5**p	0.7003	0.7003	0.0034	0.0180	0.1101	0.1217	1.82E–05	**0.0002**
miR-**26**a-**5**p	0.5506	0.5811	0.0063	0.0263	0.2939	0.3085	1.19E–05	**0.0002**
let-**7**d-**3**p	0.0323	0.1325	0.1489	0.3352	0.4834	0.4834	0.0172	0.0364

Note: *P* values for gender and AD were calculated based on *t* test; *P* values for age and MMSE were calculated based on Pearson's product moment correlation coefficient. *P* values were adjusted by the Benjamini-Hochberg procedure. Adjusted *P* values < 0.05 are indicted in orange and those < 0.005 are put in bold with blue background. AD, Alzheimer's disease; MMSE, Mini-Mental State Examination.

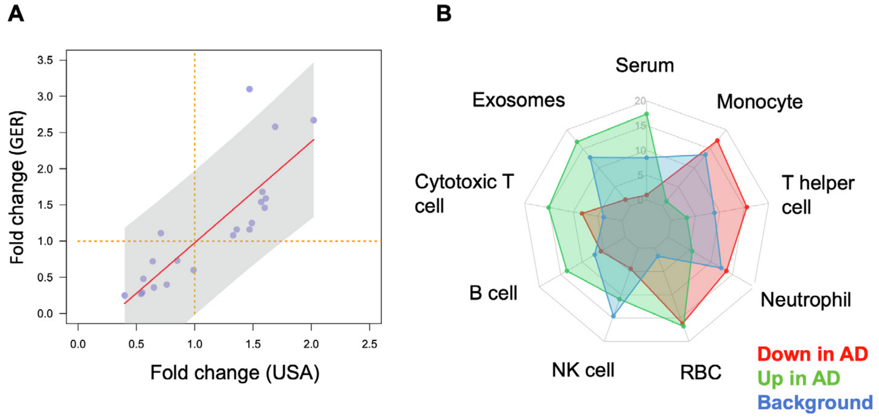


Figure 3 Differentially-expressed miRNAs are concordantly expressed in the German and the US cohorts and belong to specific blood compounds

A. Fold change in the USA cohort compared to the GER cohort. The X- and Y-axes represent the fold change between AD and HC on a $\log_2$ scale for the USA and GER patient cohorts, respectively. Each miRNA is represented by one dot. The dashed orange line is the segregation between up- and down-regulation. miRNAs in the upper right or lower left quadrant are concordantly up- or downregulated in AD compared to HC in both cohorts, respectively. The solid red line is a linear regression fit and the shaded area is the 95% confidence interval of that fit. **B.** Radar chart showing the blood compound distribution. The plot shows the relative abundance of up-regulated, down-regulated, and all miRNAs in different blood compounds. Since the relative abundance is provided, it is more appropriate to compare the different groups within one specific compound rather than comparing different compounds to each other.

are up-regulated in AD compared to controls concordantly in both cohorts. Of 21 miRNAs, only miR-4482-3p was down-regulated in the GER cohort, but up-regulated in the USA cohort. The differences in abundance levels of this miRNA in AD compared to controls were, however, not significant, neither in the GER nor in the USA cohort, nor in the combined analysis. Thus, miR-4482-3p likely represents a single false positive marker from the initial deep-sequencing based miRNA discovery study. In contrast, the results for the remaining 20 miRNAs were concordant between the USA and the GER cohort. Furthermore, eleven of these miRNAs were nominally significant in both cohorts, when analyzing the USA cohort and the GER cohort separately, and remained significant in the combined analysis. These significant miRNAs include miR-103a-3p, miR-107, miR-1285-5p, miR-139-5p, miR-1468-5p, miR-17-3p, miR-28-3p, miR-361-5p, miR-5006-3p, miR-5010-3p, and miR-532-5p.

Up- and down-regulated miRNAs are expressed in different blood compounds

We asked whether the miRNAs that are up- and down-regulated are expressed to the same amount in different blood cell types, serum or exosomes. To this end we made use of a public miRNA blood cell type atlas [20]. For the up- and down-regulated miRNAs we then compared the average expression in the different compounds and compared them to the background distribution of all human miRNAs (Figure 3B). Interestingly, we observed a highly specific pattern. miRNAs up-regulated in AD were expressed mostly in serum, exosomes, cytotoxic t-cells, and b-cells while those that were down-regulated in AD were expressed in monocytes and t-helper cells. These results suggest a complex regulatory pattern of miRNAs in the different blood cell compounds which would have been likely not observed if only a specific blood cell type or serum would have been investigated.

Machine learning facilitates accurate diagnosis of AD

To obtain more accurate diagnostic results, molecular markers can be considered as "weak learners" that can be combined by machine learning approaches. For our present data set, we explored common statistical and deep learning approaches including support vector machines, decision trees, neural networks and gradient boosted trees and others using five repeated runs of a ten-fold cross validation. While the performance of all approaches was similar (data not shown), the best results were obtained by gradient boosted trees. Compared to other classifiers, gradient boosted trees have the additional advantage that missing values do not have to be imputed. In the classification, two scenarios were modeled: First, the diagnosis of AD patients with unaffected controls (HC) as background group, and second, the diagnosis of AD patients with all controls, *i.e.*, HC, OND, and MCI combined, as background group. In the first and apparently less complex scenario the gradient boosted tree model reached an area under the curve (AUC) of 87.6% (**Figure 4**A). For the second and more complex case, an AUC of 83.5% was reached (Figure 4B). A further advantage of the gradient boosted tree models is that sensitivity and specificity can be well balanced and traded-off. Depending on whether a diagnosis trimmed for sensitivity or for specificity is required *e.g.*, in screening tests, as confirmatory tests or tests for enrollment for clinical studies, a sensitive or a specific model can be chosen.

Feature importance values for each miRNA based on the relative gain obtained via their splits were extracted from both models using the method provided by LightGBM (Table S3) According to this metric, miR-17-3p had the highest importance value in both models, followed by miR-5010-3p. For the model comparing AD to all controls, the next most important miRNAs were let-7d-3p, miR-26b-5p, and miR-28-3p. For the model comparing to unaffected controls, miR-361-5p, let-7d-3p, and miR-532-5p were the next most important features. Interestingly, let-7d-3p and miR-26b-5p were not significantly associated with AD on their own, suggesting that their discriminative power might come from the combination with other miRNAs or their association with different stages of the disease. For example, miR-26b-5p was recently reported to be likely deregulated early in AD, even before the appearance of clinical symptoms [21].

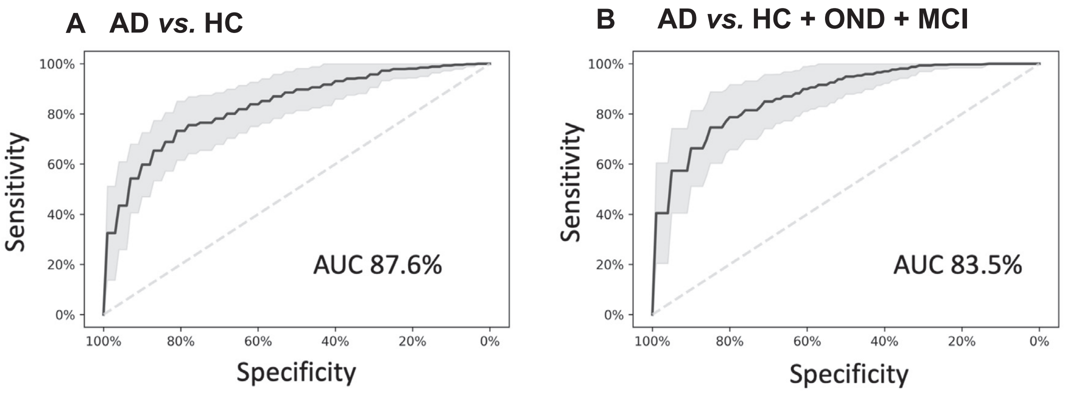
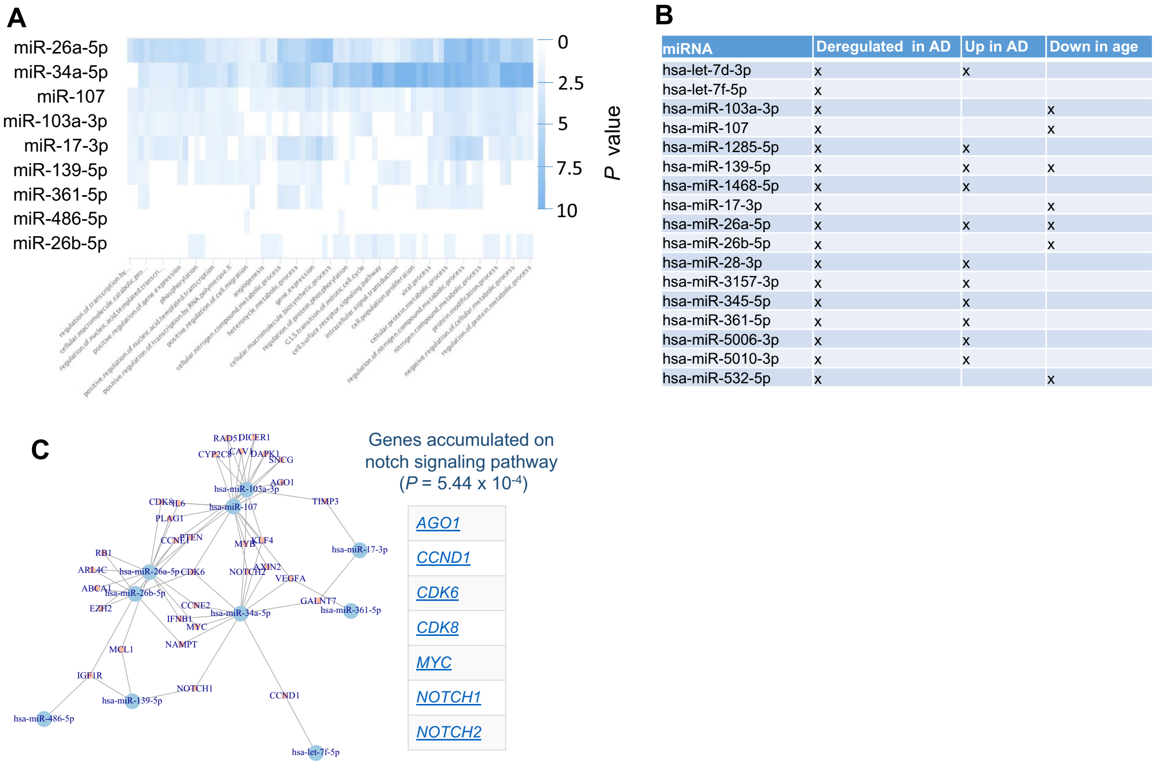


Figure 4 miRNA classifiers show a high diagnostic performance to detect AD
Diagnostic performance of the miRNA classifiers. **A.** ROC AUC for the diagnosis of AD patients compared to HC. **B.** ROC AUC for the diagnosis of AD patients compared to all controls combined (HC, MCI, and OND). The black line indicates the average ROC values of all replicates and folds of the 5 × 10-fold cross-validation models, and the gray area represents the resulting standard deviation. The average AUC obtained over all replicates and folds is displayed for each classification scenario. ROC, receiver operator characteristics; AUC, area under the curve.



miRNA	Deregulated in AD	Up in AD	Down in age
hsa-let-7d-3p	x	x	
hsa-let-7f-5p	x		
hsa-miR-103a-3p	x		x
hsa-miR-107	x		x
hsa-miR-1285-5p	x	x	
hsa-miR-139-5p	x	x	x
hsa-miR-1468-5p	x	x	
hsa-miR-17-3p	x		x
hsa-miR-26a-5p	x	x	x
hsa-miR-26b-5p	x		x
hsa-miR-28-3p	x	x	
hsa-miR-3157-3p	x	x	
hsa-miR-345-5p	x	x	
hsa-miR-361-5p	x	x	
hsa-miR-5006-3p	x	x	
hsa-miR-5010-3p	x	x	
hsa-miR-532-5p	x		x

Figure 5 AD miRNAs regulate distinct pathways and form a dense regulatory core network

A. Heatmap of the miRPathDB results. The heatmap presents the negative decade logarithm of miRNAs and target pathways, and the color represents the significance values. **B.** Overview of miRNAs in significant categories. For the three significant miEAA categories we highlight the miRNAs participating in the respective categories. **C.** miRNA target network from miRTargetLink. From miRTargetLink we extracted the target network of the miRNAs and generated a representation in R using the igraph library. Each node is a miRNA/gene and an edge means that the miRNA targets that gene. As an example of an enrichment of target genes, the genes on the Notch pathway are shown on the right side of the network.

miRNAs are enriched in specific functional categories

To get insights into the targeting of the dysregulated miRNAs, we performed different miRNA target analyses. First, we individually searched for each miRNA those pathways that are enriched with target genes of that miRNA. The result is presented as heat map in **Figure 5**A. Most significant pathways were computed for miR-34a-5p miR-26a-5p followed by miR-107. Among the pathways, many transcription regulated categories have been observed. This result is however to be expected since the main biological function of miRNAs is to regulate the gene expression.

To get more insights, we next performed a miRNA Enrichment analysis [22]. Following adjustment for multiple testing, we identified three categories to be significantly enriched including "Dys-regulation in AD" ($P = 4.8 \times 10^{-8}$), "Up-regulation in AD" ($P = 0.00018$), and "Age" ($P = 0.02$). Two of three categories were directly related to AD. Also this is an expected result for miRNAs that were known to be associated with AD. In addition, these miRNAs are negatively correlated with age. Although this was a weak correlation, it still suggests that the abundances of these miRNAs are lower in older patients. Figure 5B presents for each miRNA in the signature on which categories it has been observed. Performing an enrichment analysis for each of the three miRNAs clusters indicated in Figure 2B, we found cluster A to be especially enriched with miRNAs that are "up-regulated in AD" ($P = 4.9 \times 10^{-6}$) while for cluster B the only significant category was "down-regulated in AD" ($P = 0.04$).

In a third analysis we analyzed all target genes of the miRNAs that had strong evidence in the miRTarBase and were extracted from miRTargetLink. This analysis highlighted that for most miRNAs in our signature, target genes that have been experimentally validated are known. The target network shown in Figure 5C highlighted a dense structure. This network was enriched for genes associated with AD including ABCA1, DAPK1, IGF1R, and VEGFA according to the national institute of aging (NIA). Likewise, "DNA damage response" represented by CCND1, CCNE1, CCNE2, CDK6, MYC, RAD51, and RB1 was over represented. Moreover, the genes in that network were also enriched for the notch signaling pathway.

Discussion

In the current study we present results of our ongoing efforts to develop a diagnostic test for AD patients based on circulating miRNA profiles extracted from blood cells.

As Figure 1 and Table S1 highlight, the samples were largely homogenous with respect to the age and gender distribution. With respect to other characteristics the cohort was however heterogenous (*e.g.*, the origin of the samples from two continents, different diagnostic procedures to identify the patients, potentially different treatment regimens, or a spectrum of patients with higher and lower MMSE values). This heterogeneity helps us to understand whether the deregulation in miRNA patterns between AD patients and controls is of general nature and helps to assess whether *e.g.*, miRNAs are associated with the MMSE state.

The current outcomes are consistent with our previous studies in the US and Germany on smaller cohorts. In contrast to the previous studies relying on deep sequencing, we here applied RT-qPCR as molecular profiling technique that can be more easily driven towards application in clinical care. In the context of the known variability and the bias introduced by sample integrity and sample treatment [23–25] in deep sequencing data, RT-qPCR offers a promising alternative for routine application. But also for RT-qPCR experiments, there is a debate whether RNA samples with low integrity, *i.e.*, low RIN values, compromise miRNA expression data [26,27]. In our study, we also measured RIN as quality criterion for RNA integrity of the samples. The markers that we validated in this study seem to play partially an important role in different diseases. As an example, our most significant marker miR-532-5p is not only correlated and functionally associated to cancer [28–30]. The miRNA and its target network is also associated to sporadic amyotrophic lateral sclerosis [31]. Further, the miR-532-5p has also been discovered in exosomes of multiple sclerosis patients [32] and in exosomes of patients with the geriatric frailty syndrome [33]. Also, our analyses indicate a very essential role of exosome derived miRNAs.

The results of the two cohorts from the US and from Germany were highly concordant. As to be expected by the selection of AD-associated miRNAs for this study, the miRNAs and the target genes of the miRNAs were both significantly associated with the development of AD. Our test that is highly reproducible can be applied with a model based on specificity, sensitivity or trimmed for overall performance. The quality of the results is indicated by an AUC of 87.6% for the comparison between AD and unaffected controls, and an AUC of 83.5% for a comparison between AD and a combined group of unaffected controls, MCI patients and patients with OND. It is known that complex statistical learning approaches can lead to overfitting, especially considering the curse of dimensionality [34] and the fact that usually many more features (p) are measured as compared to the number of patients (n), the $p \gg n$ problem. In our study we however measured $p = 21$ markers and $n = 465$ individuals. Further, we even select small subsets of these markers for our models and perform comprehensive re-sampling to prevent potential overfitting. Although the de-regulation of miRNAs was generally concordant between the GER and the USA cohort, miRNAs have shown differences in the expression level in the two cohorts. This might be due to technical reasons such as shipment, other batch effects or biological differences. Despite this fact, the statistical learning approach succeeded to separate AD and controls in the GER and the USA cohort. In sum, the performance of our diagnostic solution compares well to other recently-developed tests, such as the plasma amyloid marker introduced by Nakamura and co-workers [4]. While already such single "omics" tests have a large potential, the targeted combination of few representatives from different "omics" classes can add even more diagnostic information, supporting clinicians in detecting AD patients in time. One challenge of respective studies is that the clinical diagnosis may be imperfect. The MCI patients that are an important second control group besides the unaffected controls may have already early forms of AD that are not yet detected with the current diagnostic means.

A pathway based analysis of miRNAs and target genes indicated a functional role of the miRNAs. This is further supported by a different blood compound distribution of those miRNAs that are up- and down-regulated in AD. Respective pathway analyses have however always considered with caution, especially when small sets of miRNAs are considered. Although the results of the analysis seem to be reasonable, a potential bias is hard to be excluded. *e.g.*, we picked already miRNAs known from literature to be associated with AD. An enrichment of AD related miRNAs itself is thus an expected result. Similarly, also the target gene analyses might be biased for miRNAs and target genes that are in the focus of many research groups.

As for other omics types, confounders including age and gender potentially influence also the results of miRNA biomarker studies [35]. To minimize the influence of such confounders, our cohorts largely show similar age and gender distribution (Table 1). In addition, we investigated the influence of the age and gender on the miRNA profiles. Except for a very modest influence of age, we found no evidence for an influence of these confounders on the miRNA pattern. Notably, miRNAs that are down-regulated in AD were partially expected to be lower expressed with increasing age in a normal population. Among the many different candidates for minimally-invasive and potentially early stage tests for AD, our study indicates that circulating miRNAs likely in combination with other blood-born omics profiles will contribute to stable tests applicable to specific diagnostic questions with regard to this highly complex disease.

Materials and methods

Overview of the study

In the current study we included patients from the US [8] and Germany [16] that were partially collected within the longitudinal Tübinger Erhebung von Risikofaktoren zur Erkennung von Neurodegeneration (TREND) study. From the former studies we included those individuals, where a sufficient amount of high-quality RNA was left for analysis. In detail, 169 individuals from our initial study (36%) [8], 107 individuals from the second study (23%) [16], as well as 189 newly collected individuals (41%) were included in the study. The studies were approved by the institutional review boards of Charité – Universitätsmedizin Berlin (EA1/182/10), or the ethical committee of the Medical Faculty of the University of Tuebingen (Nr. 90/2009BO2). All subjects gave written informed consent. Besides AD patients and HC, patients with OND such as Parkinson's disease (PD), schizophrenia or bipolar disorder were included and grouped together, termed OND. Further, patients with MCI were included to evaluate the specificity of the miRNA markers for AD. For each of the four groups and separately for the USA and GER cohorts, total number, age, gender distribution, and MMSE value are presented in Table 1. Moreover, from one individual, 12 technical replicates were measured continuously during the project as process control.

miRNA marker set selection

From our two previous studies [8,16] we selected the top miRNAs that were concordant between the two studies, and also checked for evidence that the miRNAs are associated with AD in literature. A final set of 21 miRNAs was selected. These are listed in Supplemental Table 4 where additional selection criteria are provided. In more detail, 17 miRNAs were significantly associated with AD in our first study, 14 miRNAs were significant in our second study. miR-34a-5p was not detected in our previous studies by NGS but in a study by Cosin-Tomas [36]. Further, this miRNA is one of our main targets regulating calcium signaling, NFKappaB pathway and T-cell killing and is down-regulated significantly in aging [37,38]. miR-151-3p is one of the most stable miRNAs in our studies as well as miR-486-5p, which is a red blood cell miRNA that serves as positive control [20].

RNA extraction and quality control

Total RNA from PAX-Gene Blood Tubes (Catalog No. 762165, BD Biosciences, Franklin Lakes, NJ) was isolated using the Qiacube robot with the PAXgene Blood miRNA Kit (Catalog No. 763134, Qiagen, Hilden, Germany) according to manufacturer's instructions. In the tubes, 2.5 ml blood are collected, typically yielding around 1 mg total RNA. RNA quantity and quality were assessed using Nanodrop (Thermo Fisher Scientific) and RNA Nano 6000 Bioanalyzer Kit (Catalog No. 5067-1511, Agilent Technologies, Santa Clara, CA). Mean RNA integrity number (RIN) value of the RNA samples was 7.5 (STDEV 1.4).

RT-qPCR

Quantification of miRNAs was performed using miScript PCR system and custom miRNA PCR arrays (all reagents from Qiagen, Hilden, Germany). Custom miRNA PCR arrays were designed in 96-well plates to measure the expression of 21 human miRNAs and RNU48 as well as RNU6 as two endogenous controls in duplicates. Two process controls (miR-TC for RT efficiency, PPC for PCR efficiency) were included as single probes. A total of 100 ng total RNA was used as input for reverse transcription reaction using miScriptRT-II kit according to manufacturer's recommendations in 20 μl total volume (Catalog No. 218161). Subsequently, 1 ng cDNA was used per PCR reaction. PCR reactions with a total volume of 20 μl were setup automatically using the miScript SYBR Green PCR system (Catalog No. 218076) in a Qiagility pipetting robot (Qiagen, Hilden, Germany) according to manufacturer's instructions. Data from samples that failed the quality criteria for the process controls was excluded, leaving expression data from 465 samples available for analysis. For process control over the course of the project, eleven technical replicates of one cDNA sample were measured throughout the course of the project to estimate technical reproducibility. We computed 55 pair-wise correlation coefficients between any pair of the replicates and found a median correlation of 0.996, indicating high technical reproducibility of our assay.

Statistical approaches

From the Cq values, delta Cq values in relation to the endogenous control (RNU48) were computed. Mean delta Cq value per individual was scaled to zero. Missing values were not imputed. As estimate of the expression on a linear scale, 2^{deltaCq} values were computed. For multi group comparisons, Analysis of Variance (ANOVA) was performed. Since the miRNA data and partially the response variable were not always normally distributed according to Shapiro Wilk tests, we performed for the pair-wise comparisons and for the correlation analysis parametric as well as non-parametric tests. For pair-wise comparisons, both, parametric t-test and non-parametric Wilcoxon Mann-Whitney test were calculated. If not mentioned explicitly and where applicable, all P values were adjusted for multiple testing by the Benjamini-Hochberg approach. For correlating miRNAs to the age and the MMSE value, the P value was computed based on parametric Pearson's product moment correlation coefficient as well as non-parametric Spearman Correlation. To find enrichment of miRNAs in specific blood compounds we used data of an NGS based blood cell miRNA repertoire [20]. Each miRNA was normalized to 100% and the different expression ratios in the different blood compounds were compared to each other.

miRNA target analysis

We performed three different approaches on miRNA target analysis. First, for each single miRNA the target pathways have been extracted from miRPathDB [39] and the CustomHeatmap tool was used to find miRNAs that target at least 5 pathways and pathways targeted by at least 5 miRNAs from biological GO processes. Next, we performed a so-called miRNA set enrichment analysis relying on the hypergeometric distribution using MIEAA [22]. Here, the miRNAs in the study were compared to the background distribution of all miRNAs and the procedure was repeated for the dysregulated miRNAs. All pathways with an adjusted P value below 0.05 were considered to be significant. Finally, we used the miRTargetLink tool [40] to extract the experimentally validated targets of the miRNAs. In this analysis only the strong target category from miRTarBase has been used to obtain specific results. From that data we computed a network using the R igraph package and performed an enrichment analysis of the target genes in that network.

Machine learning

A prediction model based on the RT-qPCR Cq values was developed using gradient boosted trees from the LightGBM framework (version 2.1.0). Since not all miRNAs were consistently measured for all patients, tree-based methods are particularly suited for this task, as they can handle missing values and no imputation is required. LightGBM ignores the missing values when computing the splits of the trees and assigns all samples with missing values to the side that reduces the loss most. The performance of the model was assessed using five repetitions of stratified ten-fold cross-validation using scikit-learn 0.19.1 with Python 3.6.4 [41]. Each repetition was initiated with an integer seed (0–4). Thus, in total 50 combinations of different training and validation sets were considered. The reported ROC AUC corresponds to the average performance over all repetitions and folds of the model, on data not used for training. The models were manually tuned (*i.e.*, no grid search was performed) over the number of leaves (testing ranges between 5 and 50), number of estimators (between 40 and 120), learning rate (0.01 to 0.2), and depth (3 to no restriction). The final model comparing patients with AD to all controls uses 30 leaves, a learning rate of 0.1 and 100 estimators. The model comparing patients with AD to unaffected controls uses 9 leaves, a learning rate of 0.05 and 100 estimators. The depth of both models was not restricted. Gradient boosted trees outperformed other tree-based methods such as random forests, or classifiers as Support Vector Machines or Neural Networks (data not shown). As an input for the classification task, the expression matrix of the delta Cq values has been used.

Data availability

The full data set is available as Table S1 without any restrictions.

Authors' contributions

NL measured the samples and supported the interpretation of data; TF and FK interpreted and analyzed the data; MG, WM, CS, AKvT, CD, FM, US, and DB supported the study conceptionally, added to the study protocol and enrolled patients; VK added to the clinical interpretation of the data; CB supported the data analysis and contributed to drafting the manuscript; SD and SG supported to measure and interpret the data; HPL, EM, AK were the PIs of the study, contributed to writing and correcting the manuscript and to data analysis and interpretation.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by the Alzheimer Forschungs Initiative (AFI) (Grant No. AFI-Grant# 15013). This study was also supported by internal funds of Saarland University. We acknowledge support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) and Saarland University within the funding programme Open Access Publishing.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2019.09.004.

References

[1] Querfurth HW, LaFerla FM. Alzheimer's disease. N Engl J Med 2010;362:329–44.

[2] Weuve J, Hebert LE, Scherr PA, Evans DA. Deaths in the United States among persons with Alzheimer's disease (2010–2050). Alzheimers Dement 2014;10:e40–6.

[3] Murphy MP. Amyloid-Beta solubility in the treatment of Alzheimer's disease. N Engl J Med 2018;378:391–2.

[4] Nakamura A, Kaneko N, Villemagne VL, Kato T, Doecke J, Dore V, et al. High performance plasma amyloid-beta biomarkers for Alzheimer's disease. Nature 2018;554:249–54.

[5] Mapstone M, Cheema AK, Fiandaca MS, Zhong X, Mhyre TR, MacArthur LH, et al. Plasma phospholipids identify antecedent memory impairment in older adults. Nat Med 2014;20:415–8.

[6] Lunnon K, Sattlecker M, Furney SJ, Coppola G, Simmons A, Proitsi P, et al. A blood gene expression marker of early Alzheimer's disease. J Alzheimers Dis 2013;33:737–53.

[7] Fransquet PD, Lacaze P, Saffery R, McNeil J, Woods R, Ryan J. Blood DNA methylation as a potential biomarker of dementia: a systematic review. Alzheimers Dement 2018;14:81–103.

[8] Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, et al. A blood based 12-miRNA signature of Alzheimer disease patients. Genome Biol 2013;14:R78.

[9] Casanova R, Varma S, Simpson B, Kim M, An Y, Saldana S, et al. Blood metabolite markers of preclinical Alzheimer's disease in two longitudinally followed cohorts of older individuals. Alzheimers Dement 2016;12:815–22.

[10] Pichler S, Gu W, Hartl D, Gasparoni G, Leidinger P, Keller A, et al. The miRNome of Alzheimer's disease: consistent downregulation of the miR-132/212 cluster. Neurobiol Aging 2017;50: e1–10.

[11] Ren RJ, Zhang YF, Dammer EB, Zhou Y, Wang LL, Liu XH, et al. Peripheral blood microRNA expression profiles in Alzheimer's disease: screening, validation, association with clinical phenotype and implications for molecular mechanism. Mol Neurobiol 2016;53:5772–81.

[12] Denk J, Oberhauser F, Kornhuber J, Wiltfang J, Fassbender K, Schroeter ML, et al. Specific serum and CSF microRNA profiles distinguish sporadic behavioural variant of frontotemporal dementia compared with Alzheimer patients and cognitively healthy controls. PLoS One 2018;13:e0197329.

[13] Yang TT, Liu CG, Gao SC, Zhang Y, Wang PC. The serum exosome derived microRNA-135a, -193b, and -384 were potential Alzheimer's disease biomarkers. Biomed Environ Sci 2018;31:87–96.

[14] Hu YB, Li CB, Song N, Zou Y, Chen SD, Ren RJ, et al. Diagnostic value of microRNA for Alzheimer's disease: a systematic review and meta-analysis. Front Aging Neurosci 2016;8:13.

[15] Nagaraj S, Zoltowska KM, Laskowska-Kaszub K, Wojda U. microRNA diagnostic panel for Alzheimer's disease and epigenetic trade-off between neurodegeneration and cancer. Ageing Res Rev 2019;49:125–43.

[16] Keller A, Backes C, Haas J, Leidinger P, Maetzler W, Deuschle C, et al. Validating Alzheimer's disease micro RNAs using next-generation sequencing. Alzheimers Dement 2016;12:565–76.

[17] Leidinger P, Brefort T, Backes C, Krapp M, Galata V, Beier M, et al. High-throughput qRT-PCR validation of blood microRNAs in non-small cell lung cancer. Oncotarget 2016;7:4611–23.

[18] Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T, Pallasch C, et al. Distribution of miRNA expression across human tissues. Nucleic Acids Res 2016;44:3865–77.

[19] Keller A, Leidinger P, Bauer A, Elsharawy A, Haas J, Backes C, et al. Toward the blood-borne miRNome of human diseases. Nat Methods 2011;8:841–3.

[20] Juzenas S, Venkatesh G, Hubenthal M, Hoeppner MP, Du ZG, Paulsen M, et al. A comprehensive, cell specific microRNA catalogue of human peripheral blood. Nucleic Acids Res 2017;45:9290–301.

[21] Swarbrick S, Wragg N, Ghosh S, Stolzing A. Systematic review of miRNA as biomarkers in Alzheimer's disease. Mol Neurobiol 2019;56:6156–67.

[22] Backes C, Khaleeq QT, Meese E, Keller A. miEAA: microRNA enrichment analysis and annotation. Nucleic Acids Res 2016;44: W110–6.

[23] Backes C, Sedaghat-Hamedani F, Frese K, Hart M, Ludwig N, Meder B, et al. Bias in high-throughput analysis of miRNAs and implications for biomarker studies. Anal Chem 2016;88:2088–95.

[24] Ludwig N, Becker M, Schumann T, Speer T, Fehlmann T, Keller A, et al. Bias in recent miRBase annotations potentially associated with RNA quality issues. Sci Rep 2017;7:5162.

[25] Backes C, Leidinger P, Altmann G, Wuerstle M, Meder B, Galata V, et al. Influence of next-generation sequencing and storage conditions on miRNA patterns generated from PAXgene blood. Anal Chem 2015;87:8910–6.

[26] Becker C, Hammerle-Fickinger A, Riedmaier I, Pfaffl MW. mRNA and microRNA quality control for RT-qPCR analysis. Methods 2010;50:237–43.

[27] Jung M, Schaefer A, Steiner I, Kempkensteffen C, Stephan C, Erbersdobler A, et al. Robust microRNA stability in degraded RNA preparations from human tissue and cell samples. Clin Chem 2010;56:998–1006.

[28] Yamada Y, Arai T, Kato M, Kojima S, Sakamoto S, Komiya A, et al. Role of pre-miR-532 (miR-532-5p and miR-532-3p) in regulation of gene expression and molecular pathogenesis in renal cell carcinoma. Am J Clin Exp Urol 2019;7:11–30.

[29] Xie X, Pan J, Han X, Chen W. Downregulation of microRNA-532-5p promotes the proliferation and invasion of bladder cancer cells through promotion of *HMGB3*/Wnt/beta-catenin signaling. Chem Biol Interact 2019;300:73–81.

[30] Wei H, Tang QL, Zhang K, Sun JJ, Ding RF. miR-532-5p is a prognostic marker and suppresses cells proliferation and invasion by targeting *TWIST1* in epithelial ovarian cancer. Eur Rev Med Pharmacol Sci 2018;22:5842–50.

[31] Liguori M, Nuzziello N, Introna A, Consiglio A, Licciulli F, D'Errico E, et al. Dysregulation of microRNAs and target genes networks in peripheral nlood of patients with sporadic Amyotrophic Lateral Sclerosis. Front Mol Neurosci 2018;11:288.

[32] Selmaj I, Cichalewska M, Namiecinska M, Galazka G, Horzelski W, Selmaj KW, et al. Global exosome transcriptome profiling reveals biomarkers for multiple sclerosis. Ann Neurol 2017;81:703–17.

[33] Ipson BR, Fletcher MB, Espinoza SE, Fisher AL. Identifying exosome-derived microRNAs as candidate biomarkers of Frailty. J Frailty Aging 2018;7:100–3.

[34] Barbour DL. Precision medicine and the cursed dimensions. NPJ Digit Med 2019;2:4.

[35] Meder B, Backes C, Haas J, Leidinger P, Stahler C, Grossmann T, et al. Influence of the confounding factors age and sex on microRNA profiles from peripheral blood. Clin Chem 2014;60:1200–8.

[36] Cosin-Tomas M, Antonell A, Llado A, Alcolea D, Fortea J, Ezquerra M, et al. Plasma miR-34a-5p and miR-545-3p as early biomarkers of alzheimer's disease: potential and limitations. Mol Neurobiol 2017;54:5550–62.

[37] Hart M, Walch-Ruckheim B, Krammes L, Kehl T, Rheinheimer S, Tanzer T, et al. miR-34a as hub of T cell regulation networks. J Immunother Cancer 2019;7:187.

[38] Hart M, Walch-Ruckheim B, Friedmann KS, Rheinheimer S, Tanzer T, Glombitza B, et al. miR-34a: a new player in the regulation of T cell function by modulation of NF-kappaB signaling. Cell Death Dis 2019;10:46.

[39] Backes C, Kehl T, Stockel D, Fehlmann T, Schneider L, Meese E, et al. miRPathDB: a new dictionary on microRNAs and target pathways. Nucleic Acids Res 2017;45:D90–6.

[40] Hamberg M, Backes C, Fehlmann T, Hart M, Meder B, Meese E, et al. MiRTargetLink–miRNAs, genes and interaction networks. Int J Mol Sci 2016;17:564.

[41] Pedregosa F, Varoquaux G, Gramfort A, Michel V, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.

3.4 *miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems*

Published online 6 May 2020 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* ***W521–W528***
doi: 10.1093/nar/gkaa309

miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems

Fabian Kern 1,†, Tobias Fehlmann 1,†, Jeffrey Solomon1, Louisa Schwed1, Nadja Grammes 1, Christina Backes 1, Kendall Van Keuren-Jensen2, David Wesley Craig 3, Eckart Meese 4 and Andreas Keller 1,5,6,*

1Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, 2Neurogenomics Division, Translational Genomics Research Institute, Phoenix, AZ 85004, USA, 3Institute of Translational Genomics, University of Southern California, Los Angeles, CA 90033, USA, 4Department of Human Genetics, Saarland University, 66421 Homburg, Germany, 5School of Medicine Office, Stanford University, Stanford, CA 94305, USA and 6Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA 94304, USA

Received March 05, 2020; Revised April 06, 2020; Editorial Decision April 17, 2020; Accepted April 22, 2020

ABSTRACT

Gene set enrichment analysis has become one of the most frequently used applications in molecular biology research. Originally developed for gene sets, the same statistical principles are now available for all omics types. In 2016, we published the miRNA enrichment analysis and annotation tool (miEAA) for human precursor and mature miRNAs. Here, we present miEAA 2.0, supporting miRNA input from ten frequently investigated organisms. To facilitate inclusion of miEAA in workflow systems, we implemented an Application Programming Interface (API). Users can perform miRNA set enrichment analysis using either the web-interface, a dedicated Python package, or custom remote clients. Moreover, the number of category sets was raised by an order of magnitude. We implemented novel categories like annotation confidence level or localisation in biological compartments. In combination with the miRBase miRNA-version and miRNA-to-precursor converters, miEAA supports research settings where older releases of miRBase are in use. The web server also offers novel comprehensive visualizations such as heatmaps and running sum curves with background distributions. We demonstrate the new features with case studies for human kidney cancer, a biomarker study on Parkinson's disease from the PPMI cohort, and a mouse model for breast cancer. The tool is freely accessible at: https://www.ccb.uni-saarland.de/mieaa2.

INTRODUCTION

Transcriptomics designates an indispensable set of techniques to study gene expression, often in a genome-wide manner, as the backbone of modern molecular biology and clinical research. The innumerable amount of classical bulk-sequencing datasets is further augmented by the recent advancements in high-resolution single-cell approaches. Since gene expression is constituted by many biological factors, experimental focus has been enlarged to include the regulatory non-coding transcriptome (ncRNAs), i.e. to RNA classes that regulate messenger RNAs (mRNAs) either directly or indirectly. Among these, microRNAs (miRNAs) are small non-coding RNAs, typically 18-25 nucleotides in length, loaded into proteins of the AGO-family to build RNA-induced silencing complexes (RISC) (1). Gene regulation through the RISC complex is facilitated by one or two mature ($-5p$; $-3p$) miRNA arms, arising from one or several transcribed precursors (2). Besides other modes of action, activated complexes target preferentially 3′-untranslated regions of mRNAs to induce either catalytic cleavage or translation repression. Hence, profiling miRNA expression contributes to the understanding of gene regulation and potentially portrays cellular states. To date, numerous studies highlight their informative role in disease detection, sub-type classification, or progression, such as for cancer (3), neurodegenerative (4), or metabolic disorders (5) with a variety of bio-specimens (6).

Considering that several thousands of miRNAs have already been discovered, many novel miRNA candidates have been additionally proposed (7), while the total number of human miRNAs is estimated to be 2300 (8). Finding differences in expression for miRNAs is similar to mRNAs

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

and therefore non-trivial. Differential gene expression studies often lead to dozens, hundreds, or even thousands of deregulated genes. Thus, large scale studies often make use of the functionality of gene set enrichment analysis (GSEA) (9). GSEA can further reduce large amounts of information towards a significant set of molecular functions, biological properties, or pathways of genes. In principle, a user inputs either a set or ordered list of genes and the tool runs the required statistical algorithms and provides background datasets to compare against.

Similar functionality was also implemented for other omics types, including proteomics, metagenomics or epigenomics. An in-depth review of gene set analysis methods for data other than mRNAs demonstrates the increasing interest and demand of the community in respective tools (10). We previously developed a statistical approach tailored for both miRNA precursor and mature miRNA input, the miRNA enrichment analysis and annotation tool (miEAA) (11). Here, we present an update of this tool that includes more categories, supports nine additional species, has new statistical functionality and offers a standardised Application Programming Interface (API) to facilitate the inclusion in modern data analysis workflows (12).

Given the growing interest in miRNAs, other tools with similar functionality to miEAA exist. The pioneering tool providing functionality for miRNA enrichment was TAM (13), which covers in it's latest version 2.0 (14) as many as 1238 human miRNA categories obtained from manual literature review of ∼9000 scientific manuscripts, along with new query and visualization features. In addition to the over- and under-representation analysis, users can compare the correlation of two miRNA lists under different disease conditions. Another important tool with similar functionality is miSEA (microRNA Set Enrichment Analysis) (15). It facilitates the selection of a large set of microRNA categories, including family classification, disease association, and genome coordinate. Furthermore, custom miRNA sets can be defined by the user. All kinds of enrichment tools rely on high quality sets of miRNA categories that were either obtained by curation of scientific literature or collected from specific databases. For instance, curated miRNA annotations can be obtained from miRBase (16) or miRCarta (17), miRNA–target interactions from miRTarBase (18), miRNA–pathway associations from miRPathDB (19), tissue-specific miRNAs from the human TissueAtlas (20), or miRNA-disease associations from HMDD (21) or MNDR (22), many of which were updated in the last two years. Further specialized annotations like miRNA and transcription factor interactions from TransmiR (23), miRNA sub-cellular localisations collected in RNALocate (24), or extra-cellular circulating miRNAs contained in miRandola (25) provide target categories for comprehensive enrichment analysis.

MATERIALS AND METHODS

In miEAA 2.0, we provide support for ten species whereas the first release of miEAA only supported *Homo sapiens*, 31 new category sets, and updates to our pre-existing datasets. To unify data preprocessing, we implemented an automated pipeline using Snakemake (26), Python 3.6, and the pandas (27) Python package facilitating data collection and filtering steps. For each species and their corresponding data sources our pipeline performs the same basic process, consisting of downloading the datasets, cleaning and updating the miRNA and precursor identifiers, transforming the results into a Gene Matrix Transposed (GMT) file, and creating background reference sets. Files were copied to the web server without further modification.

Data collection

Novel datasets were obtained to build our enrichment categories, consisting of Gene Ontology (28), miRTarBase 8.0 (18), KEGG (29), miRandola 2017 (25), miRPathDB 2.0 (19), TissueAtlas (20), MNDR v2.0 (22), NPInter 4.0 (30), RNALocate v2.0 (24), SM2miR (31), TAM 2.0 (14) and TransmiR v2.0 (23). Further annotations for cell-type and tissue specific expression of miRNAs and precursors were derived from three dedicated atlas publications (32,33) (10.1101/430561). Other pre-existing datasets have been updated, including HMDD v3.0 (21) and miRBase v22.1 (16). We retained the rest of our pre-existing datasets, namely miRWalk2.0 (34), published age and gender dependent miRNAs and distribution of miRNAs in immune cells (11). Most of the datasets contain miRNAs or precursors for *H. sapiens*. When available, we also utilise the data to derive categories representing the non-human organisms. Raw datasets were obtained either through a direct download or via an API. In particular, the QuickGO and KEGG datasets are compiled by querying corresponding REST APIs.

Category data preprocessing

First, data from QuickGO was mapped back to miRBase using RNAcentral (35). NCBI Gene was used in conjunction with miRTarBase to produce the indirect annotations. With the aid of the miRBaseConverter R package (36), miRNA and precursor names were translated to the latest version of miRBase. For KEGG Pathways and GO Annotations (direct and indirect through target genes from miRTarBase) we only keep miRNAs for which functional MTI support is available. In the MNDR diseases category set, we exclude HMDD data as it is precursor based, and MNDR is for mature miRNAs. To determine tissue-specific expression we computed the tissue specificity index (20) and applied a threshold filtering at 0.75.

Web server, statistics, and API implementation

The miEAA web server was built using a dockerized Django Web Framework v2.1, which exposes a web-API using the Django REST framework. The celery software was used as the job scheduler. Frontend libraries comprise Highcharts, dataTables, jquery, and Bootstrap. *P*-value correction methods were implemented using the R stats package. As gene set enrichment analysis (GSEA) implementation we provide an un-weighted variant of the algorithm. This implies the amount by which the running sum is changed in each step is constant, corresponding to a Kolmogorow–Smirnow test. This approach enables to compute the exact *P*-value without requiring permutations of either the case / control labels, or the miRNA lists (37). As an exception, the static GSEA running sum plots are computed by randomly permuting the test set 100 times and traversing the running sum for each random permutation. If the absolute maximal deviation from zero is positive, miEAA assumes an enrichment on top of the ordered list and results are shown in red colour to denote an enrichment. If the absolute maximal deviation from zero is negative, miEAA assumes an enrichment at the end of the ordered list and results are displayed in green color to denote an inverse enrichment, i.e. a depletion. Alongside our new API we provide a lightweight Python package, as well as a command line interface (CLI) tool, supporting Python 3.5 or higher. These are made freely available through the Python Package Index (pip) and through the *ccb-sb* conda channel. The already existing miRNA to precursor and miRBase converters were upgraded to miRBase v22.1. The former offers new output modes to simplify the review of ambiguous conversion results and proper down-stream usage.

Case studies

Raw and reads per million miRNA mapping (rpmmm) normalized miRBase v21 precursor counts and metadata of kidney renal clear cell carcinoma case and control samples were obtained from The Cancer Genome Atlas (TCGA). Since multiple sequencing results might be associated with the same sample ID in TCGA, we kept only one result file for each sample by preferring files from H over R over T analytes and selecting the aliquot with the highest plate number and / or lexicographical sorting order. Subsequently, miRNAs with fewer than 5 raw reads in less than 50% of either case or control samples were discarded from the analysis. All remaining miRNA counts were $\log_2$-scaled. Effect size was calculated using the implementation of Cohen's d from the R package effsize. Lists of precursor names, either selected by statistical significance or ordered by effect size, were converted from miRBase v21 to v22.1 using the online miRBase converter feature of miEAA. The list of all precursors from miRBase v21, converted to v22.1, were used as a reference set. The configured parameters included default precursor category sets without the *PubMed ID* and *TransMiR Tissues* sets, BH-FDR adjustment to a significance level of 0.05 with independently adjusted *P*-values per category set, and a minimum of 2 required hits per subcategory.

For the second case study, raw Agilent microarray data and sample metadata was downloaded from NCBI's GEO using accession ID GSE117000. Array parsing and probe signal processing was performed identically to the description in the first publication of miEAA (11). Subsequently, all counts were quantile-normalized and $\log_2$-transformed. All further down-stream analyses were performed analogous to the first case-study described above.

To provide a non-cancer case study we evaluated the performance of miEAA on a high-resolution dataset of small non-coding RNAs in whole blood (38). This dataset is freely available from the Parkinson's Progression Markers Initiative (PPMI) data portal. In summary, for 1600 individuals up to five blood samples from a time frame of over three years were acquired and sequenced for sncRNAs. We quantified all human miRBase v22 precursors from the 4340 sequencing samples. Raw counts were normalized to reads per million (rpm) and precursors were filtered analogously to the criteria defined for the TCGA case study. Next, we compared the miRNA precursor profiles of 2337 Parkinson's samples to 1538 age-matched controls. For this case study we also mapped back the precursors to miRBase v21 to perform a detailed comparison of enrichment results to TAM 2.0.

RESULTS

Overview on miEAA 2.0

In the following, changes and novelties introduced by the second major release of miEAA are described. Since all annotations of miRNAs to categories and databases are with respect to the miRNA reference database, miRBase, we converted the datasets to match its latest public version 22.1. This also affects the miRBase-version and miRNA-to-precursor converters, the former of which was designed to be fully backwards compatible. Moreover, both ORA and GSEA algorithms accept lists of either precursors or miRNAs, from *H. sapiens*, *Mus musculus*, *Rattus norvegicus*, *Arabidopsis thaliana*, *Bos taurus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Gallus gallus* and *Sus scrofa*. In total, 134 525 categories from 16 published databases/resources are available to test against. A detailed breakdown of the counts by source and organism, on database and category set level, are available from Supplementary Table S1 and S2, respectively. For the precursor annotations, we curated family assignments, re-computed genomic clusters of miRNA genes, updated the chromosomal locations for human, and added all similar categories for other species. We also updated the category set representing PubMed IDs of manuscripts that contributed miRNA entries to miRBase. This feature has both, a biological and technical aspect. From the technical view, miRNAs could have been reported by the original paper due to experimental bias. In case a new input query is enriched for respective miRNAs it could be due to the same kind of bias. From a biological perspective, a study might have found miRNAs in the context of a disease. If such a manuscript is identified in a similar context in miEAA, additional evidence for the validity can be inferred. All species except *A. thaliana* are annotated with a new category listing high confidence precursors according to miRBase criteria. For human data, we transferred the disease annotations from HMDD to the new major release v3. We added associations from MNDR to allow disease comparisons against HMDD, and incorporated functional RNA interactions from NPInter. Lastly, novel categories such as the cellular localisation of miRNAs and regulatory interactions between miRNAs and transcription factors were incorporated from RNALocate and TransmiR, respectively. For the mature miRNAs, comparable changes apply as for the precursors in the cases of miRBase, MNDR, NPInter, and RNALocate-derived category sets. The gap between annotations of miRNA properties and their function is filled by categories on target genes taken from miRTarBase. Moreover, known miRNA to drug assocations are provided from SM2miR. To facilitate target-based enrichment of molecular pathways or biological function, we computed enrichments on target genes of miRNAs using Gene Ontology and KEGG. As an alternative for end-users, pre-computed significant enrichments of miRNAs associated with pathways provided by miRPathDB were made available for analysis. As the data from miRPathDB already involves a statistical pre-filtering, we implemented a new list of expert categories to highlight the underlying differences. Manually curated classifications from miRandola about known circular or extracellular miRNAs are also integrated. Finally, new annotations for cell-type and tissue-specific precursors and miRNAs have been integrated. Supposedly, the substantially enlarged number of categories might increase the average runtime of our algorithms, especially for the computationally intensive GSEA. Therefore, we profiled and improved our GeneTrail-based implementation to be three times faster, on average (39).

We raised the available number of statistical parameter settings as well. First, users can request unadjusted or adjusted *P*-values using six published techniques to account for multiple hypothesis testing on the same dataset. In addition to the classical Bonferroni and Benjamini–Hochberg False discovery rate (BH-FDR) procedures, the adjustments proposed by Benjamini-Yekuteli, Hochberg, Holm and Hommel can be selected. Moreover, the default behavior of miEAA to correct *P*-values database / category setwise was extended by a *P*-value pooling approach. In summary, the well-established alternatives for *P*-value correction can support highly customized research setups where alternate levels of stringency are required (40).

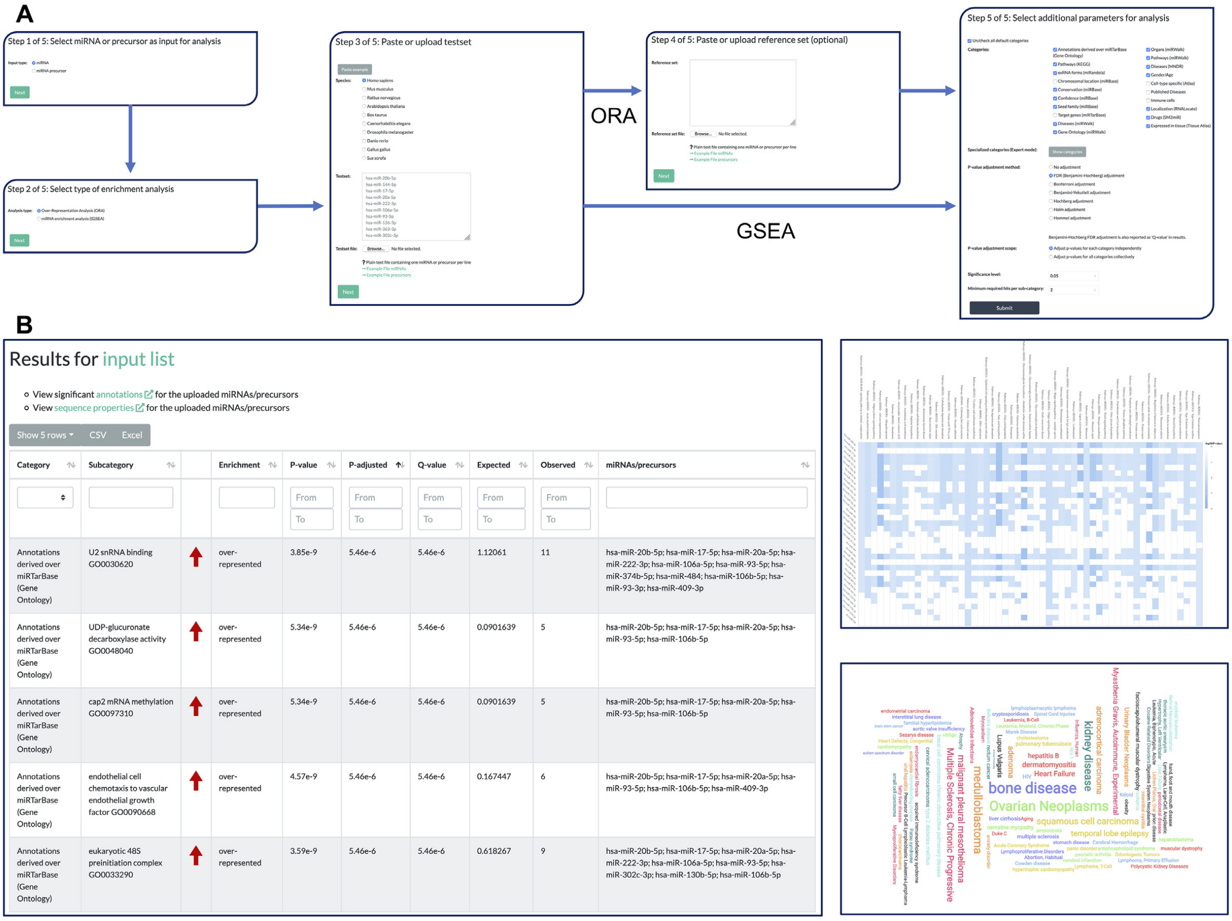
We also evaluated new visualization features for the output of enrichment analyses to provide a simple overview and to improve comprehension. As a result, we made existing graphs interactive and implemented enrichment graphs with simulated background distributions for GSEA as well as automatic word cloud and heatmap plots for all enrichment algorithms. Word clouds display the names of obtained categories while scaling the size of the terms relatively to the number of hits that occurred (on a linear or logarithmic scale) and allow one to qualitatively compare the categories. On top of that, category to miRNA heatmaps depict log-transformed *P*-values for the hits obtained. This feature permits to compare the similarity of enriched / depleted categories with respect to associated miRNAs or precursors in a simple fashion. The workflow of miEAA and example visualizations are displayed in Figure 1. Finally, we enhanced the general accessibility of miEAA through the implementation of a public API and a Python package, for which more details are provided below.

Case study 1: Human kidney renal clear cell carcinoma

As the first case-study of miEAA 2.0, we acquired 591 human miRNA-seq samples from the kidney renal clear cell carcinoma (KIRC) project of TCGA, which can be divided into 520 Primary tumor (PT) and 71 Solid tissue normal (STN) samples. Sample information can be found in Supplementary Table S3. Of the 1881 precursors from miRBase v21, 321 are consistently detected in at least 50% of the samples for each biogroup. Among these, 282 were differentially expressed between PT and STN according to the FDR-adjusted wilcoxon test *P*-values ($P < 0.01$). Overrepresentation analysis of the precursors resulted in 541 significantly enriched and seven significantly depleted (FDR-adjusted; $P < 0.05$) categories. As shown in Figure 2A, a subset of precursors is ubiquitously present in significant categories, while others seem to be more specific. The top 10 categories sorted by increasing *P*-value are associated with cancer, including renal cell carcinoma. Also, the observed over expected ratio (123/48.6) indicates a strong enrichment ($P = 2.80 \times 10^{-38}$) of the de-regulated precursors with kidney and other types of cancer. A miRNA set enrichment analysis, using the list of detected precursors and sorted by effect size, revealed 253 enriched and 40 depleted categories. Here, the miRNA gene cluster 147, 189, 704 : 147, 284, 728 on the X chromosome is the most depleted category ($P = 8.64 \times 10^{-10}$), an observation that is in line with the depletion of precursor family hsa-mir-506. Interestingly, the list of highly enriched terms contains many transcription factors, the top 5 being *HEY1*, *WDR5*, *ELF1*, *BRD4* and *FLI1*.

Case study 2: mouse model for breast cancer progression

To showcase the novel support for model organisms in miEAA, we selected a dataset from GEO where circulating miRNAs from a breast-cancer mouse model were measured with microarrays (41). The dataset comprises 36 samples from mutation-carrier (NeuT+) and age-matched wildtype (NeuT–) mice that were collected at the premalignant, preinvasive and invasive stages of the disease. In this particular study, agilent microarrays probed with miRNAs from miRBase v19 were used on mice's plasma extracted RNA samples. Sample information can be found in Supplementary Table S4. Following a detection threshold procedure similar to our first case study, 212 miRNAs remained for differential expression analysis. Of these, mmu-miR-6243 had to be discarded as a result of mapping the identifiers from miRBase v19 to v22.1, which we performed with the miEAA miRBase version converter. Subsequently, we applied GSEA on the list of miRNAs sorted by decreasing effect size between the premalignant and the invasive stage, for NeuT+ and NeuT- samples separately. Strikingly, the former run returned 311 significant categories, while the latter returned none. Overall, many more categories seemed to be depleted ($N = 301$) than enriched ($N = 9$), suggesting a wide-spread up-regulation of molecular pathways as miRNAs get down-regulated in NeuT+. For example, we found Macrophage differentiation ($P = 2.54 \times 10^{-5}$), Vasculature development ($P = 1.60 \times 10^{-4}$), and VEGF signaling pathway ($P = 0.0016$) to be depleted, which might be a signal for the increased tumor burden of NeuT+ mice at the invasive breast cancer stage. Moreover, we evaluated GSEAs for the comparison of NeuT+ and NeuT- at all three stages. While the first two setups returned a rather unspecific set of categories with all *P*-values located close to the significance boundary, the last comparison yielded many interesting results. First, observations were in line with the group-wise comparison along the age dimension, because all categories are depleted, i.e. no enrichments at the top of the sorted list. Further, the results show that several dozen conserved miRNAs ($P = 4.53 \times 10^{-5}$) are down-regulated in the NeuT+ model at the invasive stage. More significant categories we



Step 1 of 5: Select miRNA or precursor as input for analysis

Input type: ● miRNA ○ miRNA precursor

Next

Step 2 of 5: Select type of enrichment analysis

Analysis type: ● Over-Representation Analysis (ORA) ○ miRNA enrichment analysis ((G)SEA)

Next

Step 3 of 5: Paste or upload testset

Paste example

Species: ● Homo sapiens ○ Mus musculus ○ Rattus norvegicus ○ Arabidopsis thaliana ○ Bos taurus ○ Caenorhabditis elegans ○ Drosophila melanogaster ○ Danio rerio ○ Gallus gallus ○ Sus scrofa

Testset:
hsa-miR-20b-5p
hsa-miR-144-5p
hsa-miR-17-5p
hsa-miR-20a-5p
hsa-miR-222-3p
hsa-miR-106a-5p
hsa-miR-93-5p
hsa-miR-126-3p
hsa-miR-363-3p
hsa-miR-302c-3p

Testset file: Browse... No file selected.

? Plain text file containing one miRNA or precursor per line
→ Example File miRNAs
→ Example File precursors

Next

Step 4 of 5: Paste or upload reference set (optional)

Reference set:

Reference set file: Browse... No file selected.

? Plain text file containing one miRNA or precursor per line
→ Example File miRNAs
→ Example File precursors

Next

Step 5 of 5: Select additional parameters for analysis

☑ Un/check all default categories

Categories: ☑ Annotations derived over miRTarBase (Gene Ontology) ☑ Pathways (KEGG) ☑ exRNA forms (miRandola) □ Chromosomal location (miRBase) ☑ Conservation (miRBase) ☑ Confidence (miRBase) ☑ Seed family (miRBase) □ Target genes (miRTarBase) ☑ Diseases (miRWalk) ☑ Gene Ontology (miRWalk) ☑ Organs (miRWalk) ☑ Pathways (miRWalk) ☑ Diseases (MNDR) ☑ Gender/Age □ Cell-type specific (Atlas) □ Published Diseases □ Immune cells ☑ Localization (RNALocate) ☑ Drugs (SM2miR) ☑ Expressed in tissue (Tissue Atlas)

Specialized categories (Expert mode): Show categories

P-value adjustment method: ○ No adjustment ● FDR (Benjamini-Hochberg) adjustment ○ Bonferroni adjustment ○ Benjamini-Yekutieli adjustment ○ Hochberg adjustment ○ Holm adjustment ○ Hommel adjustment

Benjamini-Hochberg FDR adjustment is also reported as 'Q-value' in results.

P-value adjustment scope: ● Adjust p-values for each category independently ○ Adjust p-values for all categories collectively

Significance level: 0.05

Minimum required hits per sub-category: 2

Submit

Results for input list

- View significant annotations for the uploaded miRNAs/precursors
- View sequence properties for the uploaded miRNAs/precursors

Show 5 rows CSV Excel

Category	Subcategory		Enrichment	P-value	P-adjusted	Q-value	Expected	Observed	miRNAs/precursors
				From To	From To	From To	From To	From To	
Annotations derived over miRTarBase (Gene Ontology)	U2 snRNA binding GO0030620	↑	over-represented	3.85e-9	5.46e-6	5.46e-6	1.12061	11	hsa-miR-20b-5p; hsa-miR-17-5p; hsa-miR-20a-5p; hsa-miR-222-3p; hsa-miR-106a-5p; hsa-miR-93-5p; hsa-miR-374b-5p; hsa-miR-484; hsa-miR-106b-5p; hsa-miR-93-3p; hsa-miR-409-3p
Annotations derived over miRTarBase (Gene Ontology)	UDP-glucuronate decarboxylase activity GO0048040	↑	over-represented	5.34e-9	5.46e-6	5.46e-6	0.0901639	5	hsa-miR-20b-5p; hsa-miR-17-5p; hsa-miR-20a-5p; hsa-miR-93-5p; hsa-miR-106b-5p
Annotations derived over miRTarBase (Gene Ontology)	cap2 mRNA methylation GO0097310	↑	over-represented	5.34e-9	5.46e-6	5.46e-6	0.0901639	5	hsa-miR-20b-5p; hsa-miR-17-5p; hsa-miR-20a-5p; hsa-miR-93-5p; hsa-miR-106b-5p
Annotations derived over miRTarBase (Gene Ontology)	endothelial cell chemotaxis to vascular endothelial growth factor GO0090668	↑	over-represented	4.57e-9	5.46e-6	5.46e-6	0.167447	6	hsa-miR-20b-5p; hsa-miR-17-5p; hsa-miR-20a-5p; hsa-miR-93-5p; hsa-miR-106b-5p; hsa-miR-409-3p
Annotations derived over miRTarBase (Gene Ontology)	eukaryotic 48S preinitiation complex GO0033290	↑	over-represented	3.59e-9	5.46e-6	5.46e-6	0.618267	9	hsa-miR-20b-5p; hsa-miR-17-5p; hsa-miR-20a-5p; hsa-miR-222-3p; hsa-miR-106a-5p; hsa-miR-93-5p; hsa-miR-302c-3p; hsa-miR-130b-5p; hsa-miR-106b-5p

Figure 1. miEAA workflow and exemplary results. (**A**) Each miRNA/precursor enrichment analysis consists of at most five steps. First, users should select whether they want to perform enrichment on precursors or miRNAs. Second, the enrichment algorithm, i.e. either ORA or GSEA must be selected. Next, the desired test set can be defined either through a textbox or a file upload. The fourth step only appears for ORAs where custom background reference sets can be inserted or uploaded. This is optional since miEAA provides pre-computed reference sets for all categories. Lastly, the set of categories and databases as well as statistical parameters should be selected. (**B**) Typical result view for an ORA. Users can sort, select, filter, and export the obtained enrichment results interactively. Moreover, several visualizations of the results are provided for each run, such as the precursor/miRNA to category heatmap and the category word cloud.

found such as exosome ($P = 2.31 \times 10^{-5}$) and circulating ($P = 0.0086$) miRNAs, breast cancer ($P = 0.0094$, Figure 2(b)), microRNAs in cancer ($P = 0.028$), and PI3K-Akt signaling pathway ($P = 0.028$) can be associated with the research setup of this exemplifying study.

Case study 3: Parkinson's Biomarkers from PPMI and comparison to TAM 2.0

At last we aimed to test a non-cancer disease (Parkinson's), to present a direct comparison between TAM 2 and miEAA 2.0. We compared the raw *P*-values of the tools to exclude an influence of the size of available categories. A direct comparison highlighted 72 hits by both tools (additional 70 reported only by TAM and 144 only by miEAA). Very similar but not exactly matching category names (e.g. *Alzheimer's* versus *Alzheimers* or *Carcinoma, Lung, Non-Small-Cell* versus *Carcinoma, Lung. Non-Small-Cell*) had to be matched manually. After matching those, several ambiguously defined categories remained, e.g. *Human Immunodeficiency Virus Infection* in miEAA and *Acquired Immunodeficiency Syndrome* in TAM and that had to be mapped. As a result, the overlap increased to 94 hits. Asking whether the overlap between the output of the two tools is larger for the categories with higher significance than expected, we performed a DynaVenn analysis of the result sets ordered by increasing *P*-value (42). Selecting the 32 most significant miEAA sets and the 30 most significant TAM sets we observed an overlap of 23 categories ($P = 10^{-8}$), indeed suggesting better comparable results for the most significant categories. Also, when comparing the miRNA hits for the obtained categories we observed very similar results. Alzheimer's Disease was covered by 10 miRNAs in miEAA and nine in TAM with *P*-values of 3.31×10^{-4} and 2.19×10^{-3}, respectively. We also observed the function category of TAM to be advantageous in this case, revealing direct hits such

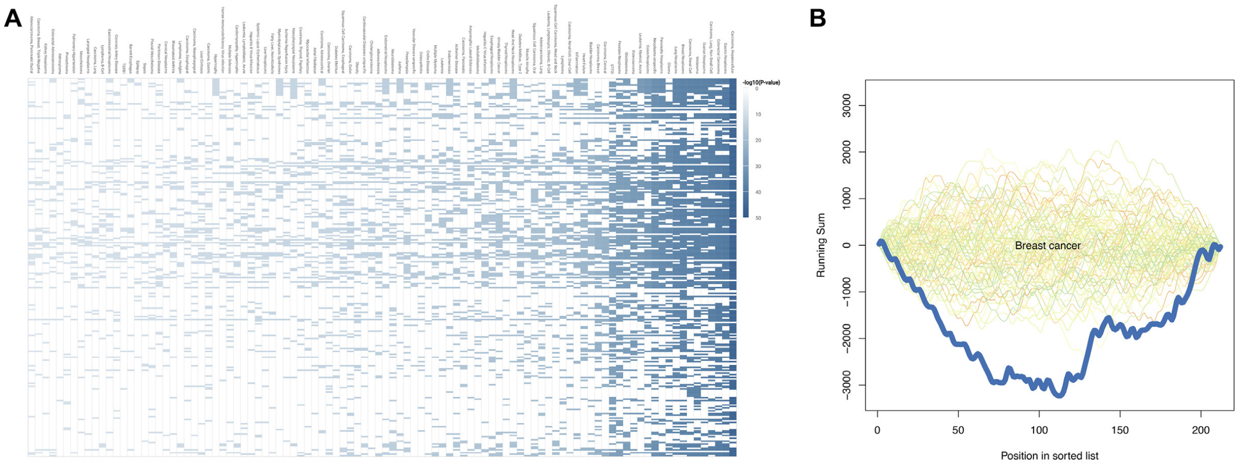


Figure 2. Web server visualisation of case study results. (**A**) Category (x-axis) to precursor (y-axis) heatmap with $-\log_{10}$-scaled enrichment *P*-values for the first case study. (**B**) GSEA plot with simulated background distributions (green to orange lines) and actual depletion for breast cancer (dark blue line) observed during evaluation of the second case study.

as *Aging*, which remained partially hidden in miEAA. On the other hand, miEAA seems to have slight advantages in the disease-associated categories, reporting 176 entries compared to 106 in TAM. This extended list contains among others *Parkinson's Disease* which was covered by three miRNAs in TAM and missing the alpha level while being covered by six miRNAs in miEAA and thus being significant ($P = 0.019$). The full list of results obtained from both tools in direct comparison is shown in Supplementary Table S5. Besides the case study benchmark, we performed a detailed feature comparison with respect to 22 criteria between our tool and TAM that is shown in Supplementary Table S6.

New data export and browsable API

All data, results, and interactive plots shown on the web server are exportable to common data formats. To support the trend towards the development of reproducible and automated data analysis pipelines (12), miEAA hosts a public, browsable API offering the same functionality as the web site, allowing one to access the miRNA converters and statistical algorithms remotely. This functionality is further augmented by a full-featured Python package with API library code and a command-line interface (CLI). For example, a regular workflow as performed on the website can be accomplished with three sequential calls to the web API or one call to the CLI. We provide code examples in the common data science programming languages Python and R to demonstrate this use-case. We also implemented the interface to solve two recurring problems in biological data analysis. First, reproducibility of statistical experiments can be improved, because usage of the versioned API in the context of a workflow manager such as Snakemake (26) or Nextflow fosters self-documenting research setups (43). Second, oftentimes the analysis of miRNA high-throughput data involves the comparison of multiple biogroups, timepoints or other annotation variables. By using our API and the package, multiple runs of miEAA can be performed at ease while minimising the time spent for set up and results aggregation.

DISCUSSION

Statistical tools for biological enrichment analysis are a key to understanding data from high-throughput omics assays. However, the performance primarily depends on the quality of the underlying annotations and the statistical soundness. We show that new developments in the miRNA research field yielded an unprecedented set of biological categories, covering most aspects of miRNA properties and function, with cross-species analysis becoming increasingly important. On the other side, as with every statistical framework applied on biological data, assumptions are not always met and findings should be assessed critically in the light of further validation experiments. The novel release of miEAA attempts to cover these aspects by enhancing the set of available categories both quantitatively and qualitatively as well as through offering more (stringent) approaches for *P*-value correction. Also, a major limitation of some datasets concerns the availability of mature miRNA identifiers, as only precursor names were available for some of the sources. However, especially in the context of diseases, mature miRNA resolution is preferable to match the biological selectivity for one major miRNA arm being expressed. Datasets incorporated in miEAA were compiled either automatically or manually. The competitor tool TAM uses a fewer number of high-quality annotations. In particular, an advantage of TAM arises from the manual curation of datasets (14). The case study on Parkinson's disease highlighted the results of miEAA 2.0 and TAM 2.0 to be similar whereas individual advantages in usability, functionality, or scope in the one or the other tool remain.

We have demonstrated the capability of miEAA to yield novel biological results in cancer research. For the kidney renal clear cell carcinoma case study, we found a depletion of the mir-506 precursor family, which has been observed before in other types of cancers (44,45). Many interactions to transcription factors were also found for the up-regulated miRNAs, suggesting an increased regulatory burden due to the exceeding transcriptional up-regulation observed in cancer. For example, HEY1, which is a transcriptional repressor has been characterised to be up-regulated in renal cell carcinomas (46). For the mouse breast cancer progression study, we illustrated the backwards compatibility of miEAA with respect to miRBase. The overall observed depletion of pathway regulating miRNAs in mice agrees with our first case study. Moreover, the significant categories like vasculature development that are associated with morphogenesis, resemble an increased tumor burden of NeuT+ mice, which was previously confirmed with a large human RNA-seq dataset on breast cancer (47). In both case studies, we observed many associations with other types of cancers or diseases. While this may speak for a molecular and biological similarity, a certain publication bias, e.g. for cancer, is a confounding factor that skews the statistics (14).

Establishing a standardized nomenclature is an on-going challenge in miRNA research. Results of the implemented manual converters are more accurate as compared to automated mappings since the naming schemes changed along the different releases. miEAA supports an exact mapping of old (e.g. miR*) to new nomenclature which would be ambiguous using automatic conversion (e.g. hsa-miR-499a-3p could be converted to hsa-miR-499a-3p or hsa-miR-499b-3p). Similar ambiguity issues would arise by performing a case insensitive miRNA to precursor mapping ('miR' to 'mir'), in case multiple precursors with the same miRNA exist (for example hsa-let-7a-5p is annotated in three precursors). Finally, we sought to improve accessibility of miEAA and developed a web-API in combination with a Python package. These features enhance its usability in other applications for miRNA research, for example to annotate functional sub-graphs in regulatory network analysis (48). In conclusion, miEAA 2.0 is a flexible, comprehensive, and highly accessible tool for high-throughput miRNA annotation and enrichment analysis.

DATA AVAILABILITY

miEAA 2.0 is freely available at https://www.ccb.uni-saarland.de/mieaa2. No login is required. Example code for API-usage and the pre-compiled Python package are freely available from https://github.com/Xethic/miEAA-API.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the authors of the utilized GEO dataset for providing their microarray samples to the general public. The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. We would like to express gratitude towards all specimen donors and research groups involved in the sample acquisition.

FUNDING

Michael J. Fox foundation [14446]. Internal funds of Saarland University. Funding for open access charge: Michael J. Fox Foundation for Parkinson's Research.

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2018) Metazoan MicroRNAs. *Cell*, **173**, 20–51.
2. Kern,F., Backes,C., Hirsch,P., Fehlmann,T., Hart,M., Meese,E. and Keller,A. (2019) What's the target: understanding two decades of in silico microRNA-target prediction. *Brief. Bioinform*, doi:10.1093/bib/bbz111.
3. Cantini,L., Bertoli,G., Cava,C., Dubois,T., Zinovyev,A., Caselle,M., Castiglioni,I., Barillot,E. and Martignetti,L. (2019) Identification of microRNA clusters cooperatively acting on epithelial to mesenchymal transition in triple negative breast cancer. *Nucleic Acids Res.*, **47**, 2205–2215.
4. Ludwig,N., Fehlmann,T., Kern,F., Gogol,M., Maetzler,W., Deutscher,S., Gurlit,S., Schulte,C., von Thaler,A.K., Deuschle,C. *et al.* (2019) Machine learning to detect Alzheimer's disease from circulating Non-coding RNAs. *Genomics Proteomics Bioinformatics*, **17**, 430–440.
5. Thomou,T., Mori,M.A., Dreyfuss,J.M., Konishi,M., Sakaguchi,M., Wolfrum,C., Rao,T.N., Winnay,J.N., Garcia-Martin,R., Grinspoon,S.K. *et al.* (2017) Adipose-derived circulating miRNAs regulate gene expression in other tissues. *Nature*, **542**, 450–455.
6. 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.
7. Fehlmann,T., Backes,C., Alles,J., Fischer,U., Hart,M., Kern,F., Langseth,H., Rounge,T., Umu,S.U., Kahraman,M. *et al.* (2018) A high-resolution map of the human small non-coding transcriptome. *Bioinformatics*, **34**, 1621–1628.
8. 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.*, **47**, 3353–3364.
9. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
10. Mora,A. (2019) Gene set analysis methods for the functional interpretation of non-mRNA data—genomic range and ncRNA data. *Brief. Bioinform*, doi:10.1093/bib/bbz090.
11. Backes,C., Khaleeq,Q.T., Meese,E. and Keller,A. (2016) MiEAA: MicroRNA enrichment analysis and annotation. *Nucleic Acids Res.*, **44**, W110–W116.
12. Perkel,J.M. (2019) Workflow systems turn raw data into scientific knowledge. *Nature*, **573**, 149–150.
13. Lu,M., Shi,B., Wang,J., Cao,Q. and Cui,Q. (2010) TAM: A method for enrichment and depletion analysis of a microRNA category in a list of microRNAs. *BMC Bioinformatics*, **11**, 419.
14. Li,J., Han,X., Wan,Y., Zhang,S., Zhao,Y., Fan,R., Cui,Q. and Zhou,Y. (2018) TAM 2.0: Tool for MicroRNA set analysis. *Nucleic Acids Res.*, **46**, W180–W185.
15. Çorapçıoğlu,M. and Oğul,H. (2015) miSEA: microRNA set enrichment analysis. *Biosystems*, **134**, 37–42.
16. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2018) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
17. 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.
18. Huang,H.Y., Lin,Y.C.D., Li,J., Huang,K.Y., Shrestha,S., Hong,H.C., Tang,Y., Chen,Y.G., Jin,C.N., Yu,Y. *et al.* (2019) miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic Acids Res.*, **48**, D148–D154.
19. Kehl,T., Kern,F., Backes,C., Fehlmann,T., Stöckel,D., Meese,E., Lenhof,H.P. and Keller,A. (2020) miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database. *Nucleic Acids Res.*, **48**, D142–D147.
20. 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.

21. Huang,Z., Shi,J., Gao,Y., Cui,C., Zhang,S., Li,J., Zhou,Y. and Cui,Q. (2018) HMDD v3.0: a database for experimentally supported human microRNA–disease associations. *Nucleic Acids Res.*, **47**, D1013–D1017.
22. Cui,T., Zhang,L., Huang,Y., Yi,Y., Tan,P., Zhao,Y., Hu,Y., Xu,L., Li,E. and Wang,D. (2017) MNDR v2.0: an updated resource of ncRNA–disease associations in mammals. *Nucleic Acids Res.*, **46**, D371–D374.
23. Tong,Z., Cui,Q., Wang,J. and Zhou,Y. (2018) TransmiR v2.0: an updated transcription factor-microRNA regulation database. *Nucleic Acids Res.*, **47**, D253–D258.
24. Zhang,T., Tan,P., Wang,L., Jin,N., Li,Y., Zhang,L., Yang,H., Hu,Z., Zhang,L., Hu,C. *et al.* (2016) RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res.*, **45**, D135–D138.
25. Russo,F., Di Bella,S., Vannini,F., Berti,G., Scoyni,F., Cook,H.V., Santos,A., Nigita,G., Bonnici,V., Laganà,A. *et al.* (2017) miRandola 2017: a curated knowledge base of non-invasive biomarkers. *Nucleic Acids Res.*, **46**, D354–D359.
26. Köster,J. and Rahmann,S. (2012) Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.
27. Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W., Bright,J. *et al.* (2020) SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272.
28. The Gene Ontology Consortium (2018) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
29. Kanehisa,M., Sato,Y., Furumichi,M., Morishima,K. and Tanabe,M. (2018) New approach for understanding genome variations in KEGG. *Nucleic Acids Res.*, **47**, D590–D595.
30. Teng,X., Chen,X., Xue,H., Tang,Y., Zhang,P., Kang,Q., Hao,Y., Chen,R., Zhao,Y. and He,S. (2019) NPInter v4.0: an integrated database of ncRNA interactions. *Nucleic Acids Res.*, **48**, D160–D165.
31. Liu,X., Wang,S., Meng,F., Wang,J., Zhang,Y., Dai,E., Yu,X., Li,X. and Jiang,W. (2012) SM2miR: a database of the experimentally validated small molecules' effects on microRNA expression. *Bioinformatics*, **29**, 409–411.
32. de Rie,D., Abugessaisa,I., Alam,T., Arner,E., Arner,P., Ashoor,H., Åström,G., Babina,M., Bertin,N., Burroughs,A.M. *et al.* (2017) An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat. Biotechnol.*, **35**, 872–878.
33. Minami,K., Uehara,T., Morikawa,Y., Omura,K., Kanki,M., Horinouchi,A., Ono,A., Yamada,H., Ohno,Y. and Urushidani,T. (2014) miRNA expression atlas in male rat. *Scientific Data*, **1**, 140005.
34. Dweep,H. and Gretz,N. (2015) MiRWalk2.0: A comprehensive atlas of microRNA-target interactions. *Nat. Methods*, **12**, 697.
35. The RNAcentral Consortium (2018) RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Res.*, **47**, D221–D229.
36. Xu,T., Su,N., Liu,L., Zhang,J., Wang,H., Zhang,W., Gui,J., Yu,K., Li,J. and Le,T.D. (2018) miRBaseConverter: an R/Bioconductor package for converting and retrieving miRNA name, accession, sequence and family information in different versions of miRBase. *BMC Bioinformatics*, **19**, 514.
37. Keller,A., Backes,C. and Lenhof,H.P. (2007) Computation of significance scores of unweighted gene set enrichment analyses. *BMC Bioinformatics*, **8**, 290.
38. Marek,K., Chowdhury,S., Siderowf,A., Lasch,S., Coffey,C.S., Caspell-Garcia,C., Simuni,T., Jennings,D., Tanner,C.M., Trojanowski,J.Q. *et al.* (2018) The Parkinson's progression markers initiative (PPMI)– establishing a PD biomarker cohort. *Ann. Clin. Transl. Neur.*, **5**, 1460–1477.
39. Stöckel,D., Kehl,T., Trampert,P., Schneider,L., Backes,C., Ludwig,N., Gerasch,A., Kaufmann,M., Gessler,M., Graf,N. *et al.* (2016) Multi-omics enrichment analysis using the GeneTrail2 web service. *Bioinformatics*, **32**, 1502–1508.
40. Korthauer,K., Kimes,P.K., Duvallet,C., Reyes,A., Subramanian,A., Teng,M., Shukla,C., Alm,E.J. and Hicks,S.C. (2019) A practical guide to methods controlling false discoveries in computational biology. *Genome. Biol.*, **20**, 118.
41. Chiodoni,C., Cancila,V., Renzi,T.A., Perrone,M., Tomirotti,A.M., Sangaletti,S., Botti,L., Dugo,M., Milani,M., Bongiovanni,L. *et al.* (2020) Transcriptional profiles and stromal changes reveal bone marrow adaptation to early breast cancer in association with deregulated circulating microRNAs. *Cancer Res.*, **80**, 484–498.
42. Amand,J., Fehlmann,T., Backes,C. and Keller,A. (2019) DynaVenn: web-based computation of the most significant overlap between ordered sets. *BMC Bioinformatics*, **20**, 743.
43. DI Tommaso,P., Chatzou,M., Floden,E.W., Barja,P.P., Palumbo,E. and Notredame,C. (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.
44. Li,J., Wu,H., Li,W., Yin,L., Guo,S., Xu,X., Ouyang,Y., Zhao,Z., Liu,S., Tian,Y. *et al.* (2016) Downregulated miR-506 expression facilitates pancreatic cancer progression and chemoresistance via SPHK1/Akt/NF-κB signaling. *Oncogene*, **35**, 5501–5514.
45. Zhang,L., Zhou,H. and Wei,G. (2019) miR-506 regulates cell proliferation and apoptosis by affecting RhoA/ROCK signaling pathway in hepatocellular carcinoma cells. *Int. J. Clin. Exp. Pathol.*, **12**, 1163–1173.
46. Karim,S., Al-Maghrabi,J.A., Farsi,H.M.A., Al-Sayyad,A.J., Schulten,H.J., Buhmeida,A., Mirza,Z., Al-boogmi,A.A., Ashgan,F.T., Shabaad,M.M. *et al.* (2016) Cyclin D1 as a therapeutic target of renal cell carcinoma- a combined transcriptomics, tissue microarray and molecular docking study from the Kingdom of Saudi Arabia. *BMC Cancer*, **16**, 741.
47. Tapia-Carrillo,D., Tovar,H., Velazquez-Caldelas,T.E. and Hernandez-Lemus,E. (2019) Master regulators of signaling Pathways: An application to the analysis of gene regulation in breast cancer. *Front. Genet.*, **10**, 1180.
48. Fan,Y., Siklenka,K., Arora,S.K., Ribeiro,P., Kimmins,S. and Xia,J. (2016) miRNet - dissecting miRNA-target interactions and functional associations through network-based visual analysis. *Nucleic Acids Res.*, **44**, W135–W141.

3.5 *Common diseases alter the physiological age-related blood microRNA profile*

nature COMMUNICATIONS

ARTICLE

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

Common diseases alter the physiological age-related blood microRNA profile

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

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

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

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

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

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

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

Results

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

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

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

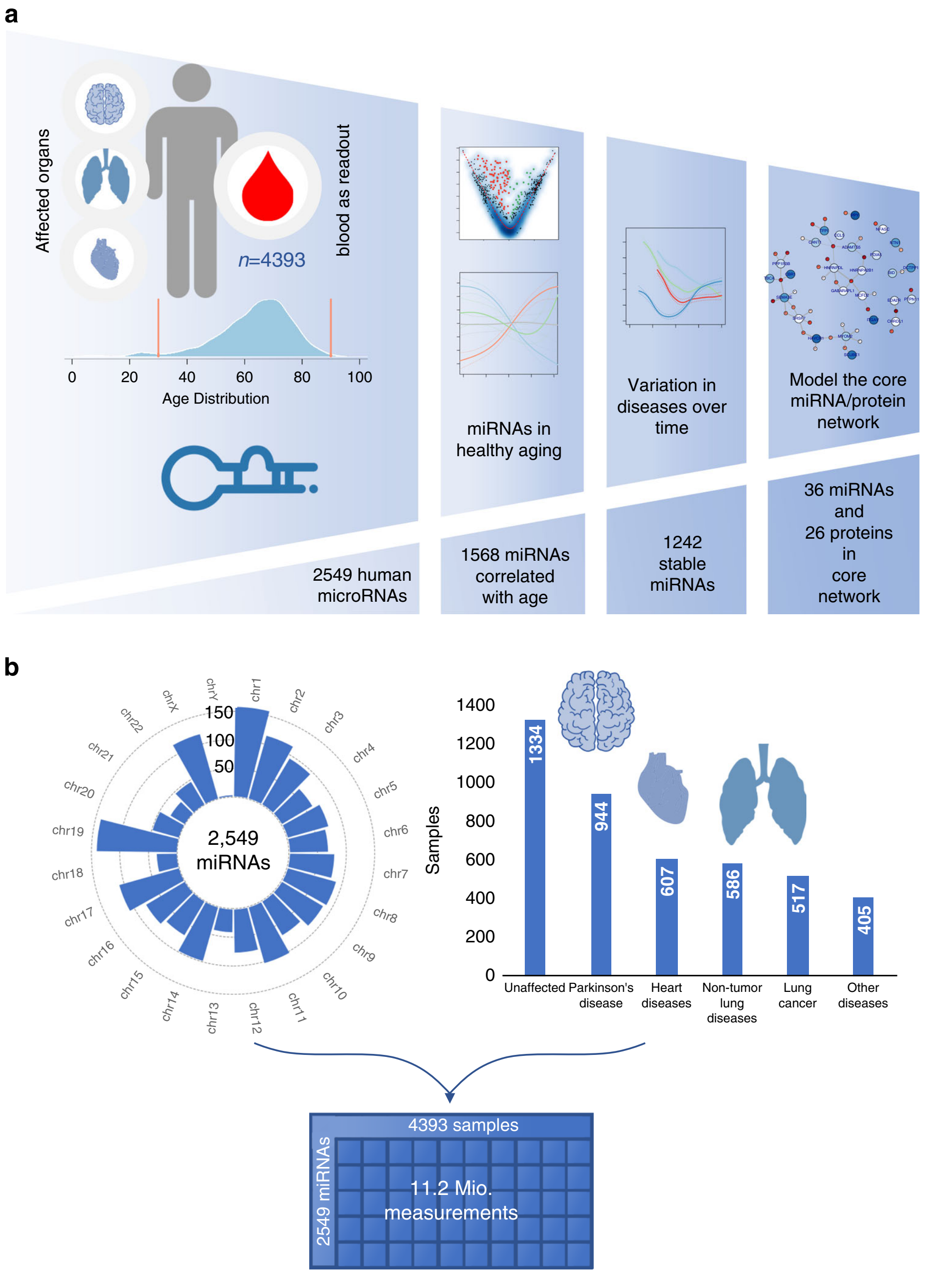
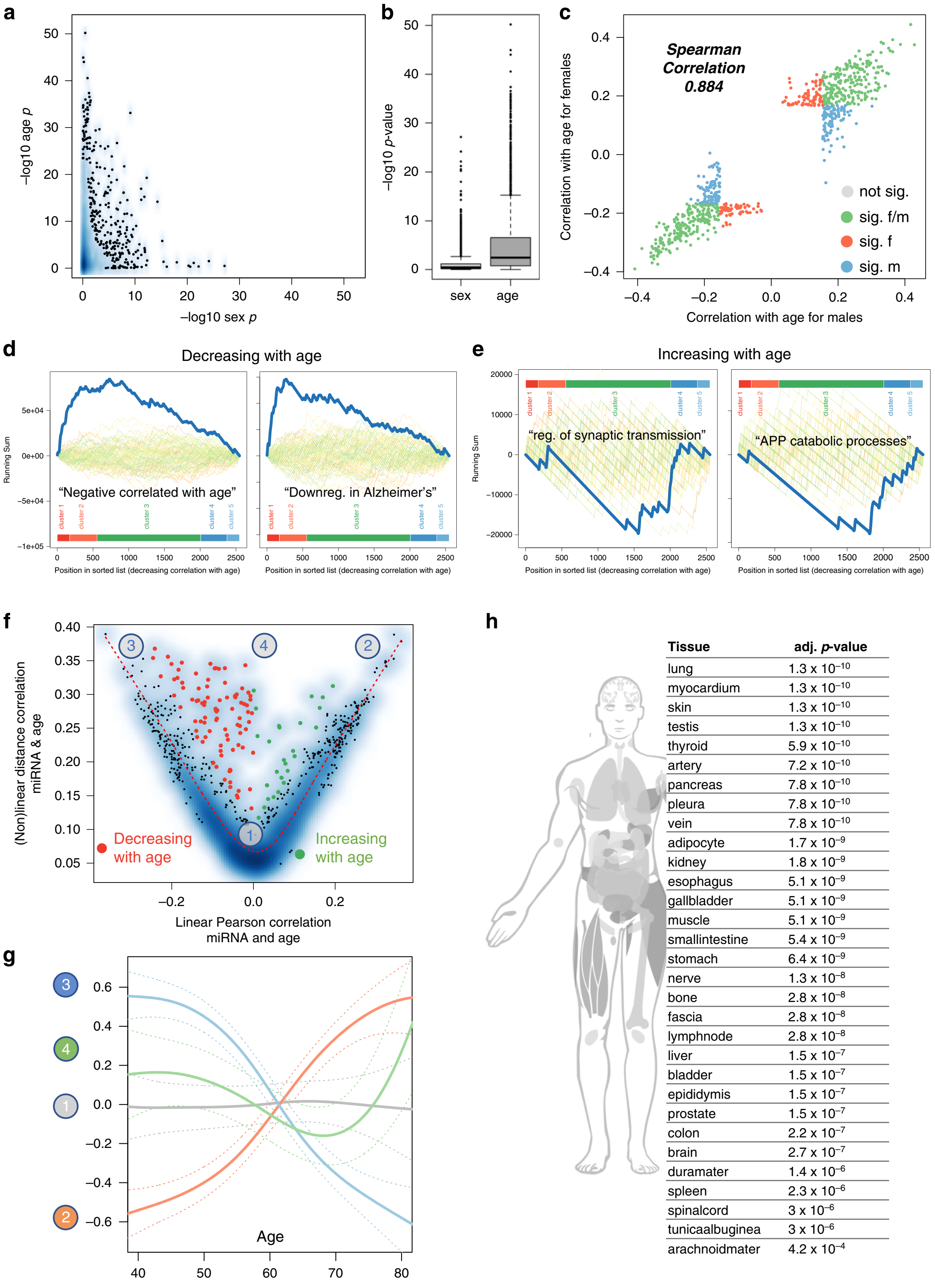


Fig. 1 Study characteristics. a Study set up and analysis workflow from high-throughput data to a specific aging network. The cohort consist of 4393 samples of which the age distribution is provided. For the 4393 samples genome wide miRNA screening using microarrays has been performed. The first analysis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

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

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



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

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

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

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

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

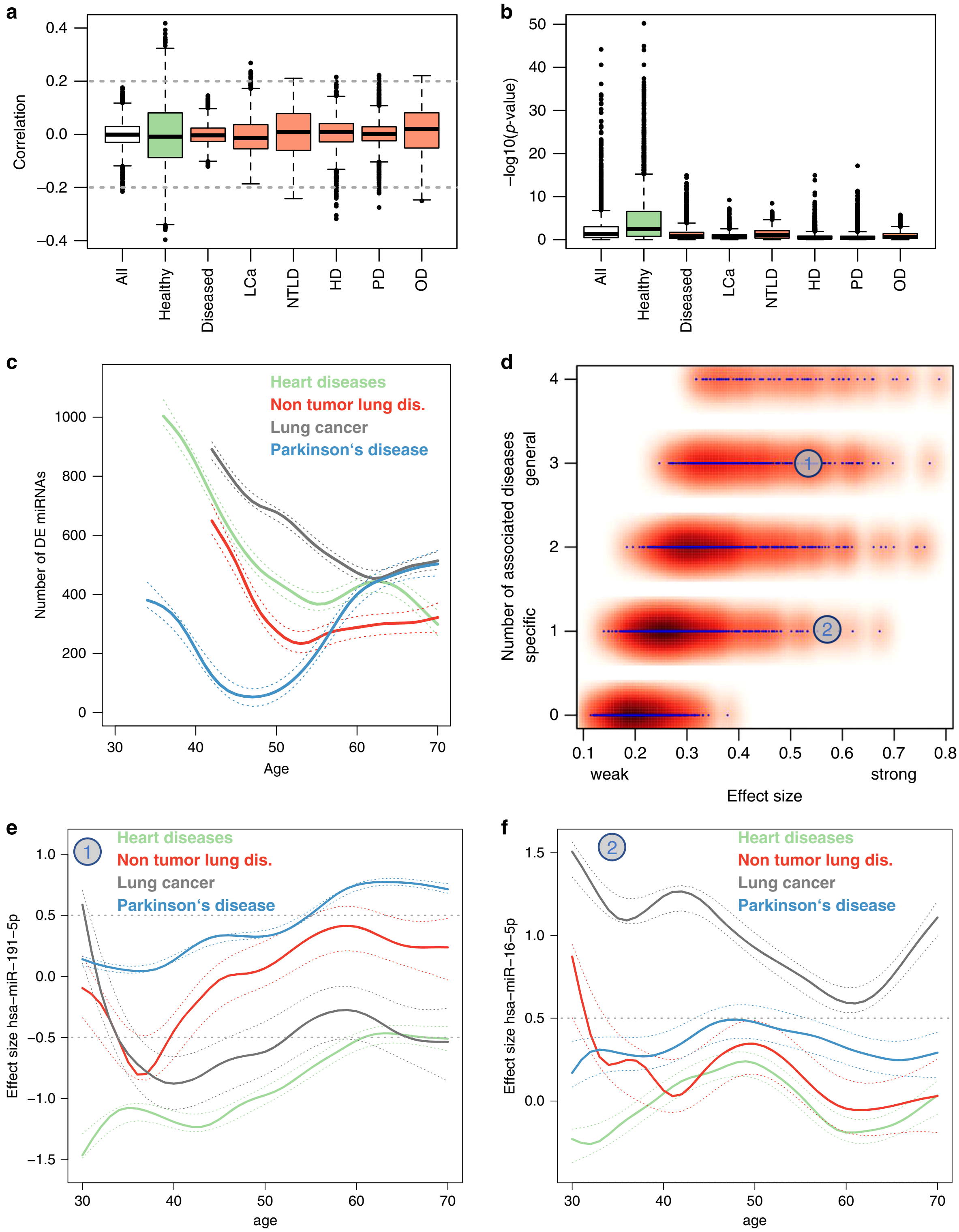
To determine the extent to which diseases affect miRNA abundance compared to healthy controls, we computed the number of differentially expressed miRNAs between cases and controls using a sliding window analysis. That is, we first compared diseased individuals aged 30–39 years to healthy individuals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

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

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

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

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



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

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

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

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

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

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

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

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

Discussion

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

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

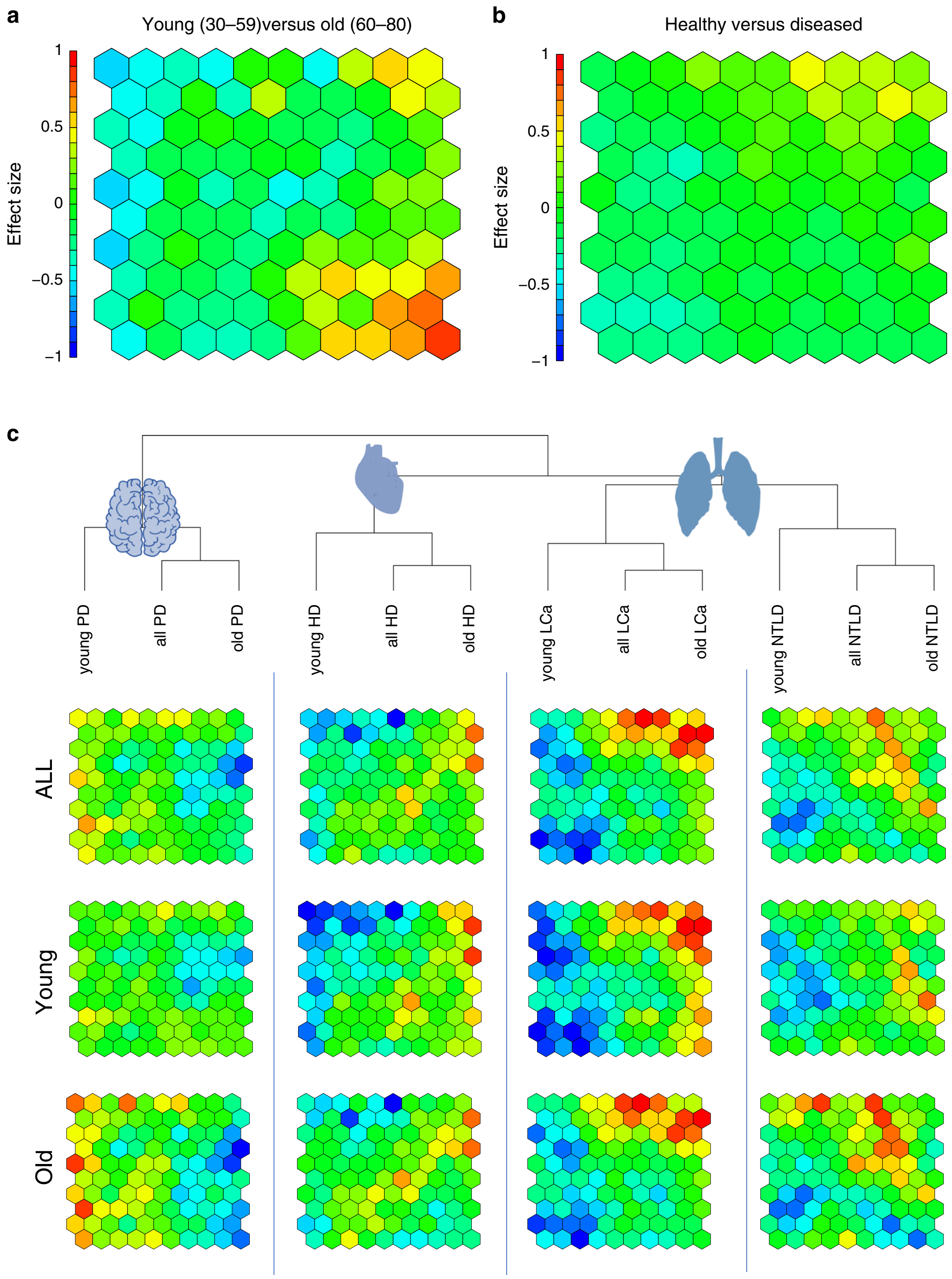


Fig. 4 Disease specificity of miRNA biomarkers. a Heat map representation of the SOM analysis as a 10 × 10 grid with 100 entries. Each cell contains at least one miRNA and up to 20 miRNAs. The full annotation of miRNAs to cells are provided in Supplementary Data 7). The cells are colored by the effect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

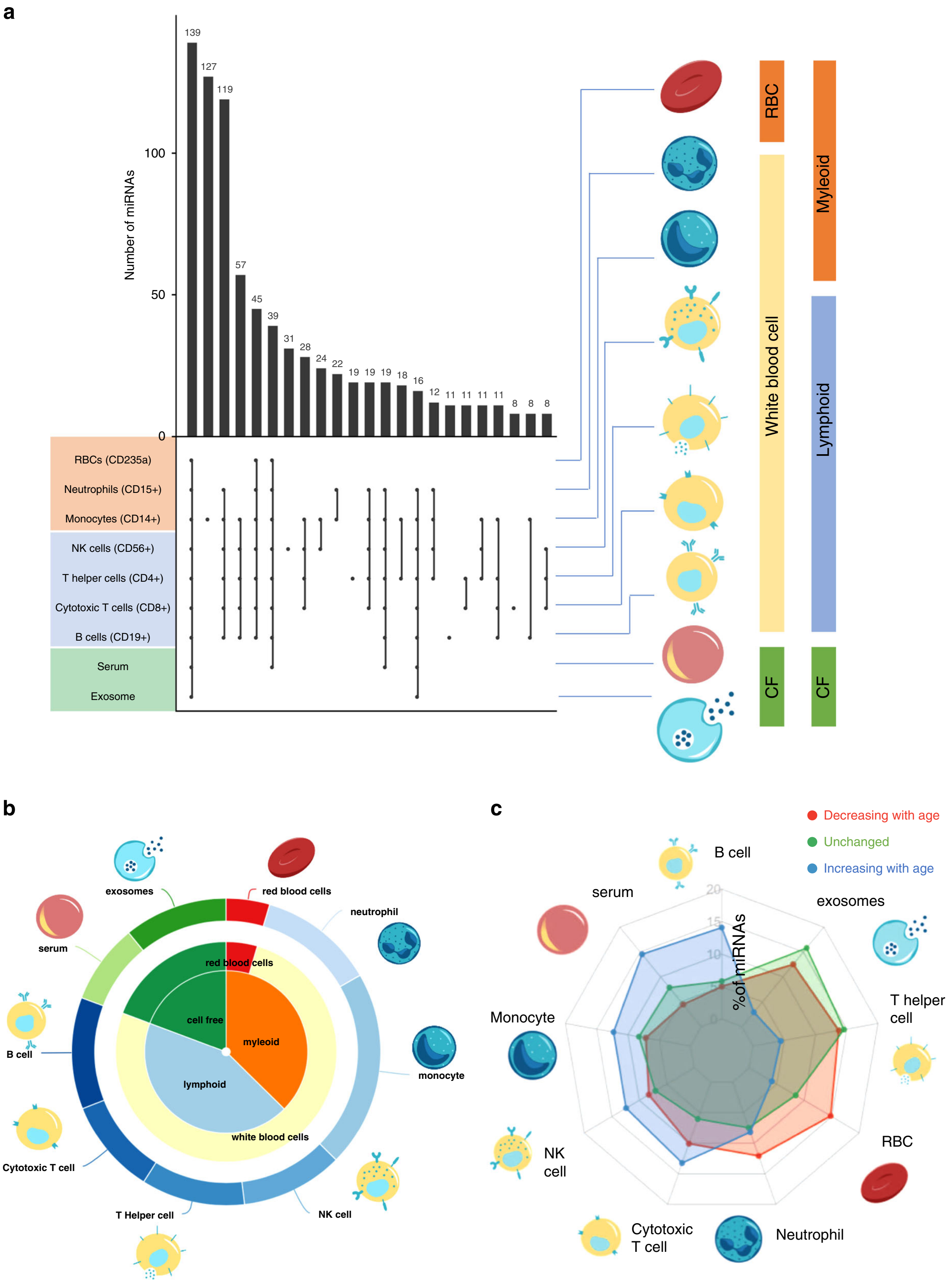
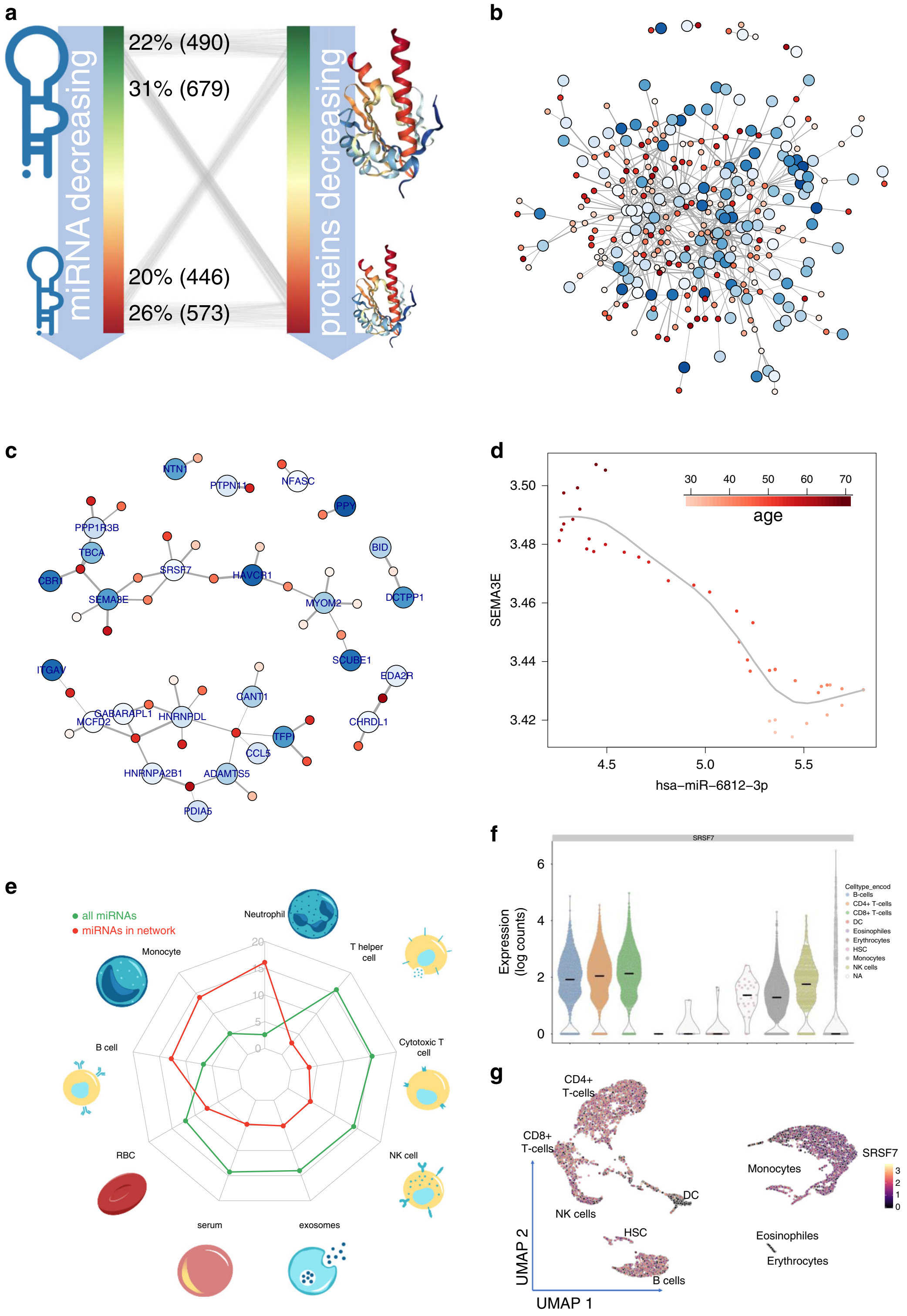


Fig. 5 Blood cell deconvolution. a The distribution of miRNAs in the different blood compounds. The rows are sorted by the blood compounds given on the right (RBC: red blood cell; CF: cell free), the columns are sorted according to a decreasing number of miRNAs. **b** Relative abundance of all miRNAs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



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

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

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

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

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

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

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

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

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

Methods

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

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

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

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

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

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

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

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

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

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

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

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

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

Data availability

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

Code availability

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

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

References

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

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

Acknowledgements

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

Author contributions

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

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

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

Additional information

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

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

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

Published online 10 December 2020

Nucleic Acids Research, 2021, Vol. 49, No. 1 ***127–144***
doi: 10.1093/nar/gkaa1161

Validation of human microRNA target pathways enables evaluation of target prediction tools

Fabian Kern[1,†], Lena Krammes[2,†], Karin Danz[3], Caroline Diener[2], Tim Kehl[4], Oliver Küchler[1], Tobias Fehlmann[1], Mustafa Kahraman[1], Stefanie Rheinheimer[2], Ernesto Aparicio-Puerta[1,5,6], Sylvia Wagner[3], Nicole Ludwig[2,7], Christina Backes[1], Hans-Peter Lenhof[4], Hagen von Briesen[3], Martin Hart[2,†], Andreas Keller[1,4,8,*,†] and Eckart Meese[2,†]

[1]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [2]Institute of Human Genetics, Saarland University, 66421 Homburg, Germany, [3]Department of Bioprocessing & Bioanalytics, Fraunhofer Institute for Biomedical Engineering, 66280 Sulzbach, Germany, [4] Center for Bioinformatics, Saarland Informatics Campus, Saarland University, 66123 Saarbrücken, Germany, [5]Department of Genetics, Faculty of Science, University of Granada, 18071 Granada, Spain, [6]Instituto de Investigación Biosanitaria ibs. Granada, University of Granada, 18071 Granada, Spain, [7]Center of Human and Molecular Biology, Saarland University, 66123 Saarbrücken, Germany and [8]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, CA, USA

Received April 25, 2020; Revised October 20, 2020; Editorial Decision November 11, 2020; Accepted November 13, 2020

ABSTRACT

MicroRNAs are regulators of gene expression. A wide-spread, yet not validated, assumption is that the targetome of miRNAs is non-randomly distributed across the transcriptome and that targets share functional pathways. We developed a computational and experimental strategy termed high-throughput miRNA interaction reporter assay (HiTmIR) to facilitate the validation of target pathways. First, targets and target pathways are predicted and prioritized by computational means to increase the specificity and positive predictive value. Second, the novel webtool miRTaH facilitates guided designs of reporter assay constructs at scale. Third, automated and standardized reporter assays are performed. We evaluated HiTmIR using miR-34a-5p, for which TNF- and TGFB-signaling, and Parkinson's Disease (PD)-related categories were identified and repeated the pipeline for miR-7-5p. HiTmIR validated 58.9% of the target genes for miR-34a-5p and 46.7% for miR-7-5p. We confirmed the targeting by measuring the endogenous protein levels of targets in a neuronal cell model. The standardized positive and negative targets are collected in the new miRATBase database, representing a resource for training, or benchmarking new target predictors. Applied to 88 target predictors with different confidence scores, TargetScan 7.2 and miRanda outperformed other tools. Our experiments demonstrate the efficiency of HiTmIR and provide evidence for an orchestrated miRNA-gene targeting.

INTRODUCTION

MicroRNAs (miRNAs) are small non coding RNAs, which regulate the gene expression post-transcriptionally (1). Specifically, miRNAs repress protein translation of target mRNAs by binding to target sequences mainly in 3′ untranslated regions (3′UTRs) and less commonly in 5′ untranslated regions or open reading frames of their target mRNAs (2,3). Aberrant expression of miRNAs is not only a hallmark of various cancers and can be detected in tumor cells and body fluids including urine, saliva, and blood (4–6), but also in solid tissue, cerebrospinal fluid, and blood of neuropathological disorders like Alzheimer's Disease and Parkinson's Disease (PD) (7–10).

While miRNA gene targeting relies on a complementary binding of the seed region to the target gene, non-canonical binding between gene and miRNA also seems to have a deterministic influence on the targeting process (11,12). The limited understanding of the true complexity of the interactions between miRNAs and genes poses substantial challenges for the computational prediction of miRNA targets. In response to this challenge, many tools have been developed including TargetScan (13), PicTar (14), miRanda (15) and other consensus methods like miRWalk (16), which

[*]To whom correspondence should be addressed. Tel: +49 174 1684638; Fax: +49 174 1684638; Email: andreas.keller@ccb.uni-saarland.de
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors and that the last three authors should be regarded as Joint Last Authors.

in turn combines the predictive power of several other predictors. The expectable number of targets per miRNA has not yet been reliably determined as a single miRNA can target between a few up to several hundred genes. Considering an overall search space of 62.5 million possible miRNA-gene interactions (25 000 human genes × 2500 human miRNAs) and the estimated number of targets of single miRNAs, a substantial class imbalance exists. Learning from imbalanced data however still poses challenges for machine learning in life sciences and beyond (17,18). When the *a priori* likelihood of a positive event gets small and the specificity is not close to an optimal value, the positive predictive value, i.e. the likelihood that a predicted event is actually positive, becomes extremely low (19).

Accumulating evidence suggests that the targetome of a miRNA is not randomly distributed across the transcriptome and that it covers genes of shared biochemical pathways. This information can support the design of prediction tools by increasing the specificity of target predictions while at the same time maintaining the sensitivity. Based on this assumption we previously developed the miRNA target pathway dictionary (20), which we subsequently extended into the miRPathDB (21), now existing in the second version (22). The wide-spread assumption that miRNAs target complex networks in an orchestrated manner to facilitate the discovery of new true positive targets has not yet been validated at scale. However, respective computational approaches, which use consensus prediction and target enrichment by pathways, motivate a systematic and standardized experimental validation of predicted targets. To validate miRNA targets, different experimental approaches exist with inherent advantages and disadvantages. One of the most common choices are reporter assays (23,24). As for the majority of similar technologies, limitations of reporter assays are known (25). In addition, manuscripts frequently report only one validated gene or small sets thereof. The miRTarBase in the most recent update 2020 (26) indicates that 6046 manuscripts describe 9679 human miRNA/gene pairs (including duplications) validated by reporter assays. Thus, on average, manuscripts validate only 1.6 targets. Additionally, 97% of the database entries are positive associations while negative results of reporter assays are frequently not reported.

To address the challenge of identifying true miRNA targets in the overall search space of 62.5 million possible miRNA-mRNA interactions, we developed an approach termed **hi**gh-**t**hroughput **m**iRNA **i**nteraction **r**eporter assay (HiTmIR). Our approach combines computational and experimental work steps into a new pipeline. In the computational part, targets are first predicted by a consensus approach relying on well-established tools. Subsequently, targets are filtered by enriched pathways or diseases using the GeneTrail (27) pathway analysis software. From the enriched targetome a novel web-based software (miRTaH) can automatically design reporter sequences for luciferase reporter assays at scale, a task that is challenging and time consuming when performed manually. The final reporter assay target sequences can be obtained from various vendors and get handled by an automated microfluidic device. Therefore, our pipeline allows to identify a higher fraction of true miRNA target interactions than previously reported in an efficient manner. The identified targets and target pathways used to benchmark a variety of target prediction tools and databases in a low, medium, and high stringency set-up have been stored in the miRATBase data warehouse. The overall workflow of our study together with the main contributions to the field are shown in Figure 1.

We applied the HiTmIR workflow to two strongly conserved miRNAs, miR-34a-5p and miR-7-5p, which are both known to be deregulated in cellular PD models and brain tissue of PD patients (28–33). While miR-34a-5p is upregulated in PD, downregulation of miR-7-5p has been previously demonstrated to effect α-synuclein and to contribute to neurodegeneration (28,34). PD is the second most common neurodegenerative disorder following Alzheimer's Disease. Its prevalence strongly increases with age, resulting in 2% of the female world population and 7% of the male world population affected being over 85 years old (35). The clinical symptoms are caused by the loss of dopaminergic neurons within the *substantia nigra pars compacta* and coupled to the accumulation of α-synuclein into intraneuronal structures, known as Lewy bodies and Lewy neurites (36,37). In the last decade, the role of deregulated miRNAs in the pathogenesis of PD has been characterized, for example by the identification of several disease associated miRNAs involved in the progression of PD (38).

MATERIALS AND METHODS

We here describe an overview of the applied methods and analyses. Further details on each of them are available in the supplement and online methods (Supplemental document).

Automated dual luciferase reporter assay

For this assay 2–2.5 × 10^4 HEK 293 T cells were seeded out per well of a 96-well plate (Eppendorf, Hamburg, Germany) by the liquid handling system epMotion 5075 (Eppendorf, Hamburg, Germany). HEK 293 T cells were transfected with 50 ng/well reporter vector with or without 3′UTR and 200 ng/well pSG5 empty vector or pSG5-miR-34a expression plasmid. Forty-eight hours after transfection cells were lysed and the cell lysates were prepared according to manual of the Dual-Luciferase® Reporter Assay System (Promega, Madison, USA) and measured with the GlowMax navigator microplate luminometer (Promega, Madison, USA).

miRNA expression plasmid and reporter constructs

The pSG5-miR-34a expression vector (Eurofins Genomics, Ebersberg, Germany) contains the nucleotides 9 151 617–9 151 816 of chromosome 1. The pSG5-miR-7 expression vector (Eurofins Genomics, Ebersberg, Germany) contains the nucleotides 88 611 724–88 612 046 of chromosome 15. For miR-34a-5p target gene validation, the sequences of the 191 3′UTRs of the TNF-, TGFB-signaling and the PD-related target genes were synthetized and the ∼490 nt long inserts were cloned into the pMIR-RNL-TK vector (Eurofins Genomics, Ebersberg). The 3′UTR sequences of

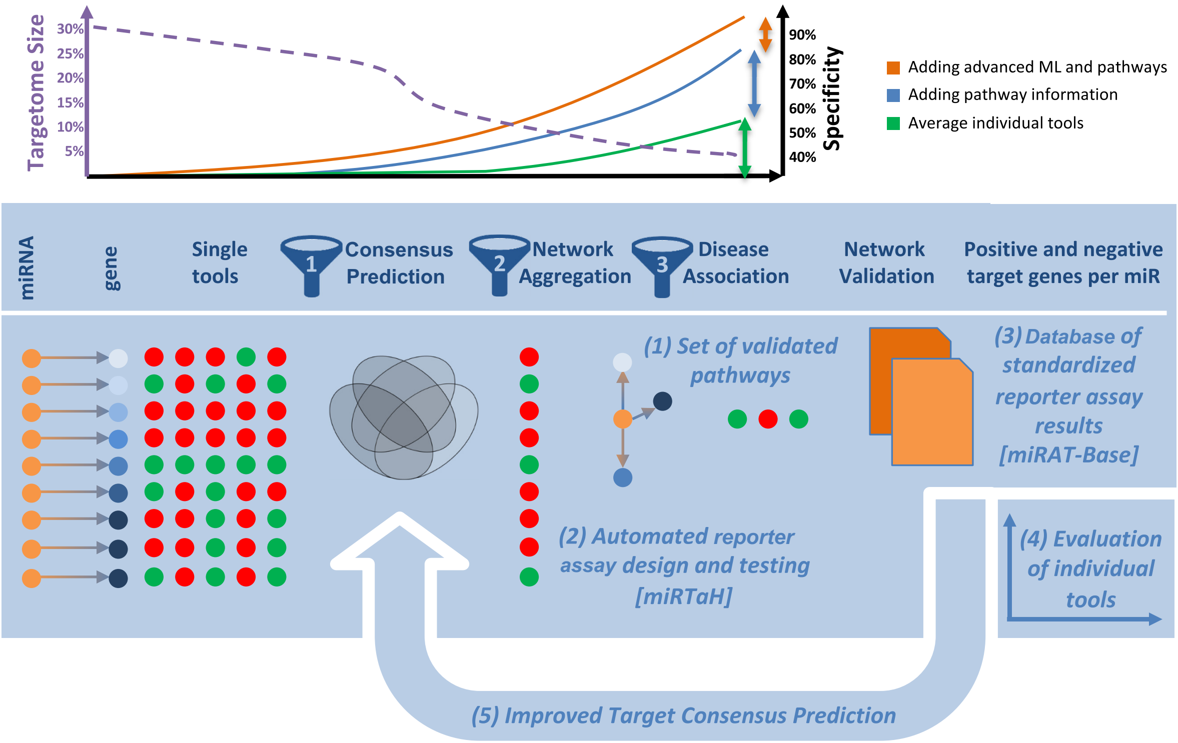

Targetome Size
30%
25%
20%
15%
10%
5%
90%
80%
70%
60%
50%
40%
Specificity
Adding advanced ML and pathways
Adding pathway information
Average individual tools
miRNA
gene
Single tools
1 Consensus Prediction
2 Network Aggregation
3 Disease Association
Network Validation
Positive and negative target genes per miR
(1) Set of validated pathways
(3) Database of standardized reporter assay results [miRAT-Base]
(2) Automated reporter assay design and testing [miRTaH]
(4) Evaluation of individual tools
(5) Improved Target Consensus Prediction

Figure 1. Study set-up, rational and contribution. The main goals of our study are to demonstrate an orchestrated targeting of miRNAs on specific pathways by experimental means and to provide novel useful resources for the scientific community. Originally, we increased the specificity of miRNA target interactions by in-silico approaches alone (green curve and green vertical arrow). By combining improved target selection strategies, we provide evidence for a higher specificity and validation rate in this study (blue curve and blue vertical arrow). We also provide evidence that in iterative improvements the specificity and validation rate can be further increased by improved target selection using advanced machine learning and pattern recognition techniques (orange curve and orange vertical arrow). Besides the main contribution of validated target pathways (1), our approach includes (2) a novel online tool miRTaH that facilitates reporter assay design at scale and (3) a database of validated pathways as well as positive and negative targets for single miRNAs. Finally, we demonstrate that a standardized target database is a valuable source for (4) evaluating the performance of individual tools and (5) improving target prediction and thus can support the development and evaluation of current and new miRNA target tools.

CREB1_1 mut, CREB1_2 mut, TNFSF14 mut, DNM1L_1 mut, DNM1L_2 mut, AKT2 mut, SMAD7 mut, BMP8B mut, SMAD2_1 mut, SMAD2_2 mut, TGFB2 mut and *EP300 mut*, with mutated binding sites were synthetized and the inserts were cloned into the pMIR-RNL-TK vector. For miR-7-5p target validation, the sequences of the 160 3′UTRs of the PD-related target genes were synthetized and the ∼690 nt long inserts were cloned into the pMIR-RNL-TK vector (BGI, Shenzhen, China).

Cell lines, tissue culture

Lund human mesencephalic (LUHMES) cells were purchased from the American Type Culture Collection (ATCC) and transfected for GFP-expression. The cells were cultured as previously described by Scholz et al. (39) in flasks pre-coated with 50 μg/ml poly-L-ornithin and 1 μg/ml Fibronectin. HEK 293T cells were cultured as described previously (40). SH-SY5Y cells were cultivated in DMEM (Life Technologies GmbH, Darmstadt, Germany) supplemented with 20% fetal bovine serum (Biochrom GmbH, Berlin, Germany), Penicillin (100 U/ml), and streptomycin (100 μg/ml). All cell lines were cultured for less than 3 months after receipt.

Differentiation of LUHMES cells

For differentiation of LUHMES cells towards dopaminergic neurons, cells were cultured in advanced DMEM/F12 (Life Technologies GmbH, Darmstadt, Germany) supplemented with 1% N2-Supplement, 2 mM L-glutamine, 1 mM dibutyryl cAMP, 2 ng/ml GDNF and 1 μg/ml tetracycline. After 48 h, cells were trypsinized and seeded with 7.5 × 10^4 cell/cm^2 in pre-coated flasks.

Neurotoxin treatment and RNA isolation

To induce a PD-like phenotype, LUHMES cells were treated with 10 μM 1-methyl-4-phenylpyridinium (MPP+; Sigma Aldrich, Munich, Germany) 6 days after initiation of differentiation for 48 hours. Control cells were supplemented with H_2O. For RNA-Isolation, cells were lysed by QIAzol Lysis Reagent (Qiagen, Hilden, Germany) and total RNA was isolated using the miRNeasy Mini Kit (Qiagen, Hilden, Germany).

Immunocytochemistry

For immunocytochemistry staining of TH and D2R, LUHMES cells were cultured and seeded on pre-coated 8-well μ-slides (ibidi GmbH, Gräfelfing, Germany) with 7.5 × 10^4 cells/cm^2. Medium was exchanged 48 hours after reseeding. The primary antibodies were diluted in PBS containing 1% bovine serum albumin and incubated at 4°C overnight. TH was stained using a polyclonal rabbit antibody (Cat# ab112, RRID: AB_297840, abcam, Cambridge, UK) and D2R was detected using a goat polyclonal antibody (Cat# ab32349, RRID: AB_2094849, abcam, Cambridge, UK). Images were taken with a Leica TCS SP8 microscope (Leica Microsystems, Wetzlar, Germany) and analyzed using LAS X software (version 3.5.5.19976, Leica Microsystems, Wetzlar, Germany).

miRNA Microarray

miRNA expression profiles after MPP+ treatment in dopaminergic neurons were monitored by using Agilent miRNA Complete Labeling and Hyb Kit as well as Agilent SurePrint G3 Human miRNA 80 × 60K Microarrays (Cat. No. G4872A, miRBase release 21.0, Agilent Technologies, Santa Clara, CA, USA) as described previously (41). The raw microarray data has been deposited at the GEO database (GSE135151).

Western blot

For western blot analysis of JNK3, SMAD2, SMAD7, CREB1, TH, CLOCK, PARK2 and GRIA4 4.5 × 10^5 SH-SY5Y cells were seeded out per well of a six well plate. After 24 hours the cells were transfected either with the Allstars Negative Control (ANC) or with hsa-miR-34a-5p miScript miRNA Mimic (MIMAT0000255: 5′UGGCAGUGUCUUAGCUGGUUGU). For endogenous miR-34a-5p inhibition, cells were transfected with miScript Inhibitor Negative Control or anti-hsa-miR-34a-5p miScript miRNA Inhibitor (MIMAT0000255: 5 ′UG GCAGUGUCUUAGCUGGUUGU). Quantification of the western blots was carried out with Image Lab Software Version 5.2.1 (Bio-Rad Laboratories Inc., Hercules, CA, USA).

Quantitative real-time PCR (qRT-PCR)

qRT-PCR was performed using miScript Primer Assay for hsa-miR-34a-5p, hsa-miR-7-5p, hsa-miR-181a-3p, hsa-miR-134-5p, hsa-miR-129-5p, hsa-miR-129-1-3p, hsa-miR-335-3p, hsa-miR-106b-3p, hsa-miR-412-5p, and Custom miScript Primer for hsa-miR-4284 (Qiagen, Hilden, Germany) and the StepOnePlus Real-Time PCR System (Applied Biosystems, Foster City, United States) following the manufacturer's protocol. RNU6B (Qiagen, Hilden, Germany) served as endogenous control. Statistical significance of differentially expressed miRNAs in MPP+ treated LUHMES as well as miR-34a-5p over-expression was analyzed by paired, two-tailed t-tests.

Automated reporter assay construct generation using miRTaH

To facilitate the bioinformatics aided design of several hundred reporter assays we implemented miRTaH (miRNA Target assay Helper). In brief, miRTaH receives a paired list of miRNAs and genes as input query and searches for known miRNA-target interactions from public databases. Next, seed binding sites for each miRNA in the corresponding target gene 3′UTRs are searched. For a list of selected pairs, the 3′UTR sequences are displayed along with the detected miRNA binding sites and potential cut sites of restriction enzymes. Long sequences can be automatically split into any number of chunks, which then can be processed independently. Finally, the tool generates a report of the generated sequence inserts to be synthesized and cloned into reporter plasmids. As organisms, our web service supports *H. sapiens* and *M. musculus*. miRTaH is freely available online (https://www.ccb.uni-saarland.de/mirtah). Further descriptions on the tool are available from the supplemental materials.

miRATBase—a database for validated targets and target pathways of miRNAs

To make the validated targets and target pathways accessible we implemented a data warehouse termed miRNA Reporter Assay Database (miRATBase). In this data warehouse we store for each miRNA the validated target pathways and the positive and negative target data sets. miRATBase is freely available online (https://www.ccb.uni-saarland.de/miratbase). In its current release, miRATBase contains over 500 target associations for four miRNAs. For each entry we also link to miRTarbase (26), miRBase (42), miRCarta (43) and MirGeneDB (44).

MiRNA target prediction

Consensus lists of predicted miRNA targets were obtained using the online interface of miRWalk 2.0 (16). The prediction tools in addition to miRWalk comprise microT v4, miRanda, mirBridge, miRDB, miRMap, miRNAMap, PicTar2, PITA, RNA22, RNAhybrid and TargetScan (45–55). Target transcripts were sorted by the number of algorithms predicting a target and aggregated on the gene level for all entries surpassing the applied cut-offs. For TargetScan the version used during study conception and implementation (6.2) was benchmarked to the currently most recent version 7.2. To this end, all miRNA targets showing a conserved and a non-conserved target site were downloaded from the TargetScan website and processed in the same manner as the targets from version 6.2. Further, aggregated predictions have been extracted from the recent mirDIP release 4.1 (56). Specifically, we made use of 25 tools in the low, medium and high stringency set-up. The final list of evaluated tools thus comprises 88 (25 × 3 + 13) prediction tools with different stringencies.

Statistical analysis

Analysis of microarray data was performed with GeneSpring (version 14.9, Agilent Technologies, Santa Clara, CA, USA). Statistical analysis of qRT-PCR and western blots was performed with Prism7.04 (GraphPad Software, La Jolla, USA) applying paired, two tailed t-tests. Quantification of the western blots was carried out with Image Lab Software Version 5.2.1 (Bio-Rad Laboratories Inc., Hercules, California, USA). Statistical analysis, including evaluation of the automated dual luciferase reporter assays, was performed with R version 3.6.3 applying two-tailed, one-sample t-tests. Heatmaps were generated using the pheatmap R package while all remaining plots were compiled with the ggplot2, cowplot and RColorBrewer packages. The association mining of predicted and validated targets was performed using the *apriori* function of the arules package. For data handling and transformations, the R packages tidyr, dplyr, stringr, data.table and openxlsx were utilized. To test the hypothesis whether 3′UTR lengths systematically influence the results we computed a ratio for each gene using the long and short assay RLUs and performed a one-sample, two-sided Student's *t*-test while setting μ equal to 1.

RESULTS

Overview on HiTmIR: a novel pipeline for validating target pathways of single miRNAs

Our HiTmIR protocol, which was applied to two miRNAs, consists of three computational filters to increase the specificity of the target prediction and to reduce the size of the predicted targetome stepwise, followed by one experimental step (Figure 2A). The first computational filter includes a consensus target prediction (16). We then performed an over-representation analysis using GeneTrail2 (27) to identify enriched target pathways. Third, we added the disease association to the pathway information. Based on the significant categories, we built a consensus target gene set to narrow the experimental search space. A novel web-service supports the design of reporter constructs that are cloned into target plasmids and subjected to systematic experimental testing. To this end, a liquid handling system was programmed to perform an automated luciferase reporter assay in a 96-well format containing the commercially obtained constructs. This pipeline allows to detect and validate complete pathways for single miRNAs, which we exemplify for miR-34a-5p and miR-7-5p. The validated target pathways as well as the positive and negative targets are stored in a data warehouse, miRATBase, a resource for testing and evaluating new target prediction tools.

Selecting microRNAs implicated in aging-related diseases to be screened with HiTmIR

To demonstrate the performance of HiTmIR we selected PD as role model. To further elucidate the role of miRNAs in PD, we differentiated lund human mesencephalic (LUHMES) cells to dopaminergic neurons and subsequently induced a PD-like phenotype using the neurotoxin MPP+ (1-methyl-4-phenylpyridinium). We verified the dopaminergic phenotype after differentiation by immunocytochemistry using tyrosine hydroxylase (TH) in combination with D2 receptor (D2R) as markers for dopaminergic neurons (Figure 2B and C). We analyzed four replicates each after stimulation with MPP+ and four according controls without MPP+ stimulation and identified 686 expressed miRNAs by genome-wide miRNA expression profiling. Following the stimulation by MPP+, we found 13 significantly deregulated miRNAs encompassing four down-regulated miRNAs including miR-7-5p and nine up-regulated miRNAs including miR-34a-5p (adjusted *t*-test *P*-values at an alpha level of 0.05) (Figure 2D and E). We validated the expression changes by qRT-PCR for 10 selected miRNAs comprising seven of the significantly deregulated miRNAs and three of the miRNAs with high fold-changes. The qRT-PCR analysis confirmed the deregulation for eight miRNAs including an up-regulation of miR-34a-5p and a down-regulation of miR-7-5p (Figure 2F, Supplemental Table S1). Since miR-34a-5p plays a crucial role in cancer and in neuropathologies, we investigated its abundance and dependency on age in blood of patients and controls. Analyzing a collection of 4393 individual blood samples (57), we examined miRNA expression of individuals who were between 30 and 80 years old (Figure 2G). We found a steady increase of miR-34a-5p expression over lifetime ($P < 2.2 \times 10^{-16}$). Since the observations suggest a prominent role of miR-34a-5p and miR-7-5p in neuropathological processes, these miRNAs were selected for systematic target pathway validation using the HiTmIR pipeline.

Three computational filters decrease the predicted targetome size to 1% of the transcriptome

The HiTmIR workflow was designed to start with a sensitive set of potential target genes, increasing the specificity in each of the computational steps (Figure 3A, Supplemental Table S2). One challenge in miRNA target prediction research are enormous sets of target genes for single miRNAs as exemplified for miR-34a-5p (Figure 3B). Seven of the 12 tools predict 20% or more of the transcriptome each. Considering the union of all target prediction algorithms basically the full transcriptome is identified as target for miR-34a-5p while each individual gene is only predicted by 2.4 of the 12 tools on average. The union of predictions thus represents a highly sensitive but very unspecific—and therefore unrealistic—representation of the targetome, calling for a more specific target set. While requiring more complex intersections, the number of targets predicted by a respective number of tools decreases significantly (Figure 3B). Around 75% of targets are already excluded by requiring an intersection of four tools to predict a gene, leaving 5198 target genes. At the same time, each of the genes is predicted on average by 5.2 tools. Still, this set is too unspecific and does likely not represent a reasonable targetome of miR-34a-5p. To add specificity, we next performed a pathway prediction as second filter step. By running an over-representation analysis in GeneTrail2 we detected a significant enrichment of target genes in 4507 pathways and biological processes (Supplemental Table S3). This analysis reduced the target gene set further by 33%. Yet again, the remaining number of 3475 genes likely represents an overestimation of the actual targetome. We then dissected targets enriched for pathways being pivotal for neurological diseases or for biological categories that have been associated with PD as a third filter.

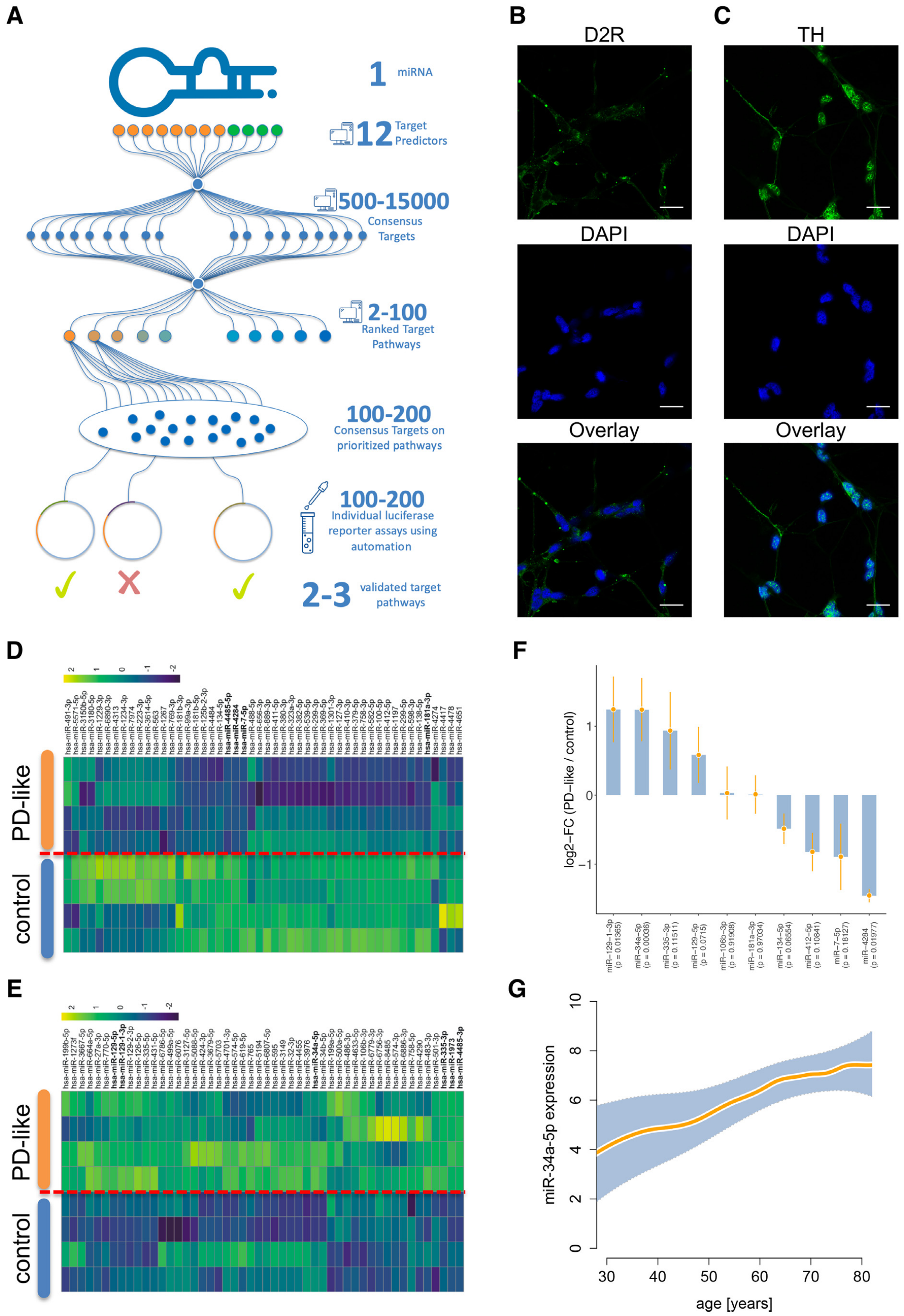


Figure 2. HiTmIR overview and representative selection of miR-34a. (**A**) Combined experimental and computational workflow of HiTmIR. Three computational steps are carried out consecutively before target gene sets are validated by an automated reporter assay. (**B**) Immunocytochemistry of D2R expression in differentiated LUHMES cells. (**C**) Immunocytochemistry of TH expression in differentiated LUHMES cells. (B, C) Expression of dopaminergic markers in differentiated LUHMES cells were analyzed by immunocytochemistry with antibodies against TH and D2R. The nuclei were visualized by DAPI staining. Scale bars are 25 μm. (**D**) Heatmap of the 50 most down-regulated miRNAs in LUHMES cells that were differentiated toward dopaminergic neurons and treated with MPP+ to induce a PD-like phenotype. (**E**) Heatmap of the 50 most up-regulated miRNAs. (D, E) Shown are z-scores of quantile-normalized expression values. (**F**) Validation of microarray results by qRT-PCR of up-regulated and down-regulated miRNAs. Bars present the $\log_2$ fold change between PD-like and controls together with the respective standard deviation. (**G**) Increased expression of miR-34a-5p in the blood of patients, spanning an age range from 20 to 80 years. The orange line shows a smoothed spline with 8 degrees of freedom and the shaded area represents the 95% confidence interval.

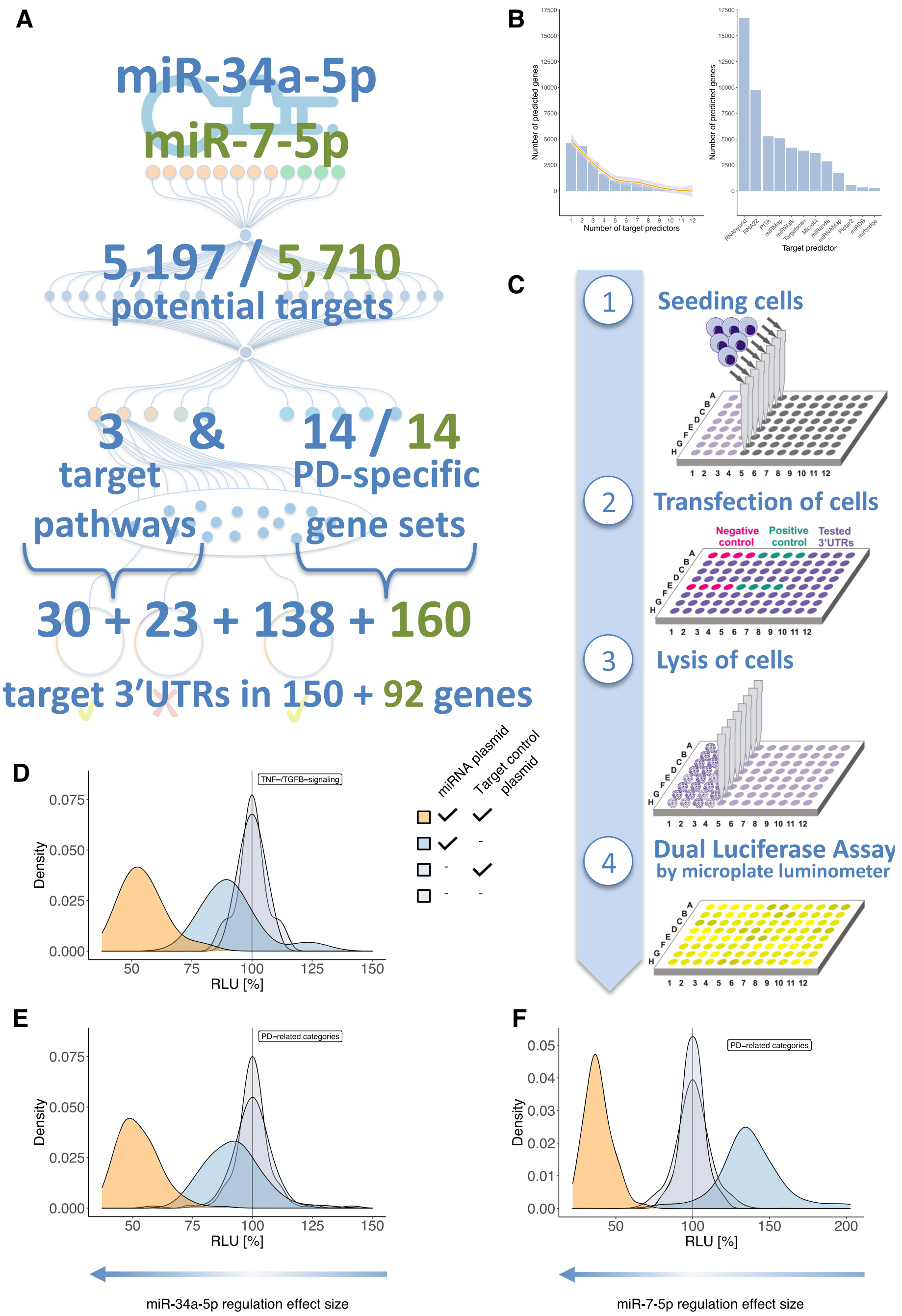


Figure 3. Application of HiTmIR to miR-34a-5p and miR-7–5p. (**A**) Adapted from the workflow in Figure 2A, the actual numbers of the application to miR-34a-5p (blue numbers) and miR-7-5p (green numbers) in the context of PD are shown. (**B**) Histogram of the number of predicted targets dependent on the number of tools predicting this target for miR-34a-5p. Most targets are predicted by one tool only. From the histogram, setting a threshold between three and five tools is a reasonable starting point because large parts of the unspecific hits are already excluded. We then set the initial number of predictions by requiring at least four tools to predict a target. The line represents a smoothed spline. The right-hand side plot of the panel displays the number of target predictions of the 12 individual tools. (**C**) The four experimental steps of the automated reporter assay required to validate target genes in a high-throughput manner. (**D**) Overview on HiTmIR results for miR-34a-5p in the TNF- and TGFB-signaling pathways. (**E**) Overview on HiTmIR results for miR-34a-5p in the PD-related categories. (**F**) Overview on HiTmIR results for miR-7-5p in the PD-related categories. (D–F) The x-axis displays the RLU while the y-axis depicts the density of experimental results. For each set, four curves of experimental transfection designs for targets of miR-34a-5p are shown; two times empty control plasmids (gray), empty miR plasmid + target control 3′UTR (light gray), miR-34a-5p plasmid + empty target control plasmid (blue), and the miR-34a-5p + target control 3′UTR plasmid (orange). The experimental transfection design for miR-7-5p was performed analogously.

Specifically, we found 45 predicted miR-34a-5p target genes in the TNF-pathway and 32 in the TGFB-pathway, both of which have been studied in connection to neurological diseases (Supplemental Table S4). We further investigated categories relevant for PD. Here, GeneTrail2 highlighted a significant enrichment of 274 initially predicted miR-34a-5p targets in 14 PD categories, 10 of which are related to dopamine.

We compared the performance of the pipeline if applied to individual tools. For all 12 tools, we thus performed the exact same pathway analysis as for the consensus prediction (Supplemental Table S5). Here, we observed a higher concordance as compared to the gene-level prediction. On average, the pathways were predicted by 8.2 tools while using the above sketched consensus approach only 5.2 tools predicted a gene ($P < 10^{-5}$). While most of the more complex KEGG pathways were covered by basically all tools (Dopaminergic synapse by all tools, TNF signaling pathway and TGF-beta signaling pathway by 11 tools), some of the smaller yet important Gene Ontology biological processes would have been missed by individual tools (Dopamine metabolism (six tools), Pink/Parkin Mediated Mitophagy (four tools) or dopamine catabolic process (three tools)). These results suggest that incorporating the information of different tools can add to the identification of relevant pathways, especially if these pathways are small.

To identify novel miR-34a-5p targets we relied on the information from the original consensus prediction but excluded all predicted target genes that did not have canonical binding sites and those targets, which were already validated by others according to the miRTarBase (58). Thereby, we obtained a final set of 150 target genes. For some of the predicted target genes, sequence analysis revealed multiple miRNA binding sites within the 3′UTR. To cover longer 3′UTRs that harbor multiple target sites, we split the sequence stretches into different segments to allow for testing of the miRNA effect on each target site separately (Supplemental Table S6). To this end, 3′UTR segments were cloned and separately tested. The respective segments were numbered consecutively starting at the 5′ end, with the number of the corresponding segment added to the plasmid name (as for example pMIR-CLOCK_1 and pMIR-CLOCK_2). In sum, we cloned 30 predicted target 3′UTRs for the TNF-pathway, 23 for the TGF-beta-pathway and 138 for genes associated with PD pathways. In generating the reporter assay constructs (cf. Supplemental Table S6) we recognized the need for a tool that automates this step and implemented the miRNA target assay helper tool miRTaH. The tool, which is freely available as web service (https://www.ccb.uni-saarland.de/mirtah), generates reporter construct sequences for arbitrary miRNA gene target pairs for *H. sapiens* and *M. musculus*. miRTaH supports binding site matching, restriction enzyme site analyses, and selection as well as modification of target sequences. The final sequences can be stored, exchanged, and downloaded easily.

We repeated the above described computational strategy for miR-7-5p. The consensus prediction yielded 5710 unique target genes (Supplemental Table S7). The analogous over-representation analysis returned 4484 pathways and functional categories (Supplemental Table S8). Since miR-7-5p is well described in the context of PD by targeting α-synuclein (34), we focused on the predicted targets for the same set of PD-related categories as screened for miR-34a-5p (Supplemental Table S9). Following the filtering with the same criteria, we generated reporter construct sequences and split 3′UTRs accordingly to a different size of ∼700 nts (Supplemental Table S10). Altogether, 150 and 92 genes were tested by automated dual luciferase assays for miR-34a-5p and miR-7-5-p, respectively.

HiTmIR performance is comparable to manual reporter assays

We tested all 351 selected target gene 3′UTRs using the experimental part of HiTmIR (Figure 3C). To control the validity of the assay, each 96-well plate contained two positive controls in variable wells to exclude positioning-effects. The miR-34a-5p positive controls of the TNF/TGFB-signaling assays showed similar RLU distributions to those of the PD-related categories (Figure 3D and E, Supplemental Table S11). Upon co-transfection with miR-34a-5p, the positive control pMIR-TCRA showed a significant down regulation of the relative luciferase activity (relative light units; RLU) to 54.7% for TNF/TGFB-assays ($P \leq 0.001$) and to 52.5% for PD related assays ($P \leq 0.001$), comparable to previous effects obtained by manual assays (59). Next, we repeated the experiments for miR-7-5p. Following co-transfection of miRNA and target plasmid we also found a clear downshift of the RLU values to a mean of 38.6% (Figure 3F, Supplemental Table S12).

HiTmIR validates 40% of miR-34a-5p targets in TNF-/TGFB-signaling pathways

Out of the 30 tested 3′UTR sequences of the TNF-signaling pathway, 12 (40%) reporter constructs showed a significant RLU down regulation upon co-transfection with miR-34a-5p (Figure 4A, Supplemental Table S13). For TGFB-signaling, 9 of 23 (39%) tested target 3′UTRs showed a significant RLU reduction (Figure 4B). To verify the direct binding of miR-34a-5p to its predicted target sites, we mutated the binding sites and performed comparative HiTmIR experiments between the wild type constructs and the mutated reporter vectors (Figure 4C and D, Supplemental Table S14). For each signaling pathway, we chose six positively tested target gene segments. In sum, we tested CREB1_1, CREB1_2, TNFRSF14, DNM1L_1, DNM1L_2 and AKT2 from TNF-signaling, and SMAD7, BMP8B, TGFB2, SMAD2_1, SMAD2_2 and EP300 from TGFB-signaling. We verified the binding of miR-34a-5p to its predicted target sites for six 3′UTRs showing a significant difference in RLU after mutation. For the non-significant cases, the assay results still suggested a trend to lower RLU values upon a knockout of binding sites.

HiTmIR validates 60% of PD-related pathways for miR-34a-5p and miR-7-5p

We applied the experimental pipeline of HiTmIR to the predicted and PD-related 3′UTR target genes of miR-34a-p and miR-7-5p (Supplemental Tables S13 and S15). Upon

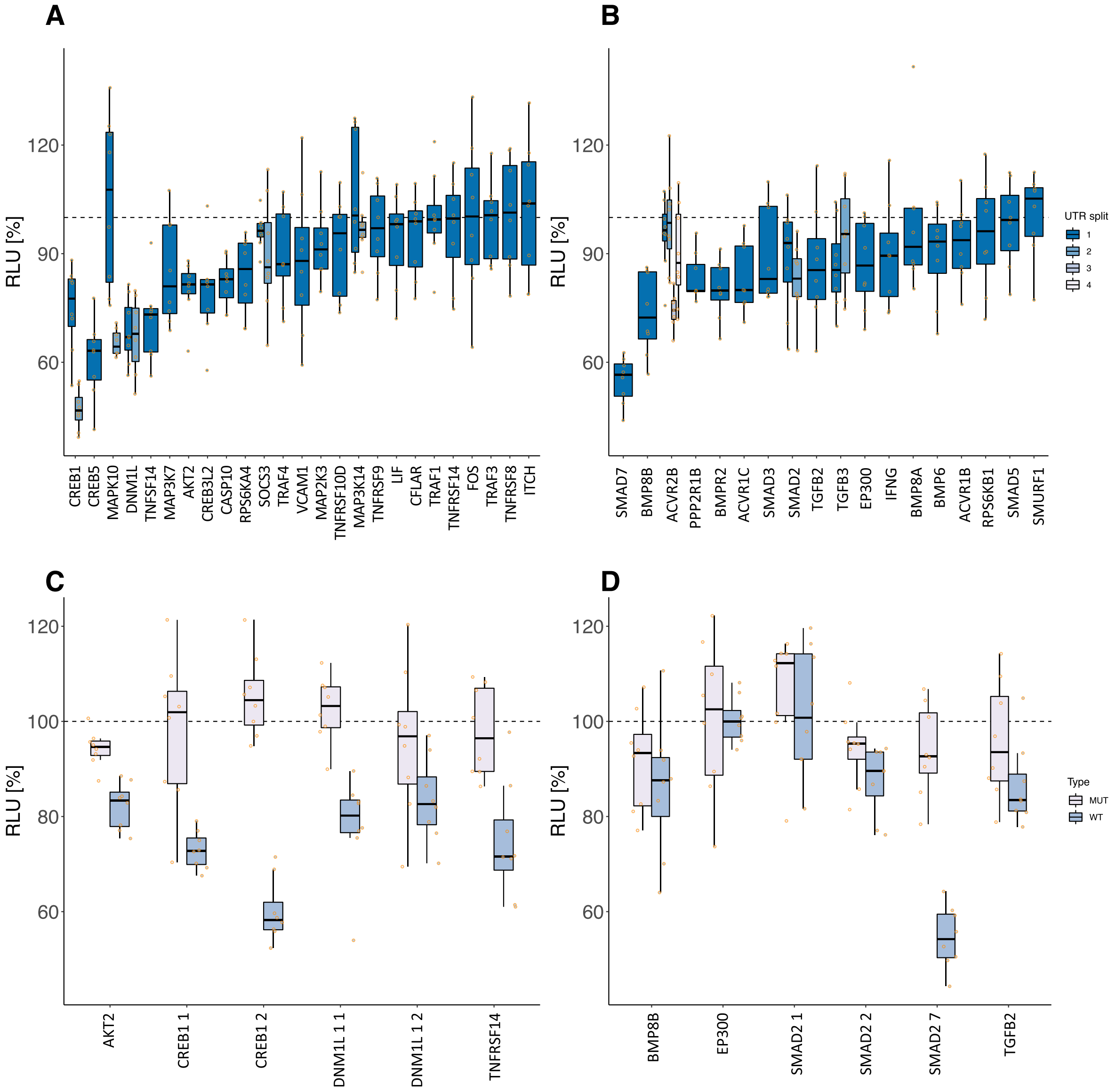


Figure 4. Detailed experimental results of HiTmIR for miR-34a-5p in TNF- and TGFB-signaling. (**A**, **B**) RLU values for eight replicates for each 3′UTR from selected target genes. The dashed line shows the normalized reference level, i.e. the expected level with no effect. (A) Results for pre-selected genes from TNF-signaling. (B) Results for pre-selected genes from TGFB-signaling. (**C**, **D**) RLU values for eight replicates for each wild-type and mutated (binding-site knock-out) 3′UTR from selected target genes. The dashed line shows the normalized reference level, i.e. the expected level with no effect. (C) HiTmIR results for binding-site knockout mutants of selected genes from TNF-signaling pathway. (D) HiTmIR results for binding-site knockout mutants of selected genes from TGFB-signaling pathway.

co-transfection with miR-34a, we detected a significant reduction ($P < 0.05$) of the RLU for 119 target 3′UTRs predicted by at least one algorithm (86.2%). Grouping the plasmids into RLU ranges, we found 51 cases in the range between 33% (KIF5C) and 70% (GSK3B_1) (Figure 5A). We observed a less pronounced decrease between 70% and 80% for 28 target 3′UTRs (Figure 5B). We next evaluated how the cut-off for the minimal number of consensus predictions potentially influences the results. Employing the cut-off, which we already used in the TNF-/TGFB-signaling validation, we observed a slight drop of the validation rate to 84.4%. However, only 39 (32.8%) genes that were predicted by at least four algorithms were removed due to non-detectable binding sites as compared to the 235 (68.7%) genes that were predicted by at least one algorithm. These results suggest an inflated false-positive rate for the genes predicted by a small number of tools only.

Of the 160 sequences tested for miR-7-5p, 106 (66.3%) were significant ($P < 0.05$). Mapping the constructs into the ranges of mean RLUs we only observed 24 targets under

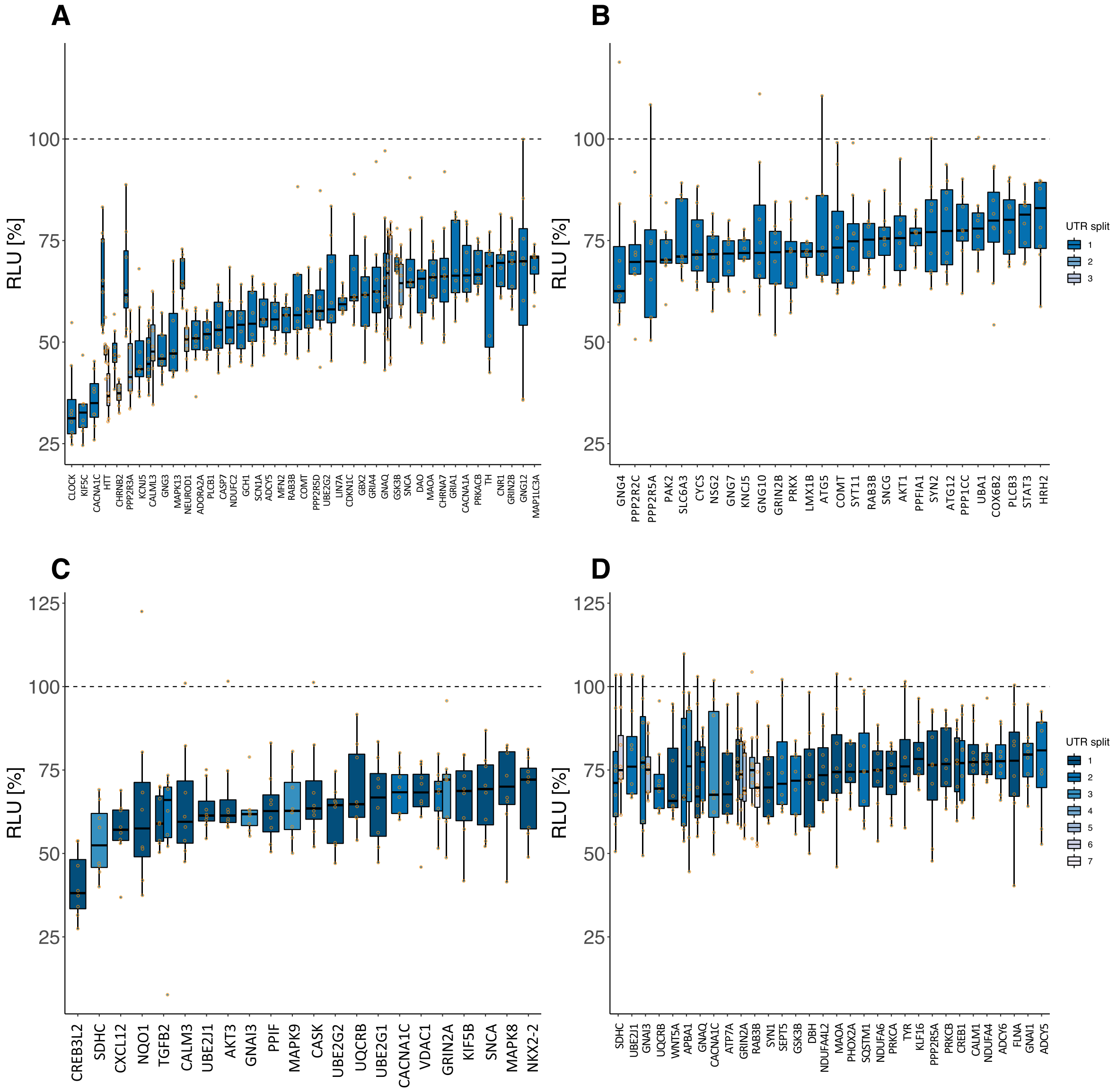


Figure 5. Detailed experimental results of HiTmIR for miR-34a-5p and miR-7–5p in PD-related categories. (A–D) RLU values for eight replicates for each 3′UTR from selected target genes. The dashed line shows the normalized reference level, i.e. the expected level with no effect. (**A**) Results for miR-34a-5p in the PD-related gene sets. Shown are the genes for which mean RLU was less or equal than 70%. (**B**) Analogous to (A) but with mean RLU between 70% and 80%. (**C**) Results for miR-7-5p in the PD-related gene sets. Shown are the genes for which mean RLU was less or equal than 70%. (**D**) Analogous to (C) but with mean RLU between 70% and 80%.

70% (Figure 5C) and 40 targets (Figure 5D) of moderate reduction. These results suggest the validation rate of HiTmIR to primarily depend on the chosen cut-offs as well as the miRNAs under investigation. To elaborate on the relation between high validation rates and the chosen cut-off (standard) parameters per miRNA, we enumerated a set of thresholds for both the minimum mean RLU and the minimum P-value cut-offs and computed the corresponding validation rates (Supplemental Table S16). We found that even with permissive cut-offs ($P < 0.005$ & mean RLU < 80%) the validation rates for the PD-related target sets of miR-34a-5p and miR-7-5p remained competitive with 55% and 35%, respectively. After showing a significant decrease of target expression upon miRNA transfection, we next asked whether the protein expression levels are decreased accordingly.

miR-34a-5p effects target protein expression in SH-SY5Y cells

To investigate the effects of miR-34a-5p targeting on the endogenous protein levels, SH-SY5Y cells were transfected by miR-34a-5p mimics or by ANC as a non-targeting control. We confirmed the over-expression of miR-34a-5p in the transfected SH-SY5Y cells by qRT-PCR (Supplemental Table S17). We next analyzed the endogenous protein levels of JNK3, SMAD7, SMAD2, CREB1, TH, CLOCK, GRIA4 and PARK2 each in three independent experiments by western blotting using specific antibodies (Supplemental Table S18). We observed significantly reduced endogenous protein levels for all tested proteins (Figure 6A–F) ranging from 46% for CREB1 ($0.001 \leq P\text{-value} \leq 0.01$) to 76% for CLOCK ($P\text{-value} \leq 0.05$) (Figure 6G). To further validate miR-34a-5p endogenous targeting, we transfected SH-SY5Y cells with miR-34a-5p inhibitor or an inhibitor control and analyzed the endogenous protein levels of JNK3, SMAD7, SMAD2, CREB1, TH, CLOCK, GRIA4 and PARK2 each in three independent experiments (Supplemental Table S19). In line with the previous observations, we found significantly induced endogenous protein levels for all of the tested proteins ranging from 118% for TH ($P\text{-value} \leq 0.05$) to 163% for CLOCK ($0.001 \leq P\text{-value} \leq 0.01$) (Figure 7).

Variation in cloned 3′UTR lengths does not lead to a systematic bias

Since the validation rates of HiTmIR varied between miR-34a-5p and miR-7-5p, we asked whether this is confounded by the fact that 3′UTR splits of varying lengths were transfected. As independent control experiments we selected nine target 3′UTRs of miR-34a-5p and created reporter constructs containing the full-length 3′UTR sequence. The full-length 3′UTR sequences (∼991 nts) were approximately two times the length of the shorter sequence chunks (∼477 nts) (Supplemental Table S20). Although several cases could be identified where the shorter 3′UTR sequence showed either a better or worse mean RLU, these differences were not significant on the overall distribution ($P = 0.9962$, cf. Materials and Methods). As a conclusion, the length of the 3′UTR reporter constructs does not significantly skew the distribution of RLU values obtained, as long as the technically upper limit (∼1500 nts) is not surpassed.

Evaluating the performance of single tools toward a more accurate consensus prediction

By design, the HiTmIR system facilitates validation of miRNA targets that are predicted and prioritized by *in silico* methods. In turn, it does not only provide a set of validated target pathways but also positive and negative sets of targets for miRNAs. These can be used to evaluate the performance of individual target predictors, utilized to test new individual tools, or used to evaluate consensus prediction. First, we calculated the performance of the individual tools that were originally contained in the target gene selection step to determine whether and how performance varies between the tools (Figure 8A). Our results suggest one set of tools (mirbridge, miRDB, miRNAMap and Pictar2) to be very specific. While this specificity is on a level we are seeking for, it here comes at the price of a sensitivity of only 9%. On the other extreme, RNAhybrid shows a sensitivity of 99.4% but also zero specicitity on our data set. As previously suggested, TargetScan (6.2) and miRanda show a well-balanced specificity and sensitivity. The only other tool that performs similarly well is MicroT v4. However, it is in the nature of successful tools that they are constantly improved. Therefore, we evaluated more recent programs (56). Altogether, 25 tools were tested and most notably for these tools low (Figure 8B and C), medium (Figure 8D and E) and high (Figure 8F and G) confidence sets of targets were acquired to evaluate the performance. Additionally, we included the 12 original tools and TargetScan 7.2. In total we evaluated 88 tools at varying levels of prediction stringency. For each of the tools, we computed the specificity, sensitivity, balanced accuracy, and other measures such as precision, recall, and the F1 score (Supplemental Table S21). As expected, the number of predicted targets generally decreases with stringency increasing. Still, the most stringent sets yield targetome sizes over 20% of the transcriptome. The high confidence set retained a sensitivity, specificity and balanced accuracy of 47%, 60% and 53%. The medium confidence set 39%, 67% and 53%, respectively. The low confidence set yielded 39%, 68% and 53%, almost identical to the medium confidence set. Most importantly, the original set we used reached 46%, 58% and 52% sensitivity, specificity and balanced accuracy, similar to the high confidence set of mirDIP (Figure 8H). The most remarkable difference between the four groups of tools was the increased sensitivity of the high confidence sets, at the cost of the lowest specificity. Of note, there was no tool that clearly outperformed all others, i.e. reaching exceptional specificity and sensitivity. The best-balanced accuracies, exceeding values of 60%, were reached for microrna.org, miRDB, miRanda and TargetScan (7.2).

We then evaluated how an updated algorithm improved the results on the example of TargetScan and compared version 6.2 (the available version when we originally implemented HiTmIR) with the most recent version 7.2. We specifically asked whether a tool update has an impact on single target genes and on the validation success rate. With respect to the original gene sets we observed an overlap of 3384 target genes, for which the newer version had an additional 1000 targets while 444 former targets were not predicted anymore. Most intriguingly, the pathway prediction was 100% concordant between TargetScan 6.2 and TargetScan 7.2 (Supplemental Table S5). In predicting more targets, we might expect also an increased false positive rate but for the genes involved in our study we observed three more true positives and two more true negative genes. For TargetScan 6.2 we computed 124 TP, 32 TN, 32 FP and 54 FN. For TargetScan 7.2 the numbers slightly changed to 126 TP (+2), 33 TN (+1), 31 FP (–1) and 52 FN (–2). The balanced accuracy improved from version 6.2 (59.8%) to 7.2 (61.2%) by 1.4% and in a non-significant manner ($P > 0.05$). Although the overall improvement is statistically not significant, the data nonetheless indicate that advancing individual target tools can improve the accuracy further. The varying performance of the single tools and limitations in consensus approaches as applied in our study also motivates the question whether the obtained wet-lab results in turn can be used to rank the prediction tools used in the first step. To this end, we concatenated the predictions of the 12 tools for miR-34a-5p and miR-7-5p to create a binary matrix. Next, we filtered for the combination of miRNA and validated targets and added a binary response vector (1 = validated,

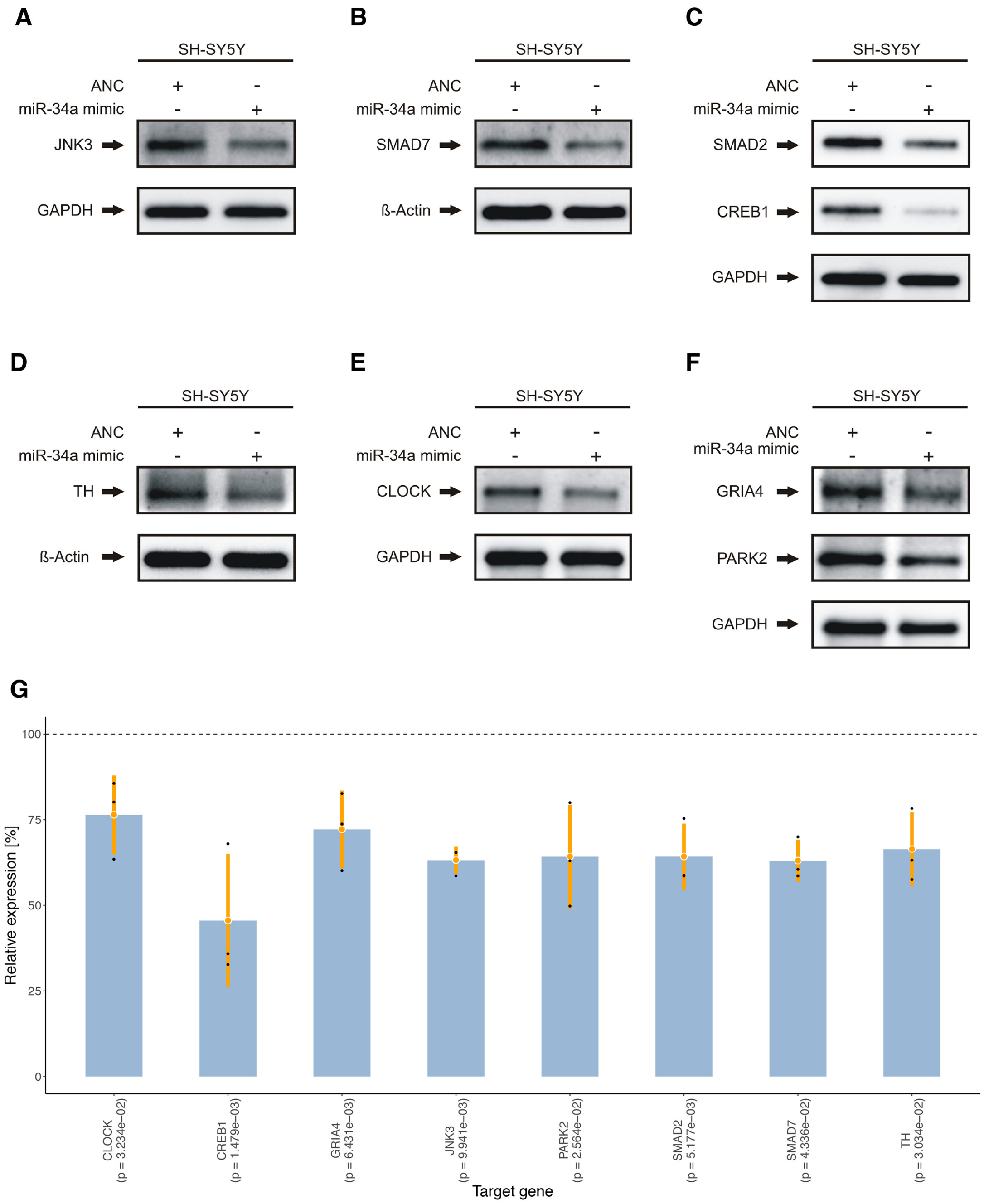


Figure 6. Western blot analysis of JNK3, SMAD7, SMAD2, CREB1, TH, CLOCK, PARK2 and GRIA4 in miR-34a-5p over-expressing cells. SH-SY5Y cells were transfected either with ANC or miR-34a-5p mimic. Forty-eight hours after transfection, the endogenous protein levels were analyzed by western blotting using specific antibodies against the aforementioned proteins. GAPDH or β-Actin served as loading control. One representative western blot out of three independent experiments is shown, respectively. All three western blots were quantified by densitometry using the Image Lab Software. (**A**) Western blot results for JNK3. (**B**) Western blot results for SMAD7. (**C**) Western blot results for SMAD2 and CREB1. (**D**) Western blot results for TH. (**E**) Western blot results for CLOCK. (**F**) Western blot results for GRIA4 and PARK2. (**G**) Combined expression analysis for genes from (A) to (F) tested by western blot analysis. The y-axis displays the relative expression levels with respect to the ANC (100%, dashed line). Each blue bar represents the triplicates (black dots) of a gene with mean (orange dot) and a range of two times the standard deviation (orange lines). *P*-values shown in parenthesis were computed using two-tailored, paired Student's *t*-tests.

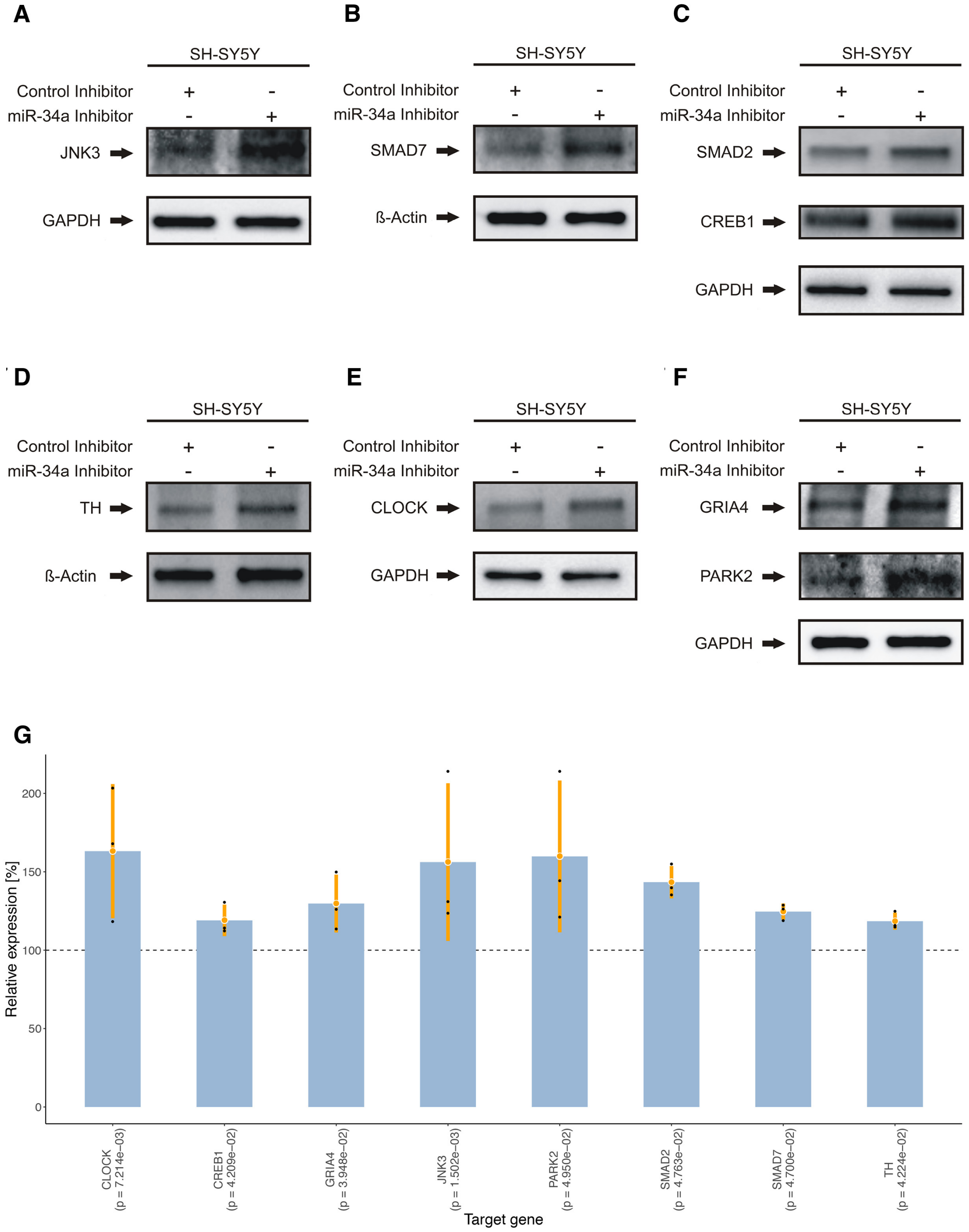


Figure 7. Western blot analysis of JNK3, SMAD7, SMAD2, CREB1, TH, CLOCK, PARK2 and GRIA4 in miR-34a-5p inhibitor transfected cells. SH-SY5Y cells were transfected either with inhibitor control or miR-34a-5p inhibitor. Forty-eight hours after transfection, the endogenous protein levels were analyzed by western blotting using specific antibodies against the aforementioned proteins. GAPDH or β-Actin served as loading control. One representative western blot out of three independent experiments is shown, respectively. All three western blots were quantified by densitometry using the Image Lab Software. (**A**) Western blot results for JNK3. (**B**) Western blot results for SMAD7. (**C**) Western blot results for SMAD2 and CREB1. (**D**) Western blot results for TH. (**E**) Western blot results for CLOCK. (**F**) Western blot results for GRIA4 and PARK2. (**G**) Combined expression analysis for genes from (A) to (F) tested by western blot analysis. The y-axis displays the relative expression levels with respect to the control inhibitor (100%, dashed line). Each blue bar represents the triplicates (black dots) of a gene with mean (orange dot) and a range of two times the standard deviation (orange lines). *P*-values shown in parenthesis were computed using two-tailored, paired Student's *t*-tests.

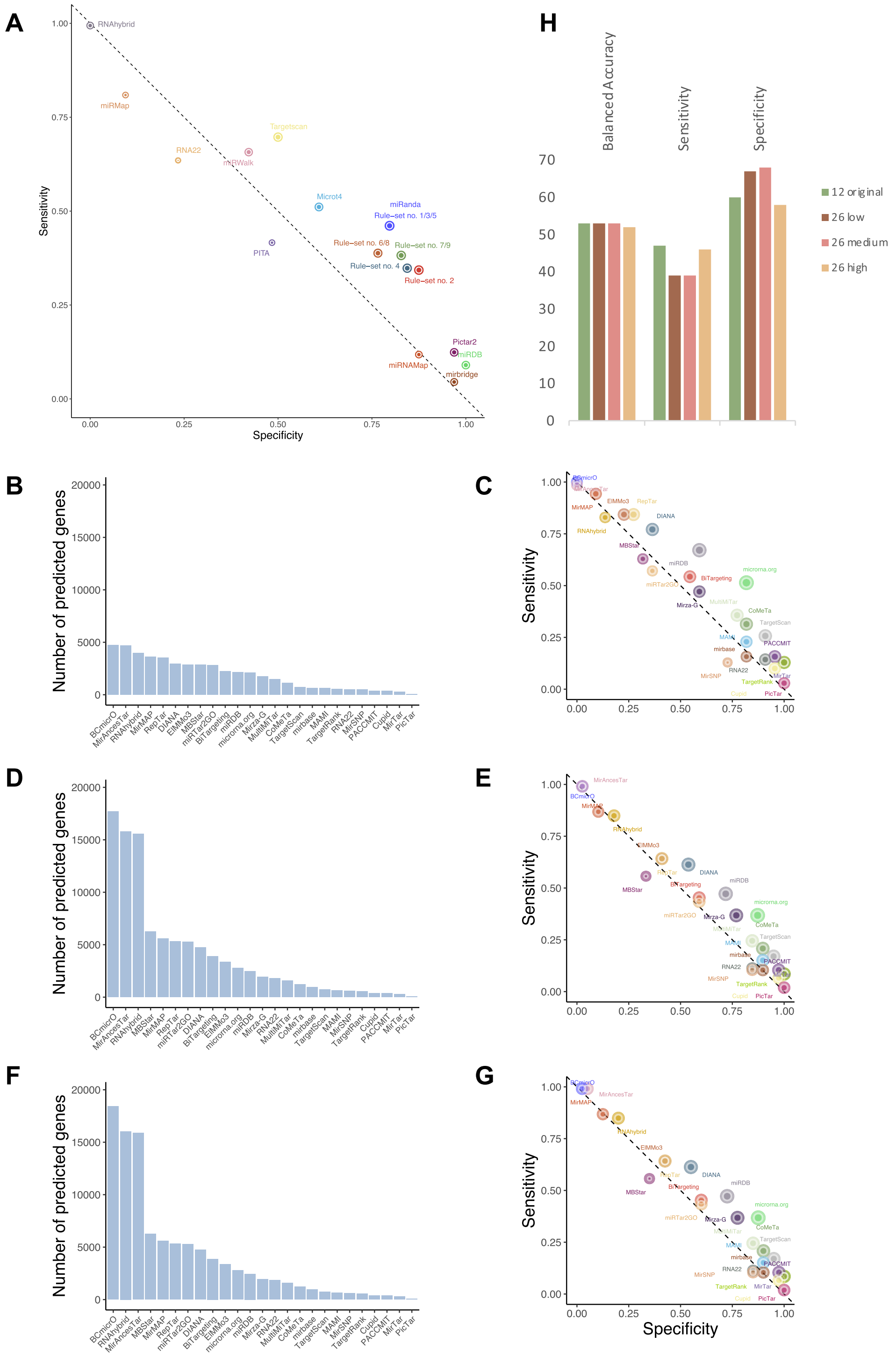


Figure 8. Performance evaluation of individual tools and association rules. (**A**) The scatter plot shows the specificity and sensitivity of the 12 individual tools and the association rules. The point size of the tools and rules correspond to the balanced accuracy. (**B**) Targetome sizes for the most stringent parameters for the new set of tools. (**C**) Specificity and sensitivity of the most stringent parameter set for the new set of tools. (**D**) Targetome sizes for the medium stringent parameters for the new set of tools. (**E**) Specificity and sensitivity of the medium stringent parameter set for the new set of tools. (**F**) Targetome sizes for the least stringent parameters for the new set of tools. (**G**) Specificity and sensitivity of the least stringent parameter set for the new set of tools. (**H**) Balanced accuracy, specificity, and sensitivity for the four tool groups presented in A, C, E, G.

0 = not validated) using the standard cut-off ($P < 0.05$) on the experimental HiTmIR results. Based on an association mining procedure, we searched for a set of rules with high confidence to indicate tools or combinations of such, which are most informative towards the outcome vector. After setting a stringent cut-off for the confidence (≥80%) and a moderate level for the minimal support (≥25%), we computed nine rules (sets of tools) that could help to improve the validation rate in a retrospective manner (Supplemental Table S22). For example, the rule to combine the predictions of miRanda and TargetScan has the largest effect on the validation rate. These results suggest that several combinations of the tools incorporated in our pipeline give a better consensus prediction. Also, this means that the likelihood of a validation to turn out positively is higher than for any other single tool or combination of such. By contrast, negating the binary values of the outcome vector and repeating the association analysis did not yield any signature with high confidence (≥0.4) or support (≥0.3). This shows that non-validated targets are not predicted systematically by any subset of tools. We recommend to potential HiTmIR users to compare the global consensus prediction with the predictions obtained from the derived signatures of tools.

DISCUSSION

With millions of theoretically possible interactions between miRNAs and mRNAs the known human miRNA targetome is far from being complete. Thus, novel methods combining high-throughput experimental and computational methods are in great demand to bring the field closer towards a comprehensive characterization of the targeting mechanisms of miRNAs. Although >100 prediction tools have been proposed, performance largely varies and even well performing tools typically report between several hundred and many thousand targets per miRNA (60). In the light of an expected low *a priori* likelihood of a miRNA targeting a gene, the specificity is of crucial importance. Considering a scenario with a low *a priori* likelihood and a specificity below 80%, the positive predictive values gets extremely low. To partially address this issue, consensus predictions of multiple predictors were used to further sharpen the set of predicted genes. Nonetheless, the methodological similarity of the approaches and their feature sets certainly influence the effectiveness of this filtering technique, still leading to high number of potential target candidates. Researchers face the situation to validate either a small set of selected candidates using traditional low-throughput techniques like reporter assays or to perform unbiased genome-wide assays that exhibit high levels of noise and complicate down-stream analysis. In addition, recent findings suggest that miRNAs orchestrate entire target pathways, an observation that has been claimed repeatedly, but never systematically been shown (59,61).

Therefore, we developed the novel HiTmIR pipeline, specifically designed to close the gap by mapping predicted targets to enriched pathways. The pipeline allows to rapidly design hundreds of recombinants based on 3′UTR sequences, which are tested using an automated parallel dual luciferase assay system. Our requirements for targets to be predicted by at least four tools followed by the filtering of enriched pathways or gene sets, improves state-of-the-art validation rates.

As for the experimental arm of our strategy, we implemented an automated dual luciferase reporter assay for high-throughput miRNA target gene validation. Although luciferase-based target validation has its inherent limitations, reporter assays provide an important piece of evidence whether a miRNA directly binds to its predicted mRNA target site. Here, we addressed two major limitations of reporter assays. First, cloned target sequences mostly do not represent the entire sequence context of the target site. Second, miRNAs are over-expressed in a non-physiological context (62). Examining the effects of different 3′UTR length on the results of reporter assays, we detected altered RLUs for varying 3′UTR lengths but no systematic bias that significantly influences the overall results. Moreover, we confirmed physiological targeting by miRNA inhibition. Using western blotting on transfected cells, we confirmed miRNA targeting for all of the proteins that were indicated as miR-34a-5p targets by reporter assays. To date, there is no gold-standard method for defining target gene regulation by miRNAs. Other high-throughput approaches like the combination of immunoprecipitation of argonaute (AGO) family members with next-generation sequencing (AGO-HITS-CLIP) do only provide evidence of miRNA-mRNA interaction but do not reflect the functional consequences (63). Comparable, high-throughput approaches that are also based on dual luciferase assays reported a significantly lower conformation rate for positive miRNA–mRNA-interactions (63,64). HiTmIR combines the computational target prediction, pathway analysis, automated reporter construct design as well as automated dual luciferase reporter assay for the identification of miRNA targets within a cellular signaling pathway and yields improved target validation rates.

To demonstrate the performance of HiTmIR we selected miR-34a-5p and miR-7-5p as use cases in the context of PD-related pathways. Besides specific evidence for an altered miRNA expression associated with PD, there is a systemic increase of miR-34a-5p with age correlating with the prevalence of neurodegenerative diseases along the lifespan. Also, the observed down-regulation of miR-7-5p has been previously described to effect α-synuclein and to contribute to neurodegeneration (34). Also in a MPTP induced PD model in mice, this miRNA was reduced (33). For both miRNAs, we showed up-scaled reporter assays to resemble the performance of manually performed experiments. Furthermore, automation allows to test batches of targets under replicable conditions. For TNF- and TGFB-signaling selected from our computational workflow, HiTmIR validated about 40% of target genes for miR-34a-5p. Validation rates were further improved for the PD-related categories, with a mean validation rate of 60% when considering both miRNAs. Moreover, we independently validated many of the targets for miR-34a-5p using binding site knockout assays and western blots with miRNA mimics and inhibitors. We then elaborated to which extent the performance depends on several parameters in the pipeline and argued that it can be miRNA specific. For the sake of simplicity, we calculated the validation rate primarily on a per 3′UTR basis as there is no gold-standard to compute it per gene. According to a technical limitation of reporter assays, several 3′UTRs had to be split into smaller constructs, an auxiliary technique that seems not to cause a systematic bias on the validation rates. Thereby, several justifiable ways exist to aggregate the HiTmIR results to compute a validation rate on the gene-level. For example, a simple rule could be to classify a gene as validated if at least one 3′UTR sequence of that gene is regulated by the chosen miRNA. Using the proposed stringent cut-offs ($P < 0.005$ & mean RLU < 80%) in combination with this rule yields a validation rate of 58.9% for miR-34a-5p and 46.7% for miR-7-5p on the gene-level for the PD-related pathways.

Our computational analysis highlighted TNF- and TGFB-pathways as target sets for miR-34a-5p and further 14 PD-related categories for miR-34a-5p and miR-7-5p. Regulation of different target genes by these miRNAs in the context of PD has been described only for a limited number of genes (30,34,65). Applying our new computational and experimental strategy HiTmIR, we demonstrate a complex regulation of cellular pathways for both miRNAs. This has been broadly claimed, but has never been proven to such an extent, especially in a disease-specific context. Via multiple points of interaction, deregulation of these miRNAs strongly impacts the signaling pathways and likely promotes cell death of dopaminergic neurons. As for example, TNF-signaling and TGFB-signaling regulate crucial processes in the central nervous system including synapse formation, synapse regulation, neurogenesis, regeneration and general maintenance of neuronal cells (66–69). Thus, a reduced TGFB-signaling by miR-34a-5p could promote nigrostriatal degeneration (68). Beyond this, we identified not only several PD-associated target genes for miR-34a-5p and miR-7-5p but also multiple targets that are crucial for dopamine metabolism and signaling. In this context, we identified the tyrosine hydroxylase, which converts L-tyrosine to L-dihydroxyphenylalanine (L-DOPA) and is a key enzyme of the dopamine metabolism as direct target of miR-34a-5p. Loss of TH is found within the striatum in 90% of postmortem samples obtained within a five-year period of diagnosis (70). As for miR-7-5p, which has been described as regulator of α-synuclein, HiTmIR identified key components of the PI3K/AKT signaling pathway like AKT3 and GSK3B as direct target genes. Balanced regulation of this signaling pathway is crucial for neuronal cell proliferation, migration, and plasticity (71). In general, the proposed pipeline allows the identification of a large number of target genes for a single miRNA in several cellular pathways and offers the possibility to discover previously hidden parts of the complex regulation network for conserved miRNAs.

Although some of the work steps of HiTmIR such as the consensus prediction and the validation by reporter assay are already described in the literature, the entire protocol, i.e. the combination of computational and experimental techniques to a systematic pipeline, is novel. With this pipeline, a new web service was developed to facilitate (i) the rapid design of potential reporter plasmid inserts by automating the steps of finding and excluding already validated targets, (ii) the search for all annotated transcripts and 3′UTRs per gene and (iii) the search for canonical binding sites in selected targets in real-time. Moreover, we incorporated functionality to split 3′ UTRs at different user-defined sequence locations and to highlight cut sites of restriction enzymes as well as a list of restriction enzymes without a cut motif in the target. These features were extensively fine-tuned and tested to improve the practical usability for massively parallel reporter assays and to reduce time intensive manual labor as much as possible. To the best of our knowledge there is no comparable free available tool published to date.

We implemented a data warehouse storing validated target pathways as well as positive and negative target gene sets. Especially negative target genes are lacking in the literature. Of 9679 reported target gene associations for *H. sapiens* in the miRTarBase, 9357 (97%) are positive and only 322 (3%) negative. In turn this highlights that negative targets are to a large extent not reported. However, such negative results are essential for developing new target predictors. Another challenge is that reporter assay results in databases such as the miRTarBase often come from heterogenous sources. Each manuscript contained in miRTarBase validates on average 1.6 target genes. This might pose challenges in the training process of individual target prediction programs. Our highly standardized positive and negative data set thus represents a valuable source to train or evaluate miRNA target prediction programs.

To further improve the sensitivity of our approach, it could be useful to include the analysis of synergistic effects due to multiple binding sites in the target 3′UTRs. As further down-stream validation strategy, miRNA target pathways additionally could be examined in a tissue-specific context (72,73). Other future developments include the extension from two miRNAs to a multitude of miRNAs that co-regulate the same signaling cascade in a systemic manner and to consider the dynamics of regulatory processes by exploring quantitative regulatory signals over time. Moreover, the setup of HiTmIR can be broadened to a more holistic approach, e.g. through testing of non-canonical binding sites.

DATA AVAILABILITY

All data shown is freely available. The LUHMES miRNA microarray data has been deposited at GEO using accession ID GSE135151.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by the European Union's Seventh Framework Program for Research, Technological Development and Demonstration [600841] and by the Michael J. Fox Foundation [14446]. The authors declare no competing financial interests.

Author contributions: M.H., L.K., H.L., H.B., A.K., E.M. conceived and designed the experiments. M.H., L.K., K.D., C.D., F.K., T.F., T.K., S.R. performed the experiments. O.K. and F.K. designed and developed the miRTaH web

service under the supervision of A.K. E.A., T.F., F.K. conceptualized and developed the miRATBase data warehouse. M.H., L.K., T.K., F.K., M.K., N.L., C.B. analyzed the data. M.H., L.K., S.W., H.L., H.B., A.K., E.M, F.K. contributed to the writing of the manuscript.

FUNDING

Michael J. Fox foundation [14446]; Internal funds of Saarland University; Instituto de Salud Carlos III [IFI16/00041, MV19/00058 to E.A.]. Funding for open access charge: Michael J Fox Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Ambros,V., Bartel,B., Bartel,D.P., Burge,C.B., Carrington,J.C., Chen,X., Dreyfuss,G., Eddy,S.R., Griffiths-Jones,S., Marshall,M. *et al.* (2003) A uniform system for microRNA annotation. *RNA*, **9**, 277–279.
2. Engels,B.M. and Hutvagner,G. (2006) Principles and effects of microRNA-mediated post-transcriptional gene regulation. *Oncogene*, **25**, 6163–6169.
3. Moretti,F., Thermann,R. and Hentze,M.W. (2010) Mechanism of translational regulation by miR-2 from sites in the 5′ untranslated region or the open reading frame. *RNA*, **16**, 2493–2502.
4. 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.
5. Peng,Y. and Croce,C.M. (2016) The role of MicroRNAs in human cancer. *Signal Transduct. Target. Ther.*, **1**, 15004.
6. 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.
7. Hoss,A.G., Labadorf,A., Beach,T.G., Latourelle,J.C. and Myers,R.H. (2016) microRNA profiles in Parkinson's disease prefrontal cortex. *Front. Aging Neurosci.*, **8**, 36.
8. Tatura,R., Kraus,T., Giese,A., Arzberger,T., Buchholz,M., Hoglinger,G. and Muller,U. (2016) Parkinson's disease: SNCA-, PARK2-, and LRRK2- targeting microRNAs elevated in cingulate gyrus. *Parkinsonism Relat. Disord.*, **33**, 115–121.
9. Pichler,S., Gu,W., Hartl,D., Gasparoni,G., Leidinger,P., Keller,A., Meese,E., Mayhaus,M., Hampel,H. and Riemenschneider,M. (2017) The miRNome of Alzheimer's disease: consistent downregulation of the miR-132/212 cluster. *Neurobiol. Aging*, **50**, 167.e1–167.e10.
10. 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.
11. McGeary,S.E., Lin,K.S., Shi,C.Y., Pham,T.M., Bisaria,N., Kelley,G.M. and Bartel,D.P. (2019) The biochemical basis of microRNA targeting efficacy. *Science*, **366**, eaav1741.
12. Hart,M., Kern,F., Backes,C., Rheinheimer,S., Fehlmann,T., Keller,A. and Meese,E. (2018) The deterministic role of 5-mers in microRNA-gene targeting. *RNA Biol.*, **15**, 819–825.
13. Agarwal,V., Bell,G.W., Nam,J.W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, e05005.
14. Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.
15. John,B., Enright,A.J., Aravin,A., Tuschl,T., Sander,C. and Marks,D.S. (2004) Human MicroRNA targets. *PLoS Biol.*, **2**, e363.
16. Dweep,H. and Gretz,N. (2015) miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nat. Methods*, **12**, 697.
17. Krawczyk,B. (2016) Learning from imbalanced data: open challenges and future directions. *Progr. Artif. Intell.*, **5**, 221–232.
18. Blagus,R. and Lusa,L. (2010) Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics*, **11**, 523.
19. Parikh,R., Mathai,A., Parikh,S., Chandra Sekhar,G. and Thomas,R. (2008) Understanding and using sensitivity, specificity and predictive values. *Indian J. Ophthalmol.*, **56**, 45–50.
20. Backes,C., Meese,E., Lenhof,H.P. and Keller,A. (2010) A dictionary on microRNAs and their putative target pathways. *Nucleic Acids Res.*, **38**, 4476–4486.
21. Backes,C., Kehl,T., Stockel,D., Fehlmann,T., Schneider,L., Meese,E., Lenhof,H.P. and Keller,A. (2017) miRPathDB: a new dictionary on microRNAs and target pathways. *Nucleic Acids Res.*, **45**, D90–D96.
22. Kehl,T., Kern,F., Backes,C., Fehlmann,T., Stockel,D., Meese,E., Lenhof,H.P. and Keller,A. (2020) miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database. *Nucleic Acids Res.*, **48**, D142–D147.
23. Ritchie,W., Rasko,J.E. and Flamant,S. (2013) MicroRNA target prediction and validation. *Adv. Exp. Med. Biol.*, **774**, 39–53.
24. Clément,T., Salone,V. and Rederstorff,M. (2015) Dual luciferase gene reporter assays to study miRNA function. *Methods Mol. Biol.*, **1296**, 187–198.
25. Sun,G. and Rossi,J.J. (2009) Problems associated with reporter assays in RNAi studies. *RNA Biol*, **6**, 406–411.
26. Huang,H.Y., Lin,Y.C., Li,J., Huang,K.Y., Shrestha,S., Hong,H.C., Tang,Y., Chen,Y.G., Jin,C.N., Yu,Y. *et al.* (2020) miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res.*, **48**, D148–D154.
27. Stöckel,D., Kehl,T., Trampert,P., Schneider,L., Backes,C., Ludwig,N., Gerasch,A., Kaufmann,M., Gessler,M., Graf,N. *et al.* (2016) Multi-omics enrichment analysis using the GeneTrail2 web service. *Bioinformatics*, **32**, 1502–1508.
28. McMillan,K.J., Murray,T.K., Bengoa-Vergniory,N., Cordero-Llana,O., Cooper,J., Buckley,A., Wade-Martins,R., Uney,J.B., O'Neill,M.J., Wong,L.F. *et al.* (2017) Loss of MicroRNA-7 regulation leads to alpha-Synuclein accumulation and dopaminergic neuronal loss in vivo. *Mol. Ther.*, **25**, 2404–2414.
29. Briggs,C.E., Wang,Y., Kong,B., Woo,T.U., Iyer,L.K. and Sonntag,K.C. (2015) Midbrain dopamine neurons in Parkinson's disease exhibit a dysregulated miRNA and target-gene network. *Brain Res.*, **1618**, 111–121.
30. Ba,Q., Cui,C., Wen,L., Feng,S., Zhou,J. and Yang,K. (2015) Schisandrin B shows neuroprotective effect in 6-OHDA-induced Parkinson's disease via inhibiting the negative modulation of miR-34a on Nrf2 pathway. *Biomed. Pharmacother.*, **75**, 165–172.
31. Rostamian Delavar,M., Baghi,M., Safaeinejad,Z., Kiani-Esfahani,A., Ghaedi,K. and Nasr-Esfahani,M.H. (2018) Differential expression of miR-34a, miR-141, and miR-9 in MPP+-treated differentiated PC12 cells as a model of Parkinson's disease. *Gene*, **662**, 54–65.
32. Kim,J., Inoue,K., Ishii,J., Vanti,W.B., Voronov,S.V., Murchison,E., Hannon,G. and Abeliovich,A. (2007) A MicroRNA feedback circuit in midbrain dopamine neurons. *Science*, **317**, 1220–1224.
33. Junn,E., Lee,K.W., Jeong,B.S., Chan,T.W., Im,J.Y. and Mouradian,M.M. (2009) Repression of alpha-synuclein expression and toxicity by microRNA-7. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 13052–13057.
34. Doxakis,E. (2010) Post-transcriptional regulation of alpha-synuclein expression by mir-7 and mir-153. *J. Biol. Chem.*, **285**, 12726–12734.
35. Collaborators,G.B.D.P.s.D. (2018) Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet. Neurol.*, **17**, 939–953.
36. Langston,J.W., Forno,L.S., Tetrud,J., Reeves,A.G., Kaplan,J.A. and Karluk,D. (1999) Evidence of active nerve cell degeneration in the substantia nigra of humans years after 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine exposure. *Ann. Neurol.*, **46**, 598–605.
37. Spillantini,M.G., Schmidt,M.L., Lee,V.M., Trojanowski,J.Q., Jakes,R. and Goedert,M. (1997) Alpha-synuclein in Lewy bodies. *Nature*, **388**, 839–840.
38. Leggio,L., Vivarelli,S., L'Episcopo,F., Tirolo,C., Caniglia,S., Testa,N., Marchetti,B. and Iraci,N. (2017) microRNAs in Parkinson's disease: from pathogenesis to novel diagnostic and therapeutic approaches. *Int. J. Mol. Sci.*, **18**, 2698.
39. Scholz,D., Poltl,D., Genewsky,A., Weng,M., Waldmann,T., Schildknecht,S. and Leist,M. (2011) Rapid, complete and large-scale generation of post-mitotic neurons from the human LUHMES cell line. *J. Neurochem.*, **119**, 957–971.

40. Hart,M., Rheinheimer,S., Leidinger,P., Backes,C., Menegatti,J., Fehlmann,T., Grasser,F., Keller,A. and Meese,E. (2016) Identification of miR-34a-target interactions by a combined network based and experimental approach. *Oncotarget*, **7**, 34288–34299.
41. Ludwig,N., Werner,T.V., Backes,C., Trampert,P., Gessler,M., Keller,A., Lenhof,H.P., Graf,N. and Meese,E. (2016) Combining miRNA and mRNA expression profiles in wilms tumor subtypes. *Int. J. Mol. Sci.*, **17**, 475.
42. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2018) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
43. Backes,C., Fehlmann,T., Kern,F., Kehl,T., Lenhof,H.-P., Meese,E. and Keller,A. (2017) miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res.*, **46**, D160–D167.
44. Fromm,B., Domanska,D., Høye,E., Ovchinnikov,V., Kang,W., Aparicio-Puerta,E., Johansen,M., Flatmark,K., Mathelier,A., Hovig,E. *et al.* (2019) MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res.*, **48**, D132–D141.
45. Maragkakis,M., Vergoulis,T., Alexiou,P., Reczko,M., Plomaritou,K., Gousis,M., Kourtis,K., Koziris,N., Dalamagas,T. and Hatzigeorgiou,A.G. (2011) DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association. *Nucleic Acids Res.*, **39**, W145–W148.
46. Betel,D., Koppal,A., Agius,P., Sander,C. and Leslie,C. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.*, **11**, R90.
47. Tsang,J.S., Ebert,M.S. and van Oudenaarden,A. (2010) Genome-wide dissection of MicroRNA functions and cotargeting networks using gene set signatures. *Mol. Cell*, **38**, 140–153.
48. Wang,X. (2016) Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics*, **32**, 1316–1322.
49. Vejnar,C.E., Blum,M. and Zdobnov,E.M. (2013) miRmap web: comprehensive microRNA target prediction online. *Nucleic Acids Res.*, **41**, W165–W168.
50. Hsu,S.-D., Chu,C.-H., Tsou,A.-P., Chen,S.-J., Chen,H.-C., Hsu,P.W.-C., Wong,Y.-H., Chen,Y.-H., Chen,G.-H. and Huang,H.-D. (2007) miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. *Nucleic Acids Res.*, **36**, D165–D169.
51. Blin,K., Dieterich,C., Wurmus,R., Rajewsky,N., Landthaler,M. and Akalin,A. (2014) DoRiNA 2.0—upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.*, **43**, D160–D167.
52. Kertesz,M., Iovino,N., Unnerstall,U., Gaul,U. and Segal,E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.
53. Loher,P. and Rigoutsos,I. (2012) Interactive exploration of RNA22 microRNA target predictions. *Bioinformatics*, **28**, 3322–3323.
54. Krüger,J. and Rehmsmeier,M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.*, **34**, W451–W454.
55. Garcia,D.M., Baek,D., Shin,C., Bell,G.W., Grimson,A. and Bartel,D.P. (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat. Struct. Mol. Biol.*, **18**, 1139–1146.
56. Tokar,T., Pastrello,C., Rossos,A.E.M., Abovsky,M., Hauschild,A.C., Tsay,M., Lu,R. and Jurisica,I. (2018) mirDIP 4.1-integrative database of human microRNA target predictions. *Nucleic Acids Res.*, **46**, D360–D370.
57. Fehlmann,T., Kahraman,M., Ludwig,N., Backes,C., Galata,V., Keller,V., Geffers,L., Mercaldo,N., Hornung,D., Weis,T. *et al.* (2020) Evaluating the use of circulating MicroRNA profiles for lung cancer detection in symptomatic patients. *JAMA Oncol.*, **6**, 714–723.
58. Chou,C.H., Shrestha,S., Yang,C.D., Chang,N.W., Lin,Y.L., Liao,K.W., Huang,W.C., Sun,T.H., Tu,S.J., Lee,W.H. *et al.* (2018) miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **46**, D296–D302.
59. Diener,C., Hart,M., Alansary,D., Poth,V., Walch-Ruckheim,B., Menegatti,J., Grasser,F., Fehlmann,T., Rheinheimer,S., Niemeyer,B.A. *et al.* (2018) Modulation of intracellular calcium signaling by microRNA-34a-5p. *Cell Death. Dis.*, **9**, 1008.
60. Kern,F., Backes,C., Hirsch,P., Fehlmann,T., Hart,M., Meese,E. and Keller,A. (2020) What's the target: understanding two decades of in silico microRNA-target prediction. *Brief. Bioinform.*, **21**, 1999–2010.
61. Hart,M., Walch-Rückheim,B., Friedmann,K.S., Rheinheimer,S., Tänzer,T., Glombitza,B., Sester,M., Lenhof,H.-P., Hoth,M., Schwarz,E.C. *et al.* (2019) miR-34a: a new player in the regulation of T cell function by modulation of NF-κB signaling. *Cell Death. Dis.*, **10**, 46.
62. Kuhn,D.E., Martin,M.M., Feldman,D.S., Terry,A.V. Jr, Nuovo,G.J. and Elton,T.S. (2008) Experimental validation of miRNA targets. *Methods*, **44**, 47–54.
63. Wolter,J.M., Kotagama,K., Pierre-Bez,A.C., Firago,M. and Mangone,M. (2014) 3′LIFE: a functional assay to detect miRNA targets in high-throughput. *Nucleic Acids Res.*, **42**, e132.
64. Ito,Y., Inoue,A., Seers,T., Hato,Y., Igarashi,A., Toyama,T., Taganov,K.D., Boldin,M.P. and Asahara,H. (2017) Identification of targets of tumor suppressor microRNA-34a using a reporter library system. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 3927–3932.
65. Alural,B., Ozerdem,A., Allmer,J., Genc,K. and Genc,S. (2015) Lithium protects against paraquat neurotoxicity by NRF2 activation and miR-34a inhibition in SH-SY5Y cells. *Front. Cell Neurosci.*, **9**, 209.
66. Montgomery,S.L. and Bowers,W.J. (2012) Tumor necrosis factor-alpha and the roles it plays in homeostatic and degenerative processes within the central nervous system. *J. Neuroimmune Pharmacol.*, **7**, 42–59.
67. Hegarty,S.V., Sullivan,A.M. and O'Keeffe,G.W. (2014) Roles for the TGFbeta superfamily in the development and survival of midbrain dopaminergic neurons. *Mol. Neurobiol.*, **50**, 559–573.
68. Tesseur,I., Nguyen,A., Chang,B., Li,L., Woodling,N.S., Wyss-Coray,T. and Luo,J. (2017) Deficiency in neuronal TGF-beta signaling leads to nigrostriatal degeneration and activation of TGF-beta signaling protects against MPTP neurotoxicity in mice. *J. Neurosci.*, **37**, 4584–4592.
69. Roussa,E., Wiehle,M., Dunker,N., Becker-Katins,S., Oehlke,O. and Krieglstein,K. (2006) Transforming growth factor beta is required for differentiation of mouse mesencephalic progenitors into dopaminergic neurons in vitro and in vivo: ectopic induction in dorsal mesencephalon. *Stem Cells*, **24**, 2120–2129.
70. Kordower,J.H., Olanow,C.W., Dodiya,H.B., Chu,Y., Beach,T.G., Adler,C.H., Halliday,G.M. and Bartus,R.T. (2013) Disease duration and the integrity of the nigrostriatal system in Parkinson's disease. *Brain* **136**, 2419–2431.
71. Jha,S.K., Jha,N.K., Kar,R., Ambasta,R.K. and Kumar,P. (2015) p38 MAPK and PI3K/AKT signalling cascades in Parkinson's disease. *Int. J. Mol. Cell Med.*, **4**, 67–86.
72. 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.
73. Kern,F., Amand,J., Senatorov,I., Isakova,A., Backes,C., Meese,E., Keller,A. and Fehlmann,T. (2020) miRSwitch: detecting microRNA arm shift and switch events. *Nucleic Acids Res.*, **48**, W268–W274.

3.7 *On the lifetime of bioinformatics web services*

Published online 3 December 2020 *Nucleic Acids Research, 2020, Vol. 48, No. 22* ***12523–12533***
doi: 10.1093/nar/gkaa1125

SURVEY AND SUMMARY

On the lifetime of bioinformatics web services

Fabian Kern [1,2,†], Tobias Fehlmann [1,2,†] and Andreas Keller [1,2,3,*]

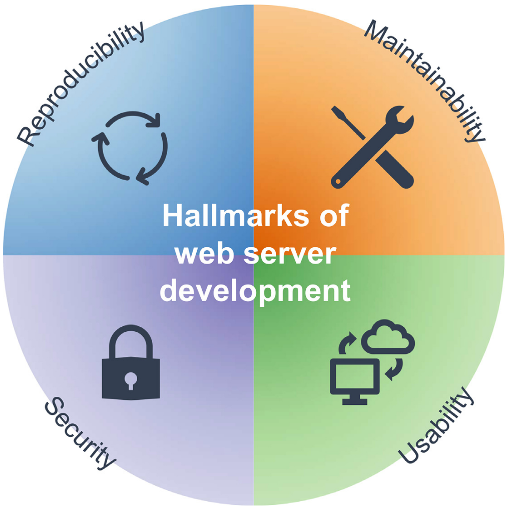
[1]Chair for Clinical Bioinformatics, Saarland University, Saarbrücken 66123, Germany, [2]Center for Bioinformatics, Saarland Informatics Campus, Saarbrücken 66123, Germany and [3]Department of Neurology and Neurological Sciences, Stanford University, Palo Alto, CA 94305, USA

Received July 31, 2020; Revised October 28, 2020; Editorial Decision November 02, 2020; Accepted November 05, 2020

ABSTRACT

Web services are used through all disciplines in life sciences and the online landscape is growing by hundreds of novel servers annually. However, availability varies, and maintenance practices are largely inconsistent. We screened the availability of 2396 web tools published during the past 10 years. All servers were accessed over 133 days and 318 668 index files were stored in a local database. The number of accessible tools almost linearly increases in time with highest availability for 2019 and 2020 (∼90%) and lowest for tools published in 2010 (∼50%). In a 133-day test frame, 31% of tools were always working, 48.4% occasionally and 20.6% never. Consecutive downtimes were typically below 5 days with a median of 1 day, and unevenly distributed over the weekdays. A rescue experiment on 47 tools that were published from 2019 onwards but never accessible showed that 51.1% of the tools could be restored in due time. We found a positive association between the number of citations and the probability of a web server being reachable. We then determined common challenges and formulated categorical recommendations for researchers planning to develop web-based resources. As implication of our study, we propose to develop a repository for automatic API testing and sustainability indexing.

GRAPHICAL ABSTRACT



INTRODUCTION

Scientific web servers and web services are frequently developed to make complex algorithms available to a broad research and user community. They have facilitated substantial contributions to the development of the current research landscape in the life sciences and biomedicine. As one example, the web service to the basic local alignment search tool BLAST, originally published by Altschul in 1990 (1) has become one of the most popular web-based tools in sequence analysis. Also, extensions for protein alignment such as Gapped BLAST and PSI-blast (2) have been made available as web services and are accessed

[*]To whom correspondence should be addressed. Tel: +49 174 1684638; Email: andreas.keller@ccb.uni-saarland.de
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

thousands of times each day. Belonging to one of the most frequently applied tools, it is evident that the web service is carefully maintained and continuously improved (3). Similarly, other successful web services such as STRING are regularly updated and maintained (4–6). Also, the European Bioinformatics Institute (EMBL-EBI) provides access to several essential analysis tools via web services that are regularly updated, extended and maintained in a sustainable manner (7).

Web services have become of such a high relevance that the journal Nucleic Acids Research (NAR) dedicates a whole special issue each year to this topic. Staring in 2003 with 131 of the most widely perceived web services from the years before (8), the annual web server issue has steadily extended its scope and has become a world-renowned resource for peer-reviewed web servers (9–11). The most recent web server issue got as much as 269 proposals of which 79 manuscripts were finally accepted after peer-review, resulting in an overall acceptance rate of 29% (12). These numbers underline the tremendous popularity but show that scientific rigorousness must be assured for online implementations as well. It became a perception that computational biology resources lack persistence and usability (13,14). Following the study by Veretnik *et al.* (13) in 2008 that investigated the availability of all NAR web servers published in the preceding 4 years, Schultheiss *et al.* presented a similar but extended analysis in 2011 (15). They found that of the 927 web servers published in NAR between 2003 and 2009, 72% were still available at their original addresses while 9% were gone offline. The study by Schultheiss *et al.* excels by a survey among all authors and a functionality test of each server providing example data. In 2017, the survey by Wren *et al.* (16) highlighted that ∼27% of URLs from web servers decayed since their original publication and that this is a relatively stable phenomena observed among scientific web tools.

Since these studies had been conducted another 1026 web instances have been published in a total of 11 issues in NAR. Here, we set to provide an up-to-date and comprehensive evaluation of the general availability of web services. Thus, we collected 2727 articles describing 2396 unique tools published by PubMed indexed journals from 2010 onwards and tested their availability over time.

The main goals of the present study are: i) to present a comprehensive analysis of the accessibility of web services; ii) to get insights into the availability dynamics of web services over a longer period of time, e.g. to understand differences across weekdays and to estimate the extent of typical downtimes; iii) to evaluate with an experiment whether and to which extent recent web services can be rescued by contacting the corresponding authors; iv) provide an analysis on the dependency between tool metadata such as service hoster or host country and site availability and v) use the observations to formulate reasons for the observed web server decay. We use this information to derive practical recommendations for web server developers to improve upon security, maintainability and user experience, that should ultimately extend the expected lifetime of web services.

MATERIALS AND METHODS

Literature search

To get a list of web servers we performed a comprehensive PubMed query using the following search term:

(((http://[title/abstract]) or (www.[title/abstract]) or (https://[title/abstract])) and ('web server'[title/abstract] or (web service[title/abstract]) or webserver[title/abstract] or 'web service'[title/abstract] or web-server[title/abstract] or web-service[title/abstract])) and (('2010/01/01'[Date–Publication]: '3000'[Date–Publication])).

The output resulted in 2727 articles. For each article, the abstract and available meta information was downloaded as CSV file and further processed. From the CSV files we extracted the primary web addresses.

Filtering of tools

The list of tools was further processed and filtered. In 2327 cases a single Uniform Resource Identifier (URL) was provided whereas in the remaining cases several URLs could be determined. This includes special cases where different web servers have been included or where a mirror URL to the same endpoint has been provided. Other examples for articles with two URLs comprise a direct link to the tutorial or to related databases. In these cases only the URL to the actual tool was selected. Another 12 tools were removed since they rather described meta analyses instead of web services in the common sense as used in this work. In addition, 37 other tools were removed because the mentioned web servers were to be deployed locally or not originally published in the linked articles. Finally, one tool was excluded since it had been retracted in the meantime (17). As a next step, we curated redirects and iteratively removed duplicated tools.

Download of landing pages

We accessed the web pages by using the *download.file* function of R with the curl method selected. With the parameter -m 30 we restricted the maximum operation time to 30 s and with the -L option up to 50 redirects were allowed.

Filtering of non-working tools

To classify web pages either in reachable or offline we extended the search beyond the typical error messages (e.g. response codes 403, 404, 406, 502 and 503). Screening manually through the non-accessible web pages we identified 44 phrases such as 'Maintenance in progress', 'has been discontinued' or 'Our server is down temporarily'. If one of the determined keywords or phrases could be detected the site was classified as non-accessible. On three non-consecutive days, the number of available tools dropped considerably (to less than 80% compared to the preceding day), potentially to technical issues at national hub nodes or internet service providers. Therefore, the affected daily counts were excluded from downstream analysis.

Additional curation steps for static analysis

After the collection of the long-term availability statistics for all tools we semi-automatically curated the entries by inspecting the URLs provided by *bio.tools* (18), searching for patterns in failed URLs and manually checking for new URLs of unavailable services. In addition, we improved the PubMed provided URLs linking to lab homepages using the address to the corresponding tool explicitly. Furthermore, we also curated tool URLs that are reachable but host different and irrelevant content. Tools that were affected by these steps were then excluded from the long-term analysis. After this final step, a total of 2396 of 2727 tools remained (Supplementary Table S1). Notably, in this set also databases with a limited web service functionality were kept. For the tools that were available but modified their host URL without a suitable scientific publication we provide separate statistics in Supplementary Table S2. Downstream analyses have been carried out using the binary matrix of $p = 2396$ tools (rows) and $n = 133$ days (columns).

Determining hosting providers

We first collected the IP addresses of all tools and retrieved usage information, hosting domain and ISP information from two IP information services, IP2Location.com and IPinfo.io. We then manually checked all non-educational and non-governmental entries for cloud hosting providers. When no IP address could be found no hosting provider was derived.

Determining institutional e-mail addresses

First, e-mail addresses of the corresponding authors were extracted from Web of Science. We then searched the hosting domains in a list of free e-mail provider domains found at https://gist.github.com/okutbay/5b4974b70673dfdcc21c517632c1f984.

Statistical analyses

All analyses have been carried out with R 3.3.2 GUI 1.68 Mavericks build (7288). To evaluate the availability of tools over time, splines from the *smooth.spline* function with 10 degrees of freedom (DF) were used. Pie charts, ridgeline, violin and bar plots were compiled using ggplot2. Clustered heat maps were generated using the *superheat* function in the superheat package. Hypothesis tests (Student's *t*-test, Wilcoxon rank-sum test) were conducted using the R stats implementations.

RESULTS

Study set-up to answer the four research questions

To reach the main goals set we extracted 6727 articles published after 2010 from PubMed. After removing duplicates and false positive hits of our literature search (e.g. meta analyses of tools) 2618 tools remained. After manually curating these (cf. 'Materials and Methods' section), 222 tools were removed, retaining the final set of 2396 articles/tools (Supplementary Table S1). The web addresses of the tools were accessed beginning on the 13 April 2020 for a total of 133 days. During this process 318 668 index pages were downloaded and stored to a local database (Figure 1A).

Static analysis of web services highlights a half-lifetime of 10 years

On 13 April, we completed the first download of all tool landing pages. As a first result 25.7% of the tools were not reachable opposing the 74.3% that were working on that day (Figure 1B). Tracking the number of tools published by year we see a generally increasing trend from ∼200 tools in 2010 to ∼300 tools in 2019 (Figure 1C). These numbers fit well to the overall growth of scientific literature and knowledge (19). With increasing time after publication, we estimate an almost linear decreasing availability of the web services. While tools published in 2019 and 2020 are available to more than 90%, this rate drops to 50% for tools published in 2010 (Figure 1D and E). Considering the source journals, we observe an uneven distribution. As explained in the introduction, the annual web server issue of NAR has become a pivotal resource in the field. Indeed, with over 550 contributions NAR is the leading journal in this regard. However, also Bioinformatics showed a large number of contributions, most likely driven by the application note manuscript category. Two other journals, PLoS One and BMC Bioinformatics reached almost 200 contributions while the remaining ones were distributed among many other journals (Figure 1F). Interestingly, we observed substantial differences in the availability of web servers depending on the journal they were published in. For example, tools published in NAR, Bioinformatics, Scientific Reports or Methods on Molecular Biology had higher long-term availability rates as compared to PLoS One or BMC Bioinformatics (Figure 1G). To further limit the influence of the time variable on these results we repeated the analysis only for the articles published in the past 5 years. Here, the trend of aforementioned differences diminished but was still noticeable (Figure 1H). This first snapshot analysis on the 2396 tools already provides interesting insights on the average lifetime of bioinformatics web services. It is however fair to speculate that these results are influenced by many factors, e.g. the actual weekday when the tools were accessed or seasonal fluctuations. To limit respective effects, we accessed the tools between 13 April and 31 August 2020.

Monitoring over time indicates short downtimes and higher availability toward the mid of the week

We have collected reliable data on the availability over time and first asked whether and how reachability varies between the tools. We found 31% could always be reached, 20.6% could never be reached and 48.4% could be reached at least once (Figure 2A). The shape of the density distribution of the percentage of days on which tools were working basically supports the existence of these three groups. Only few tools were working between 25 and 75% of the tested days (Figure 2B). For the fraction of tools that was neither consistently off- nor online, we computed the duration of consecutive downtimes. The distribution highlights that individual service outage times were rather short with the com-

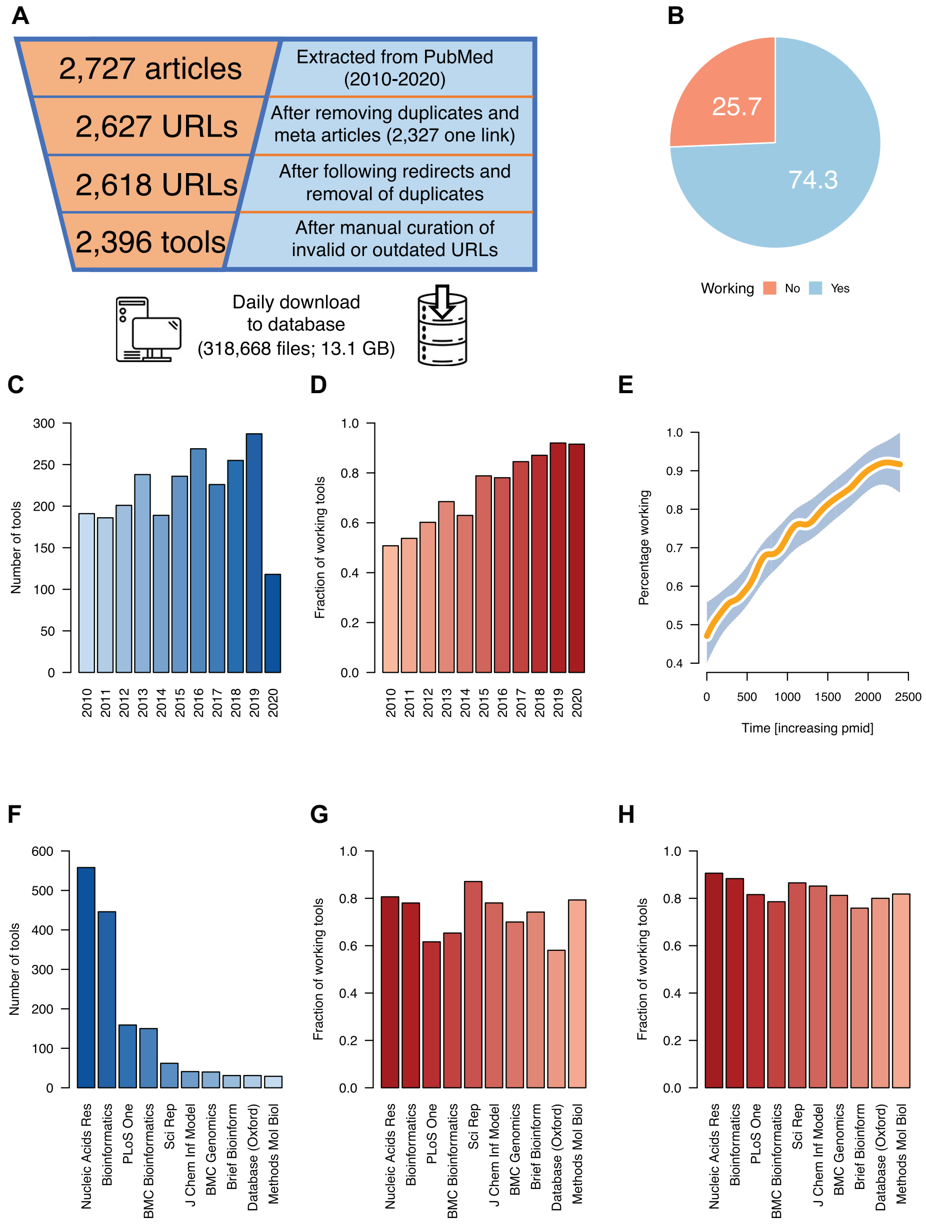


Figure 1. Study set-up and static monitoring. (**A**) Schematic representation of the conducted tool filtering steps. (**B**) Pie chart representing the number of tools that are accessible and not accessible at the start of the observation period. (**C**) Bar chart of the number of tools published by year included in our study. (**D**) Bar chart of the fraction of available tools (snapshot) per publication year. (**E**) Smoothed spline (solid orange line) with surrounding 95% confidence interval (shaded blue area) for the data presented in panel D. (**F**) Number of tools collected per journal. (**G**) Fraction of available tools per journal. (**H**) Available tools per journal restricted to manuscripts published in the past 5 years (2016–2020).

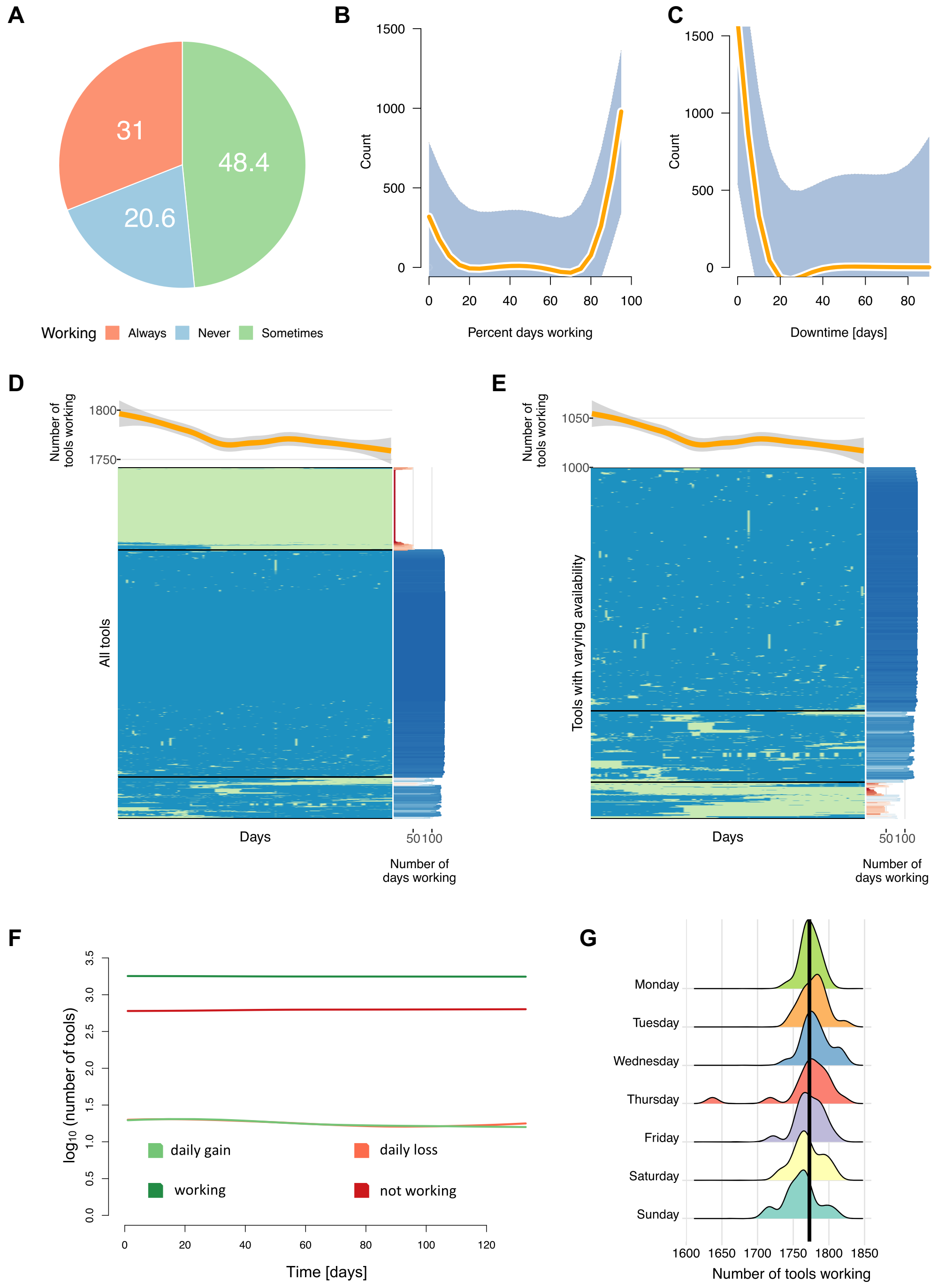


Figure 2. Dynamic monitoring over time. (**A**) Pie chart showing the distribution of tools that were tracked over time into the categories *never available*, *always available* and *sometimes available*. (**B**) Smoothed spline representation of the availability of web servers in percent of days. (**C**) Smoothed spline representation of the observed web server downtime intervals. (**D**) Heat map of the availability matrix for all tools included in the dynamic study. Blue means available, light green not available. The curve on the top represents the smoothed spline representation of the tool availability over time. The histogram on the right shows a bar for each tool proportional to the number of days it was accessible. (**E**) Clustering of those tools that belong to the category of being sometimes available. Notably, this largely corresponds to the middle cluster of panel D but includes also several tools from the other clusters. (**F**) Line chart on the availability categories of tools tracked over time and the trend of daily changes. Toward the end of the observed period we see the green and orange line (lost versus gained per day) diverging. (**G**) Ridgeline plots on the availability of tools per weekday. The solid black vertical line represents the overall mean of tools available per day.

puted median and mean downtime of 1 and 2.9 days, respectively (Figure 2C). A clustering of the tools times days availability matrix confirmed the observations on the general functionality of tools (Figure 2D and E). The heat map indicates three main clusters, one with the tools that work always or almost always, one with the tools that work never or almost never and one smaller cluster in the middle with the tools that show a more heterogenous pattern. On average, 1773 of the 2396 tools (74%) were working per day. The minimal number of 1637 tools was reached on 2 July and the maximal number of 1822 tools on Thursday, 28 April. We observed an almost continuous decrease of web server availability along the test timeframe. Deviations from this expectation could possibly be due to two reasons. First, the primary observation of decreasing tool availability over time has been performed on a 10-year horizon while we tracked only another four months, which still might be too short to observe respective long-term trends. Second, as we describe in the next section, we performed a rescue experiment after 2 weeks, which contributed to a temporarily increased tool availability. If we exclude these cases, we again observe the negative correlation between time since publication and fraction of working tools. In line with these results, we detect similar patterns for daily gains and losses of tools over time, only showing divergence toward the end of the observed period (Figure 2F). As last aspect of the analysis we assessed the dynamics on the distribution across weekdays (Figure 2G). The results suggest a tendency of higher availability of tools toward the mid of the week. On average, the lowest number of tools working was obtained on Sundays (~1761, 73.5%), while on Wednesdays the highest fraction of tools (~1781, 74.4%) was available. Although the differences are percentage-wise small, still an average of additional 20 tools were working on Wednesdays as compared to Sundays with the difference being statistically significant ($P = 0.006501$).

Rescue experiment shows that over 50% of web servers can be brought back to service

Our analysis highlights that even tools published in 2019 and 2020 exist that have lost functionality, some even few weeks after their initial publication. Especially in the light of editorial policies requiring the continued availability over at least several years (e.g. the NAR web server issue states: ‘It is expected that the website will be maintained for at least 5 years’) this observation is unexpected. To exclude likely false positives, i.e. tools that were only down for 1 or 2 days because of maintenance work, we compiled a list of tools published in 2019 and 2020 that did not work over the entire first 2 weeks of the observation period. For the resulting 47 instances we contacted the corresponding authors and asked to restore the functionality of the tool. In 57.4% of the cases we got a reply, leaving 42.6% of the enquiries unanswered (Figure 3A). However, the speed of the replies obtained was remarkable: for all but three cases the first reply was received on the same day. The latest reply occurred 3 days after the initial request and altogether, 96 e-mails were exchanged. Already one day after contacting the corresponding authors, 14 tools (29.8%) were brought back to service (Figure 3B). Although this sum slowly increased over the tracking period, we again detected a small decline in availability for the successfully recovered web servers toward the end.

Frequently cited web services invalidate URLs from scientific publications

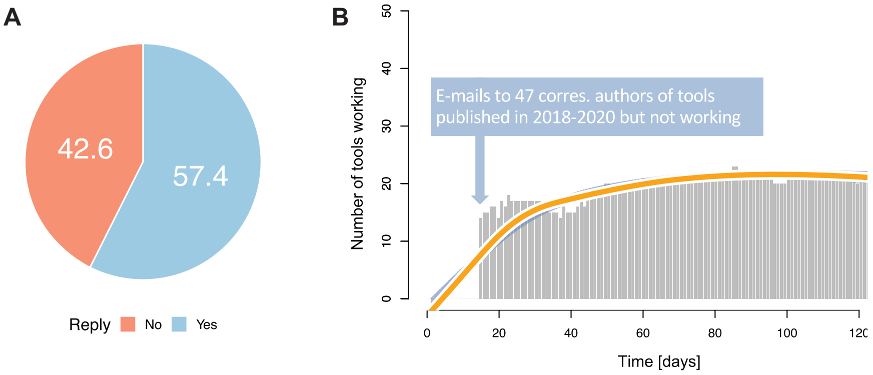
We also tracked which services modified their URL without providing a new link in a scientific publication, i.e. the 185 tools that were removed in our last filtering step (cf. Figure 1A & ‘Materials and Methods’ section). The top-ranking journals mirrored the larger distribution reported before (Figure 4A), however, the average publication year is notably shifted toward the early years considered in the study (Figure 4B). This matches our expectation for services to take several years before a new host URL is released. Nevertheless, we found that for those tools changing the URL offside the scientific literature, a higher citation was obtained on average when they were accessible at least once in our testing frame (Figure 4C), corroborating previous observations (16). We conclude that web server availability and community popularity are robust against sometimes inevitable URL modifications, an observation we largely attribute to the capabilities of modern search engines, which rapidly re-index new websites and their keywords in a few hours or days.

Tool metadata sheds light onto global web server landscape

An intriguing question is whether publication or web server metadata can be used to judge the *a priori* likelihood of a tool to be inaccessible. Therefore, we collected various features for the total 2581 tools investigated (cf. Supplementary Tables S1 and 2). First, the community has built key resources such as *bio.tools* to index and track scientific web servers along their lifetime. Interestingly, 40.5% of the tools considered are contained in *bio.tools* and an overwhelming fraction were accessible (Figure 4D). As a matter of fact, the subset of tools not contained in the service comprises more non-reachable tools, both percentage- and count-wise. By analyzing the host services, we found 71.6% of the tools to be hosted by individual research institutions and another 13.8% managed by cloud services but with overall similar accessibility rates (Figure 4E). Likewise, the distribution of corresponding contact information highlighted most e-mail addresses to be institutional but when comparing fractions the instances with non-institutional addresses are more prone to be unavailable. (Figure 4F). In fact, institutional addresses can be affected by personnel relocation and thus become unavailable, while non-institutional addresses are less likely to change. Lastly, the distribution of host countries matches the global distribution of countries by Gross National Income with the United States, China, and Germany hosting the most scientific web servers, the latter of which is closely followed by India (Figure 4G). Remarkably, many European countries do not list a single web server instance that was inaccessible in our study.

Analysis of impact reveals hallmarks of web server development

We next sought to investigate the relation between web server availability and number of citations for the respective manuscript, similar to the approach of Schultheiss *et al.*



E-mails to 47 corres. authors of tools published in 2018-2020 but not working

Figure 3. Results of rescue experiment. (**A**) Pie chart of the binarized e-mail responses. (**B**) Bar chart representing the number of tools published in 2019 and 2020 that were not working in the first 2 weeks of the tracking period. The solid orange line represents a smoothed spline with the confidence interval as surrounding blue shaded area.

Comparing the web servers in our tool collection grouped by publication date, we asked whether tools published in 2010 only, 2010 to 2011, and 2010 to 2012 have a different, i.e. higher citation count on average if they were reachable at least once in our tracking frame, as opposed to those being not reachable at all. The resulting *P*-values of 1.278×10^{-07}, 2.457×10^{-11} and 6.758×10^{-15}, and the about five times higher mean citation counts for the first group in each of the comparisons, strongly support these hypotheses. However, we reason that simply guaranteeing a long-term availability does not necessarily pay off with a high citation count, as 70 tools, which were still available and published in between 2010 and 2012, had less than 10 citations. To the contrary, only four tools that had been published in this period were not reachable at all in our time frame, even though they all received more than 100 citations. Whether the causal implication is that tools being well-cited early following publication are also subject to better long-term maintenance, or the other way around, tools that are well maintained tend to be cited more often on the long run, remains to be shown, e.g. through invited host surveys. Nevertheless, besides the quality of the work and the breadth of scope, we propose that many other factors influence the long-term impact of a web server. For example, we found that tools providing only an IP address or which were hosted in user-home directories were overall considerably less reachable (only 31 out of 68 tools (45.6%) reachable over IP and 20 out of 37 tools (54.1%) hosted in home directories were still available).

Based on our and previous findings, we collected a set of guidelines split into four categories to delineate good web server development practices targeted for beginners in the field, all of which are easy to implement and ultimately can prevent major sustainability issues (Table 1). In general, the guidelines are designed to support *reproducibility* of computational results, *security* by enforcing service integrity and privacy of user data, the *maintainability* through environment isolation and dependency minimization and *usability* via complementary ways of access, e.g. through an API, or strict documentation policies. We also ordered the specific recommendations in each category by decreasing priority to simplify selection of the most important 'DOs' and 'DO NOTs'. To enumerate on those, switching to production settings of all software components in-use, performing regular security updates, e.g. at least once every six months, and replacing standard admin access URLs and logins with hard to guess strings is essential for a reliable base level of security. To improve reproducibility, developers should encapsulate the software environment, e.g. through Docker, as much as possible, use proper code and data version control, e.g. using GIT, and publicly state any package dependencies and their version tested during development. For better maintainability, we recommend to minimize any effort that is needed to migrate the service, again by encapsulating the environment, properly fixing the software dependencies to prevent implicit updates when using package managers such as *conda* or *pip*, and document all required steps to reset the service, should it be necessary. Popular scripting languages like *Python* and *R* are especially vulnerable to implicit dependency updates as respective packages are updated at a high frequency. We also suggest developers to provide extensive sets of tutorials and example inputs or files to the user. Finally, hosting on official domain names instead of plain IP-addresses improves usability because names can be remembered and referred to significantly better than long numbers.

DISCUSSION

With increasing frequency and broader applications, the importance of bioinformatics web services and web servers is growing. This calls for an in-depth consideration on the availability and sustainability of respective services, since it might have severe consequences for research projects. In case a web-based program is used in other manuscripts to present analyses and the original tool is discontinued, later publications can be impacted by non-reproducible results. Aims of our study were to present a comprehensive analysis of the availability of web services, to get insights into the dynamics and to monitor the availability over a longer period of time, and to get an understanding whether more recent web services can be rescued by contacting the corresponding authors. There are more measures that could be

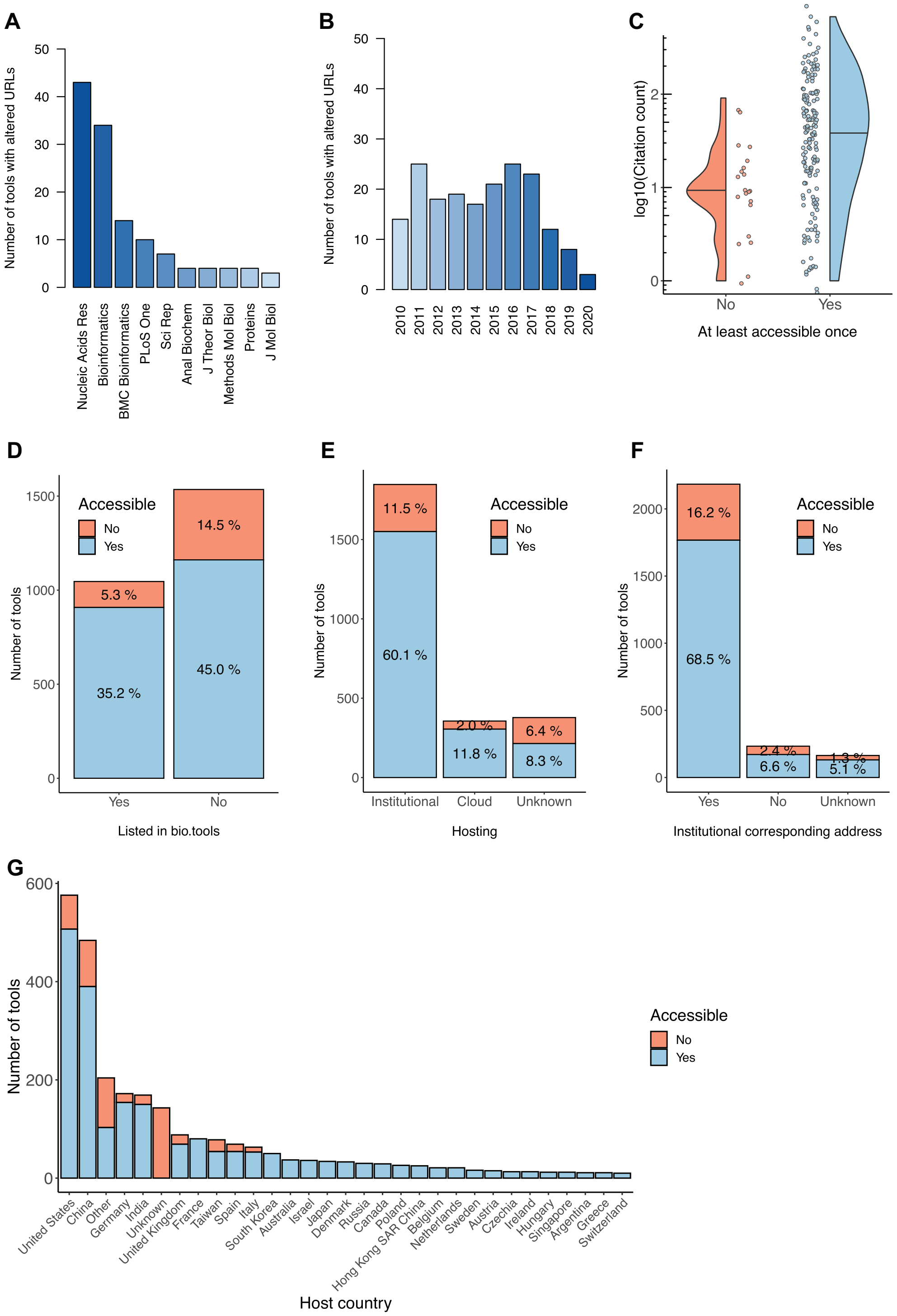


Figure 4. Distribution of publication metadata for altered (**A**–**C**) and all (**D**–**G**) tool URLs. (**A**) Bar chart for the number of tools with altered URL collected per journal. (**B**) Bar chart of the number of tools with altered URL by publication year. (**C**) Back-to-back violin-dot plot for the number of citations by accessibility status. (**D**) Stacked bar chart for the total number of tools and corresponding availability split by their presence in *bio.tools*. (**E**) Like in (D) but for determined host origins. (**F**) Like in (D) but for the type of corresponding e-mail address given in the associated publications. (**G**) Analogously to (D) but split by the web server host country. The special bars *Other* and *Unknown* summarize tools for countries with <10 web servers and indeterminable destination, respectively.

Table 1. Good practice guidelines for developing scientific web servers by category

Security	Reproducibility	Maintainability	Usability
Switch to production mode when deploying	Use virtual machines or container virtualization (e.g. Docker)	Minimize migration effort, encapsulate the software environment as much as possible	Create very detailed tutorials, one for each aspect of the web server
Perform regular security updates	Version control web server code and data	Keep all software dependencies fixed (e.g. YAML files)	Provide multiple example files
Do not use standard admin panel access domains and/or passwords	List software packages in use along with version numbers	Document internals as much as possible	Provide access over a domain name instead of an IP address
Escape user-input to prevent remote-code execution (e.g. SQL injection)	List main analysis parameters and provide timestamps in custom downloads (e.g. plots and tables)	Backup database onto external storage (e.g. user data)	Do not switch the top-level domain when publishing an update
Use DDoS protection service	Offer downloads for core data	Use popular frameworks, avoid implementing everything from scratch	Render progress bar and generate unique job ID for compute-intensive jobs
Use encryption (SSL/https)	Provide versioned subdomains and APIs	Avoid hosting in home directories, potentially depending on a user environment	Provide valid author contact details
Use valid SSL certificates to prevent malicious browser errors	Keep older versions running as archive	Include JavaScript libraries via CDN and keep a local copy as backup	Use a caching framework
Keep user submitted files private			Implement helper text messages
Set rate limits for public APIs			Use color-blind friendly palettes
Set file size limits for user uploads			Provide an (REST-ful) API in addition to standard interface
Set strict timeouts and use queue managers for compute intensive jobs			Announce maintenance slot to user before performing updates

Each column denotes a set of guidelines from the same category. Specific items in each category (column) are ordered by decreasing priority to simplify selection of the most important guidelines.

added to the analysis, e.g. usage rates, the number of update publications per tool, implementation technology and influence of international collaboration in the development of web servers. However, these aspects rather resemble a scientometric analysis (20), which does not belong to the core of our present study.

Among the most comprehensive articles on the availability of web based tools, Schultheiss *et al.* analyzed 927 web services published in the annual NAR Web Server Issues between 2003 and 2009 (15). Their test on the functionality on 77% of all tools showed that 13% were truly no longer working and for 45% of all services the functionality could be fully validated. A survey among 872 web server issue corresponding authors returned 274 replies, suggesting that the majority of tools are developed solely by students and researchers without a permanent position. Our analysis generalizes the results of the Schultheiss study. Around three times more tools were considered and also other journals than NAR were included. Additionally, we monitored the availability of web-based programs over a four-month period, which has not been performed in this manner before. Our results are nonetheless very well aligned with the observations by Schultheiss *et al.* described 10 years ago. We also provide a novel intervention experiment to demonstrate responsiveness and the estimated percentage of web servers that can be brought back to life by contacting corresponding authors.

It is important to elaborate on possible limitations of the present study. First, the literature search might already be biased since our search query requires the abstract to contain both, the keyword web service (or similar) and a web address and the strings ‘www’, ‘http’ or ‘https’. While this holds for many tools, obviously not all web servers are covered by a respective literature search leading to false negatives in our data set. A second limitation is the resolution of redirection triggers. While we followed html redirects in the download routine, other redirects were checked manually since they might also be triggered by client-side resolved JavaScript code. Whether and how redirects have been changed during the study runtime might also influence the results. A third limitation arises from the definition of availability. Many tools do not provide example files nor (RESTful-)APIs to test proper functionality in an automated fashion. In that, our analysis represents rather an upper boundary since a working main page of the web servers was already sufficient to count the tools as available. However, automatic testing the proper functionality for several thousand server instances without a common and standardized access interface is currently infeasible and requires extensive manual work. One strength of the study is at the same time a confounding factor: the rescue experiment potentially influenced the availability of tools. Likely, a substantial fraction of the 20 tools that were brought back to service by our e-mail initiative would have remained offline for a longer period of time without the intervention. Still, the 20 tools represent only a minor fraction of 0.8% of the 2581 tools included in the study. In the light of the on-going pandemic caused by SARS-CoV2, we did not detect a significant association between a reduced web server availability and the lockdown faced in most Asian and European countries between March and May of 2020. For three reasons this imaginable association is unlikely; First, our tracking frame reaches until the end of August, a time by which many universities returned to regular operations. Second, we compared availability for web servers hosted in Spain and Italy, the countries that were severely hit by the pandemic lockdown procedures and did not find an altered distribution of downtimes. Lastly, the reasons for server outage communicated by web server authors participating in the intervention experiment did not yield any COVID-19 related impact in all but one case. Similarly to the aforementioned analysis the common summer break, which is entirely contained in our tracking time-frame, did not have considerable influence on the availability rates, although it might be conceivable that the course of the summer break itself might have been altered by the SARS-CoV2 induced pandemic.

Our study raises questions about how to overcome the increasing trend of unavailable tools. First, cloud-based hosting and container-based applications such as Docker can simplify maintenance procedures and add to the reproducibility of research (21). In addition, open and comprehensive code-sharing is increasingly recognized and facilitated through major open-source platforms such as GitHub (https://github.com) and Docker hub (https://hub.docker.com). Further, recent community efforts such as *udocker* (22) promote usability of complex software tools by non-experts in multi-user environments, which closely matches most institutional compute server policies. Building upon these community efforts and our study results, we defined simple guidelines for developers that easily integrate into existing web server development workflows but are expected to substantially improve sustainability and long-term impact. Moreover, at best, one central repository would host a comprehensive list of web services. For this purpose several repositories and collections have already been established (e.g. https://www.biostars.org (23), https://bioinformaticssoftwareandtools.co.in/, or https://bio.tools/). Also, the EMBRACE Registry has been proposed as an active database for bioinformatics web services (24). Unfortunately, the web presence cannot be reached anymore (http://www.embraceregistry.net). Even though central and well-maintained databases are important, common standards and scientific guidelines become essential for large-scale data and code sharing practices. For example, ELIXIR (25) is one of the largest multi-national endeavors to integrate and coordinate computing facilities, web services, and databases across more than 220 research organizations. The FAIRsharing service (26) is a part of ELIXIR, providing community-based and reviewed standards/policies for sharing and maintaining databases. A comprehensive summary and detailed descriptions on the individual web service repositories can be found in (27).

Mechanisms for finding services automatically have already been discussed in 2008 (28) but still no perfect solution seems to exists and oftentimes manual curation is required. We suggest that a respective resource should contain at least the actual web link and a contact consisting of a full name and an e-mail address. Further, it would be desirable that web-based tools offer a well-defined API along with a standardized input file facilitating automated and daily remote tests. If testing fails, the respective contact can then be alerted automatically and mitigate the errors in due time. This could be a fair compromise to balance required efforts between the community, trying to keep the set of scientific web servers persistent, and the authors who need to provide suitable testing functionality on their services. It is conceivable for future artificial intelligence-based applications to further reduce manual intervention by automatically screening web sites to classify both availability and functionality. On the other hand, it is however also fair to mention that this task at present is implicitly performed on a large-scale by the entire research community.

As conclusion of our study we propose the timely development of a central web resource for monitoring the availability of web-based tools via automated API testing to generate on-going availability reports and statistics that serve both the web server developers and user community.

DATA AVAILABILITY

All data are freely available as supplement to the manuscript.

CODE AVAILABILITY

Computer code responsible to scan, download and aggregate web server statistics is available from https://github.com/CCB-SB/web-server-availability.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We appreciate the generous support of the corresponding authors of tools included in our rescue study.

FUNDING

Saarland University. Funding for open access charge: Internal funds.

Conflict of interest statement. None declared.

REFERENCES

1. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
2. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
3. Johnson,M., Zaretskaya,I., Raytselis,Y., Merezhuk,Y., McGinnis,S. and Madden,T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res.*, **36**, W5–W9.
4. Snel,B., Lehmann,G., Bork,P. and Huynen,M.A. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.*, **28**, 3442–3444.

5. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P. *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
6. Szklarczyk,D., Morris,J.H., Cook,H., Kuhn,M., Wyder,S., Simonovic,M., Santos,A., Doncheva,N.T., Roth,A., Bork,P. *et al.* (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.*, **45**, D362–D368.
7. McWilliam,H., Li,W., Uludag,M., Squizzato,S., Park,Y.M., Buso,N., Cowley,A.P. and Lopez,R. (2013) Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res.*, **41**, W597–W600.
8. Roberts,R.J. (2003) Editorial. *Nucleic Acids Res.*, **31**, 3289–3289.
9. Benson,G. (2017) Editorial: the 15th annual Nucleic Acids Research Web Server issue 2017. *Nucleic Acids Res.*, **45**, W1–W5.
10. Benson,G. (2018) Editorial: The 16th annual Nucleic Acids Research web server issue 2018. *Nucleic Acids Res.*, **46**, W1–W4.
11. Benson,G. (2019) Editorial: The 17th Annual Nucleic Acids Research Web Server Issue 2019. *Nucleic Acids Res.*, **47**, W1–W4.
12. Seelow,D. (2020) Editorial: the 18th annual Nucleic Acids Research web server issue 2020. *Nucleic Acids Res.*, **48**, W1–W4.
13. Veretnik,S., Fink,J.L. and Bourne,P.E. (2008) Computational biology resources lack persistence and usability. *PLoS Comput. Biol.*, **4**, e1000136.
14. Thireou,T., Spyrou,G. and Atlamazoglou,V. (2007) A survey of the availability of primary bioinformatics web resources. *Genomics Proteomics Bioinform.*, **5**, 70–76.
15. Schultheiss,S.J., Munch,M.C., Andreeva,G.D. and Ratsch,G. (2011) Persistence and availability of Web services in computational biology. *PLoS One*, **6**, e24914.
16. Wren,J.D., Georgescu,C., Giles,C.B. and Hennessey,J. (2017) Use it or lose it: citations predict the continued online availability of published bioinformatics resources. *Nucleic Acids Res.*, **45**, 3627–3633.
17. Al-Koofee,D.A.F., Ismael,J.M. and Mubarak,S.M.H. (2019) Retraction notice to 'Point mutation detection by economic HRM protocol primer design '[Biochem. Biophys. Rep. 18 (2019) 100628]. *Biochem. Biophys. Rep*, **20**, 100688.
18. Ison,J., Rapacki,K., Ménager,H., Kalaš,M., Rydza,E., Chmura,P., Anthon,C., Beard,N., Berka,K., Bolser,D. *et al.* (2016) Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Res.*, **44**, D38–D47.
19. Bornmann,L. and Mutz,R. (2015) Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inform. Sci. Technol.*, **66**, 2215–2222.
20. Scholz,S.S., Dillmann,M., Flohr,A., Backes,C., Fehlmann,T., Millenaar,D., Ukena,C., Bohm,M., Keller,A. and Mahfoud,F. (2020) Contemporary scientometric analyses using a novel web application: the science performance evaluation (SciPE) approach. *Clin. Res. Cardiol.*, **109**, 810–818.
21. Boettiger,C. (2015) An introduction to Docker for reproducible research. *SIGOPS Oper. Syst. Rev.*, **49**, 71–79.
22. Gomes,J., Bagnaschi,E., Campos,I., David,M., Alves,L., Martins,J., Pina,J., López-García,A. and Orviz,P. (2018) Enabling rootless Linux Containers in multi-user environments: the udocker tool. *Comput. Phys. Commun.*, **232**, 84–97.
23. Parnell,L.D., Lindenbaum,P., Shameer,K., Dall'Olio,G.M., Swan,D.C., Jensen,L.J., Cockell,S.J., Pedersen,B.S., Mangan,M.E., Miller,C.A. *et al.* (2011) BioStar: an online question & answer resource for the bioinformatics community. *PLoS Comput. Biol.*, **7**, e1002216.
24. Pettifer,S., Thorne,D., McDermott,P., Attwood,T., Baran,J., Bryne,J.C., Hupponen,T., Mowbray,D. and Vriend,G. (2009) An active registry for bioinformatics web services. *Bioinformatics*, **25**, 2090–2091.
25. Crosswell,L.C. and Thornton,J.M. (2012) ELIXIR: a distributed infrastructure for European biological data. *Trends Biotechnol.*, **30**, 241–242.
26. Sansone,S.-A., McQuilton,P., Rocca-Serra,P., Gonzalez-Beltran,A., Izzo,M., Lister,A.L., Thurston,M. and the, F.C. (2019) FAIRsharing as a community approach to standards, repositories and policies. *Nat. Biotechnol.*, **37**, 358–367.
27. Urdidiales-Nieto,D., Navas-Delgado,I. and Aldana-Montes,J.F. (2017) Biological web service repositories review. *Mol. Inf.*, **36**, 1600035.
28. Goble,C., Stevens,R., Hull,D., Wolstencroft,K. and Lopez,R. (2008) Data curation + process curation = data integration + science. *Brief. Bioinform.*, **9**, 506–517.

3.8 Deep sequencing of sncRNAs reveals hallmarks and regulatory modules of the transcriptome during Parkinson's disease progression

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

https://doi.org/10.1038/s43587-021-00042-6

4 Discussion

The new age of data has led to a massive increase in the number of data-driven studies ranging on an exponential scale, in contrast to traditional hypothesis-driven research. Experiencing also a massive diversification in the last ten years, bioinformatics continues to drive most modern molecular research in both academia and industry. Even though the biological phenomenon of aging has been described on the population and organism level for several decades already, much remains to be understood when it comes to the cellular pathways underlying healthy aging and age-related disease. Circulating proteins currently bear great value to be used as neurodegenerative disease biomarkers, however a general lack of advanced tools and representative human cohorts has frustrated the use of ncRNAs. In this thesis, novel methods and resources for miRNA-target and pathway analysis have been established to support these efforts, proving first success in large-scale screening applications.

Still, limitations and room for further improvements exist. First, more complex determinants for effective miRNA-target regulation than previously anticipated were identified, which is reflected by the observed varying and overall very limited performance of target prediction tools [6]. For instance, recent evidence suggests that target mRNAs may escape the miRNA induced decay through triggering a much faster miRNA turnover [414]. Moreover, lncRNAs and circular RNAs may function as so-called miRNA-sponges, effectively binding hundreds of free miRNA molecules for a timed and targeted release [415]. In that aspect, ncRNAs run a kind of competition against each other, further complicating our ability to predict protein levels based on mRNA abundance [416]. Furthermore, the mammalian miRNome is far from complete, requiring further discovery studies especially for non-model organisms [188]. Similarly, the number of validated miRNA targets is shallow in species other than human. This issue manifests in miRPathDB 2.0 where validation data was only sufficiently available for human and mouse in order to report enriched pathways. Further directed efforts are necessary to assess the degree of conservation at the pathway-level, another important criteria determining the success of biomarker discovery. In addition, a substantial imbalance towards cancer within the current body of published work on miRNAs confounds applications in other fields such as aging, justifying a more stringent analysis of functional associations derived with tools like miEAA [4].

Technology-wise, sequencing is the *de facto* standard technique for large-scale profiling of coding and ncRNAs. The applied research publications presented herein are based on miRNA-enriched bulk sequencing or microarray profiling of whole-blood samples. Small RNA sequencing on the Illumina platform is bias-afflicted by differential ligation and capture efficiency [417–419]. Further, whole-blood contains a mixture of erythroid cells, leukocytes and thrombocytes at varying proportions, leaving room for improvements to dissect miRNA expression at the individual cell type level. However, whether the success story of comprehensive RNA single-cell studies that redefined our understanding on human diseases could be repeated for sncRNAs remains to be demonstrated in the future [413; 420–423]. A recent study on total RNA profiling of single cells yielded a miRNA detection rate an order of magnitude lower than observed for standard bulk sequencing [424]. Until tremendous advancements in single-cell miRNA sequencing are accomplished, *in silico* unsupervised deconvolution of bulk data remains a technique of choice, which already revealed a cell type specific footprint of various RNAs in aging and PD [5; 8; 425]. Yet, these methods are naturally limited in their accuracy because no suitable ground-truth for miRNAs is currently available. Remarkable efforts dissecting single-cell transcriptomes of AD and PD patients, thereby comparing between the primarily affected brain regions and the peripheral system are just about to emerge [420; 426–429]. Similar advances for sncRNAs are desirable and may help to explain some of the cell type-specific mRNA perturbations observed in neurodegenerative diseases, but the throughput is currently limited by technology [430; 431]. Nonetheless, any single-cell method yields platform-specific biases, making it tricky to distinguish between true signals and technical or biological noise. While there is a strong motivation to develop equivalent protocols for miRNAs, it will certainly add another layer of complexity to the data analysis. A subsequent success of that endeavor will only be reached by developing new computational models that aid interpretation of high-dimensional data sets [62].

Gene and miRNA expression levels are modulated by a multitude of factors. It is therefore crucial to consider known confounding factors such as patient demographics, e.g. age and gender, lifestyle, or medication in analyzing larger cohorts. An evaluation of sncRNAs in the PPMI and NCER-PD cohorts yielded a significant bias of deregulated miRNAs along the lifespan [8]. However, appropriate modeling of treatment effects for miRNAs was found to be difficult due to multiple reasons. First, even though all patients were enrolled as *de novo*, i.e. being diagnosed within two years preceding enrollment, drug-naiveness was guaranteed only until the first follow-up and could then be initiated at any time. Second, treatments were apparently tuned for each patient in order to yield the best outcome, complicating a systematic comparison across the cohort. Third, only little is known whether and how miRNA expression responds to drugs, a process presumable depending on pharmacodynamic and pharmacokinetik aspects [432–434]. One tailored computational model using non-negative matrix factorization shed light onto drug-associated changes of oncomiRs following chemotherapy in breast cancer patients [435]. Together, differences in technology, sample origin and quality, cohort demographics and the heterogeneous treatments likely explain the so far rather low concordance of RNA biomarkers in neurodegenerative diseases.

4.0.1 Future directions

Future developments in basic and clinical miRNA research will certainly involve parallel breakthroughs in both experimental and computational platforms. A continuous intertwining of machine learning and artificial intelligence with traditional methods is likely. Most importantly, the ability to efficiently scale to huge amounts of data will be crucial for a long-term success of bioinformatics approaches in the field. As more and more aspects of ncRNA biogenesis and regulatory pathways are revealed, a major challenge for the community will be to unify the various models of different flavors into a logically well-defined theory. To this end, smart and integrative research based on multi-omics profiling approaches will be essential. Once a solid understanding of every single -omics field could be established, more realistic systems biology applications become feasible [407]. Moreover, better flexibility and interconnectivity between existing software implementations using open APIs would be desirable, simplifying broad access for researchers from across the life sciences. Also, open community standards for reporting reproducible results and automated pipelines will become more important than ever.

4.0.2 Conclusions

Taken together, the here presented tools and resources have created fundamental knowledge on the peripheral, cellular pathways controlled by miRNAs in aging and prevalent age-related diseases. While miEAA and miRPathDB are frequently used and cited by the community, our novel findings for progression markers in PD have triggered sustainable interest on fully characterizing the role of ncRNAs in this yet incurable disease of the elderly.

Bibliography

[1] F. Kern, C. Backes, P. Hirsch, T. Fehlmann, M. Hart, E. Meese, and A. Keller. What's the target: understanding two decades of in silico microrna-target prediction. *Brief Bioinform*, 21(6): 1999–2010, 2020. ISSN 1467-5463. doi: 10.1093/bib/bbz111. URL https://doi.org/10.1093/bib/bbz111.

[2] T. Kehl, F. Kern, C. Backes, T. Fehlmann, D. Stöckel, E. Meese, H. P. Lenhof, and A. Keller. mirpathdb 2.0: a novel release of the mirna pathway dictionary database. *Nucleic Acids Res*, 48(D1):D142–D147, 2020. ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gkz1022. URL https://doi.org/10.1093/nar/gkz1022.

[3] N. Ludwig, T. Fehlmann, F. Kern, M. Gogol, W. Maetzler, et al. Machine learning to detect alzheimer's disease from circulating non-coding rnas. *Genomics Proteomics Bioinformatics*, 17(4):430–440, 2019. ISSN 1672-0229 (Print) 1672-0229. doi: 10.1016/j.gpb.2019.09.004. URL https://doi.org/10.1016/j.gpb.2019.09.004.

[4] F. Kern, T. Fehlmann, J. Solomon, L. Schwed, N. Grammes, C. Backes, K. Van Keuren-Jensen, D. W. Craig, E. Meese, and A. Keller. mieaa 2.0: integrating multi-species microrna enrichment analysis and workflow management systems. *Nucleic Acids Res*, 48(W1):W521–W528, 2020. ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gkaa309. URL https://doi.org/10.1093/nar/gkaa309.

[5] T. Fehlmann, B. Lehallier, N. Schaum, O. Hahn, M. Kahraman, et al. Common diseases alter the physiological age-related blood microrna profile. *Nat Commun*, 11(1):5958, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19665-1. URL https://doi.org/10.1038/s41467-020-19665-1.

[6] F. Kern, L. Krammes, K. Danz, C. Diener, T. Kehl, et al. Validation of human microrna target pathways enables evaluation of target prediction tools. *Nucleic Acids Res*, 49(1):127–144, 2021. ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gkaa1161. URL https://doi.org/10.1093/nar/gkaa1161.

[7] F. Kern, T. Fehlmann, and A. Keller. On the lifetime of bioinformatics web services. *Nucleic Acids Res*, 48(22):12523–12533, 2020.

ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gkaa1125. URL https://doi.org/10.1093/nar/gkaa1125.

[8] Fabian Kern, Tobias Fehlmann, Ivo Violich, Eric Alsop, Elizabeth Hutchins, et al. Deep sequencing of sncrnas reveals hallmarks and regulatory modules of the transcriptome during parkinson's disease progression. *Nature Aging*, 1(3):309–322, 2021. ISSN 2662-8465. doi: 10.1038/s43587-021-00042-6. URL https://doi.org/10.1038/s43587-021-00042-6.

[9] Laura A. Hug, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, et al. A new view of the tree of life. *Nature Microbiology*, 1(5):16048, 2016. ISSN 2058-5276. doi: 10.1038/nmicrobiol.2016.48. URL https://doi.org/10.1038/nmicrobiol.2016.48.

[10] Tom A. Williams, Cymon J. Cox, Peter G. Foster, Gergely J. Szöllősi, and T. Martin Embley. Phylogenomics provides robust support for a two-domains tree of life. *Nature Ecology & Evolution*, 4(1):138–147, 2020. ISSN 2397-334X. doi: 10.1038/s41559-019-1040-x. URL https://doi.org/10.1038/s41559-019-1040-x.

[11] Harris A. Lewin, Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, et al. Earth biogenome project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17):4325–4333, 2018. doi: 10.1073/pnas.1720115115. URL https://www.pnas.org/content/pnas/115/17/4325.full.pdf.

[12] Anamaria Necsulea and Henrik Kaessmann. Evolutionary dynamics of coding and non-coding transcriptomes. *Nature Reviews Genetics*, 15(11):734–748, 2014. ISSN 1471-0064. doi: 10.1038/nrg3802. URL https://doi.org/10.1038/nrg3802.

[13] Sarah A. Signor and Sergey V. Nuzhdin. The evolution of gene expression in cis and trans. *Trends in Genetics*, 34(7):532–544, 2018. ISSN 0168-9525. doi: 10.1016/j.tig.2018.03.007. URL https://doi.org/10.1016/j.tig.2018.03.007.

[14] J. J. Emerson and Wen-Hsiung Li. The genetic basis of evolutionary change in gene expression levels. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1552):2581–2590, 2010. doi: 10.1098/rstb.2010.0005. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2010.0005.

[15] Peter W. Harrison, Alison E. Wright, and Judith E. Mank. The evolution of gene expression and the transcriptome–phenotype relationship. *Seminars in Cell & Developmental Biology*, 23(2):222–229, 2012. ISSN 1084-9521. doi: 10.1016/j.semcdb.2011.12.004. URL https://www.sciencedirect.com/science/article/pii/S1084952111002400.

[16] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, et al. Gencode reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1): D766–D773, 2018. ISSN 0305-1048. doi: 10.1093/nar/gky955. URL `https://doi.org/10.1093/nar/gky955`.

[17] Irene Gallego Romero, Ilya Ruvinsky, and Yoav Gilad. Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*, 13(7):505–516, 2012. ISSN 1471-0064. doi: 10.1038/nrg3229. URL `https://doi.org/10.1038/nrg3229`.

[18] Itay Tirosh, Sharon Reikhav, Avraham A. Levy, and Naama Barkai. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science*, 324 (5927):659–662, 2009. doi: 10.1126/science.1169766. URL `https://science.sciencemag.org/content/sci/324/5927/659.full.pdf`.

[19] Antoine Coulon, Carson C. Chow, Robert H. Singer, and Daniel R. Larson. Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nature Reviews Genetics*, 14 (8):572–584, 2013. ISSN 1471-0064. doi: 10.1038/nrg3484. URL `https://doi.org/10.1038/nrg3484`.

[20] Aaron D. Goldberg, C. David Allis, and Emily Bernstein. Epigenetics: A landscape takes shape. *Cell*, 128(4):635–638, 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.02.006. URL `https://doi.org/10.1016/j.cell.2007.02.006`.

[21] Giacomo Cavalli and Edith Heard. Advances in epigenetics link genetics to the environment and disease. *Nature*, 571(7766): 489–499, 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1411-0. URL `https://doi.org/10.1038/s41586-019-1411-0`.

[22] Christopher Buccitelli and Matthias Selbach. mrnas, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10):630–644, 2020. ISSN 1471-0064. doi: 10.1038/s41576-020-0258-4. URL `https://doi.org/10.1038/s41576-020-0258-4`.

[23] Vikki M. Weake and Jerry L. Workman. Inducible gene expression: diverse regulatory mechanisms. *Nature Reviews Genetics*, 11(6):426–437, 2010. ISSN 1471-0064. doi: 10.1038/nrg2781. URL `https://doi.org/10.1038/nrg2781`.

[24] Juan M. Vaquerizas, Sarah K. Kummerfeld, Sarah A. Teichmann, and Nicholas M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263, 2009. ISSN 1471-0064. doi: 10.1038/nrg2538. URL `https://doi.org/10.1038/nrg2538`.

[25] François Spitz and Eileen E. M. Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, 2012. ISSN 1471-0064. doi: 10.1038/nrg3207. URL https://doi.org/10.1038/nrg3207.

[26] Chin-Tong Ong and Victor G. Corces. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics*, 12(4):283–293, 2011. ISSN 1471-0064. doi: 10.1038/nrg2957. URL https://doi.org/10.1038/nrg2957.

[27] Dafne Campigli Di Giammartino, Andreas Kloetgen, Alexander Polyzos, Yiyuan Liu, Daleum Kim, et al. Klf4 is involved in the organization and regulation of pluripotency-associated three-dimensional enhancer networks. *Nature Cell Biology*, 21(10):1179–1190, 2019. ISSN 1476-4679. doi: 10.1038/s41556-019-0390-6. URL https://doi.org/10.1038/s41556-019-0390-6.

[28] Alex H. M. Ng, Parastoo Khoshakhlagh, Jesus Eduardo Rojo Arias, Giovanni Pasquini, Kai Wang, et al. A comprehensive library of human transcription factors for cell fate engineering. *Nature Biotechnology*, 39(4):510–519, 2021. ISSN 1546-1696. doi: 10.1038/s41587-020-0742-6. URL https://doi.org/10.1038/s41587-020-0742-6.

[29] Dafni Anastasiadi, Anna Esteve-Codina, and Francesc Piferrer. Consistent inverse correlation between dna methylation of the first intron and gene expression across tissues and species. *Epigenetics & Chromatin*, 11(1):37, 2018. ISSN 1756-8935. doi: 10.1186/s13072-018-0205-1. URL https://doi.org/10.1186/s13072-018-0205-1.

[30] Rosa Karlić, Ho-Ryun Chung, Julia Lasserre, Kristian Vlahoviček, and Martin Vingron. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, 107(7):2926–2931, 2010. doi: 10.1073/pnas.0909344107. URL https://www.pnas.org/content/pnas/107/7/2926.full.pdf.

[31] F. Schmidt, F. Kern, and M. H. Schulz. Integrative prediction of gene expression with chromatin accessibility and conformation data. *Epigenetics Chromatin*, 13(1):4, 2020. ISSN 1756-8935. doi: 10.1186/s13072-020-0327-0. URL https://doi.org/10.1186/s13072-020-0327-0.

[32] Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456 (7221):470–476, 2008. ISSN 1476-4687. doi: 10.1038/nature07509. URL https://doi.org/10.1038/nature07509.

[33] Timothy W. Nilsen and Brenton R. Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280): 457–463, 2010. ISSN 1476-4687. doi: 10.1038/nature08909. URL https://doi.org/10.1038/nature08909.

[34] Sung Ho Boo and Yoon Ki Kim. The emerging role of rna modifications in the regulation of mrna stability. *Experimental & Molecular Medicine*, 52(3):400–408, 2020. ISSN 2092-6413. doi: 10.1038/s12276-020-0407-z. URL https://doi.org/10.1038/s12276-020-0407-z.

[35] Nicky Jonkhout, Julia Tran, Martin A. Smith, Nicole Schonrock, John S. Mattick, and Eva Maria Novoa. The rna modification landscape in human disease. *RNA*, 23(12):1754–1769, 2017. doi: 10.1261/rna.063503.117. URL http://rnajournal.cshlp.org/content/23/12/1754.abstract.

[36] Sylvain Delaunay and Michaela Frye. Rna modifications regulating cell fate in cancer. *Nature Cell Biology*, 21(5):552–559, 2019. ISSN 1476-4679. doi: 10.1038/s41556-019-0319-0. URL https://doi.org/10.1038/s41556-019-0319-0.

[37] Young-Guen Kwon, Soo Young Lee, Yongwon Choi, Paul Greengard, and Angus C. Nairn. Cell cycle-dependent phosphorylation of mammalian protein phosphatase 1 by cdc2 kinase. *Proceedings of the National Academy of Sciences*, 94(6): 2168–2173, 1997. doi: 10.1073/pnas.94.6.2168. URL https://www.pnas.org/content/pnas/94/6/2168.full.pdf.

[38] Christian Pohl and Ivan Dikic. Cellular quality control by the ubiquitin-proteasome system and autophagy. *Science*, 366(6467):818–822, 2019. doi: 10.1126/science.aax3769. URL https://science.sciencemag.org/content/sci/366/6467/818.full.pdf.

[39] Sheelagh Frame, Philip Cohen, and Ricardo M. Biondi. A common phosphate binding site explains the unique substrate specificity of gsk3 and its inactivation by phosphorylation. *Molecular Cell*, 7(6):1321–1327, 2001. ISSN 1097-2765. doi: 10.1016/S1097-2765(01)00253-2. URL https://doi.org/10.1016/S1097-2765(01)00253-2.

[40] Sha Yu and V. Narry Kim. A tale of non-canonical tails: gene regulation by post-transcriptional rna tailing. *Nature Reviews Molecular Cell Biology*, 21(9):542–556, 2020. ISSN 1471-0080. doi: 10.1038/s41580-020-0246-8. URL https://doi.org/10.1038/s41580-020-0246-8.

[41] Luca Grumolato, Guizhong Liu, Phyllus Mong, Raksha Mudbhary, Romi Biswas, Randy Arroyave, Sapna Vijayakumar, Aris N. Economides, and Stuart A. Aaronson. Canonical

and noncanonical wnts use a common mechanism to activate completely unrelated coreceptors. *Genes & Development*, 24(22):2517–2530, 2010. doi: 10.1101/gad.1957710. URL http://genesdev.cshlp.org/content/24/22/2517.abstract.

[42] Angela Arensdorf, Danilo Diedrichs, and Thomas Rutkowski. Regulation of the transcriptome by er stress: non-canonical mechanisms and physiological consequences. *Frontiers in Genetics*, 4(256), 2013. ISSN 1664-8021. doi: 10.3389/fgene.2013.00256. URL https://www.frontiersin.org/article/10.3389/fgene.2013.00256.

[43] Yufeng Yao, Qiulun Lu, Zhenkun Hu, Yubin Yu, Qiuyun Chen, and Qing K. Wang. A non-canonical pathway regulates er stress signaling and blocks er stress-induced apoptosis and heart failure. *Nature Communications*, 8(1):133, 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-00171-w. URL https://doi.org/10.1038/s41467-017-00171-w.

[44] Peiwei Huangyang and M. Celeste Simon. Hidden features: exploring the non-canonical functions of metabolic enzymes. *Disease Models & Mechanisms*, 11(8), 2018. ISSN 1754-8403. doi: 10.1242/dmm.033365. URL https://doi.org/10.1242/dmm.033365.

[45] Sherif Rashad, Teiji Tominaga, and Kuniyasu Niizuma. The cell and stress-specific canonical and non-canonical trna cleavage. *bioRxiv*, page 2020.02.04.934695, 2020. doi: 10.1101/2020.02.04.934695. URL https://www.biorxiv.org/content/biorxiv/early/2020/07/29/2020.02.04.934695.full.pdf.

[46] Nils Eling, Michael D. Morgan, and John C. Marioni. Challenges in measuring and understanding biological noise. *Nature Reviews Genetics*, 20(9):536–548, 2019. ISSN 1471-0064. doi: 10.1038/s41576-019-0130-6. URL https://doi.org/10.1038/s41576-019-0130-6.

[47] Florian Schmidt and Marcel H Schulz. On the problem of confounders in modeling gene expression. *Bioinformatics*, 35(4):711–719, 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty674. URL https://doi.org/10.1093/bioinformatics/bty674.

[48] John Toubia, Vanessa M. Conn, and Simon J. Conn. Don't go in circles: confounding factors in gene expression profiling. *The EMBO Journal*, 37(11):e97945, 2018. ISSN 0261-4189. doi: 10.15252/embj.201797945. URL https://doi.org/10.15252/embj.201797945.

[49] Jennifer Listgarten, Carl Kadie, Eric E. Schadt, and David Heckerman. Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences*, 107(38):16465–16470, 2010. doi: 10.1073/

pnas.1002425107. URL https://www.pnas.org/content/pnas/107/38/16465.full.pdf.

[50] Princy Parsana, Claire Ruberman, Andrew E. Jaffe, Michael C. Schatz, Alexis Battle, and Jeffrey T. Leek. Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biology*, 20(1):94, 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1700-9. URL https://doi.org/10.1186/s13059-019-1700-9.

[51] Li Tong, Po-Yen Wu, John H. Phan, Hamid R. Hassazadeh, Wendell D. Jones, et al. Impact of rna-seq data analysis algorithms on gene expression estimation and downstream prediction. *Scientific Reports*, 10(1):17925, 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-74567-y. URL https://doi.org/10.1038/s41598-020-74567-y.

[52] Sahin Naqvi, Alexander K. Godfrey, Jennifer F. Hughes, Mary L. Goodheart, Richard N. Mitchell, and David C. Page. Conservation, acquisition, and functional impact of sex-biased gene expression in mammals. *Science*, 365 (6450):eaaw7317, 2019. doi: 10.1126/science.aaw7317. URL https://science.sciencemag.org/content/sci/365/6450/eaaw7317.full.pdf.

[53] Yixing Han, Shouguo Gao, Kathrin Muegge, Wei Zhang, and Bing Zhou. Advanced applications of rna sequencing and challenges. *Bioinformatics and Biology Insights*, 9s1: BBI.S28991, 2015. doi: 10.4137/bbi.S28991. URL https://journals.sagepub.com/doi/abs/10.4137/BBI.S28991.

[54] Julia Alles, Tobias Fehlmann, Ulrike Fischer, Christina Backes, Valentina Galata, et al. An estimate of the total number of true human mirnas. *Nucleic acids research*, 2019. doi: 10.1093/nar/gkz097. URL https://doi.org/10.1093/nar/gkz097.

[55] Yaron E. Antebi, Nagarajan Nandagopal, and Michael B. Elowitz. An operational view of intercellular signaling pathways. *Current Opinion in Systems Biology*, 1:16–24, 2017. ISSN 2452-3100. doi: 10.1016/j.coisb.2016.12.003. URL https://www.sciencedirect.com/science/article/pii/S2452310016300233.

[56] Daphne Tsoucas, Rui Dong, Haide Chen, Qian Zhu, Guoji Guo, and Guo-Cheng Yuan. Accurate estimation of cell-type composition from gene expression data. *Nature Communications*, 10(1):2975, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-10802-z. URL https://doi.org/10.1038/s41467-019-10802-z.

[57] Sarah Kim-Hellmuth, François Aguet, Meritxell Oliva, Manuel Muñoz-Aguirre, Silva Kasela, et al. Cell type–specific

genetic regulation of gene expression across human tissues. *Science*, 369(6509):eaaz8528, 2020. doi: 10.1126/science.aaz8528. URL https://science.sciencemag.org/content/sci/369/6509/eaaz8528.full.pdf.

[58] Dylan Kotliar, Adrian Veres, M. Aurel Nagy, Shervin Tabrizi, Eran Hodis, Douglas A. Melton, and Pardis C. Sabeti. Identifying gene expression programs of cell-type identity and cellular activity with single-cell rna-seq. *eLife*, 8:e43803, 2019. ISSN 2050-084X. doi: 10.7554/eLife.43803. URL https://doi.org/10.7554/eLife.43803.

[59] Abhijeet Rajendra Sonawane, John Platig, Maud Fagny, Cho-Yi Chen, Joseph Nathaniel Paulson, Camila Miranda Lopes-Ramos, Dawn Lisa DeMeo, John Quackenbush, Kimberly Glass, and Marieke Lydia Kuijjer. Understanding tissue-specific gene regulation. *Cell Reports*, 21(4):1077–1088, 2017. ISSN 2211-1247. doi: 10.1016/j.celrep.2017.10.001. URL https://www.sciencedirect.com/science/article/pii/S2211124717314183.

[60] François Aguet, Andrew A. Brown, Stephane E. Castel, Joe R. Davis, Yuan He, et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017. ISSN 1476-4687. doi: 10.1038/nature24277. URL https://doi.org/10.1038/nature24277.

[61] Erick Armingol, Adam Officer, Olivier Harismendy, and Nathan E. Lewis. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, 22(2):71–88, 2021. ISSN 1471-0064. doi: 10.1038/s41576-020-00292-x. URL https://doi.org/10.1038/s41576-020-00292-x.

[62] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J. McCarthy, Stephanie C. Hicks, et al. Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):31, 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-1926-6. URL https://doi.org/10.1186/s13059-020-1926-6.

[63] Ning Zhou, Xiao-Bing Zhang, Cheng Chen, Xin-Yu Chen, Bo Kang, Jian-Qin He, Guo-Zhong Gong, Ying-Jie Wang, and Yan-Wen Zhou. Cellular context- and protein level-dependent interaction of pluripotency factor oct4a with multiple octamer motifs of the same target gene. *Life Sciences*, 248:117461, 2020. ISSN 0024-3205. doi: 10.1016/j.lfs.2020.117461. URL https://www.sciencedirect.com/science/article/pii/S0024320520302095.

[64] Aviv Regev, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, et al. The human cell atlas. *eLife*, 6:e27041, 2017. ISSN 2050-084X. doi: 10.7554/eLife.27041. URL https://doi.org/10.7554/eLife.27041.

[65] Ron Milo, Paul Jorgensen, Uri Moran, Griffin Weber, and Michael Springer. Bionumbers–the database of key numbers in molecular and cell biology. *Nucleic acids research*, 38(Database issue):D750–D753, 2010. ISSN 1362-4962 0305-1048. doi: 10.1093/nar/gkp889. URL `https://pubmed.ncbi.nlm.nih.gov/19854939https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2808940/`.

[66] Dongxue Wang, Basak Eraslan, Thomas Wieland, Björn Hallström, Thomas Hopf, et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Molecular Systems Biology*, 15(2):e8503, 2019. ISSN 1744-4292. doi: 10.15252/msb.20188503. URL `https://www.embopress.org/doi/abs/10.15252/msb.20188503`.

[67] Mar Gonzàlez-Porta, Adam Frankish, Johan Rung, Jennifer Harrow, and Alvis Brazma. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology*, 14(7):R70, 2013. ISSN 1474-760X. doi: 10.1186/gb-2013-14-7-r70. URL `https://doi.org/10.1186/gb-2013-14-7-r70`.

[68] Shin Lin, Yiing Lin, Joseph R. Nery, Mark A. Urich, Alessandra Breschi, et al. Comparison of the transcriptional landscapes between human and mouse tissues. *Proceedings of the National Academy of Sciences*, 111(48):17224–17229, 2014. doi: 10.1073/pnas.1413624111. URL `https://www.pnas.org/content/pnas/111/48/17224.full.pdf`.

[69] Kai Wang, Hong Li, Yue Yuan, Alton Etheridge, Yong Zhou, David Huang, Paul Wilmes, and David Galas. The complex exogenous rna spectra in human plasma: An interface with human gut biota? *PLOS ONE*, 7(12):e51009, 2012. doi: 10.1371/journal.pone.0051009. URL `https://doi.org/10.1371/journal.pone.0051009`.

[70] Anubrata Ghosal. Secreted bacterial rna: an unexplored avenue. *FEMS Microbiology Letters*, 365(7), 2018. ISSN 0378-1097. doi: 10.1093/femsle/fny036. URL `https://doi.org/10.1093/femsle/fny036`.

[71] Hongwei Liang, Lei Huang, Jingjing Cao, Ke Zen, Xi Chen, and Chen-Yu Zhang. Regulation of mammalian gene expression by exogenous micrornas. *WIREs RNA*, 3(5):733–742, 2012. ISSN 1757-7004. doi: 10.1002/wrna.1127. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/wrna.1127`.

[72] Ryan M. Allen, Shilin Zhao, Marisol A. Ramirez Solano, Wanying Zhu, Danielle L. Michell, et al. Bioinformatic analysis of endogenous and exogenous small rnas on lipoproteins. *Journal of Extracellular Vesicles*, 7(1):1506198, 2018. ISSN null. doi: 10.1080/20013078.2018.1506198. URL `https://doi.org/10.1080/20013078.2018.1506198`.

[73] Kemal Avican, Jehad Aldahdooh, Matteo Togninalli, A. K. M. Firoj Mahmud, Jing Tang, Karsten M. Borgwardt, Mikael Rhen, and Maria Fällman. Rna atlas of human bacterial pathogens uncovers stress dynamics linked to infection. *Nature Communications*, 12(1):3282, 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23588-w. URL `https://doi.org/10.1038/s41467-021-23588-w`.

[74] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001. ISSN 1476-4687. doi: 10.1038/35057062. URL `https://doi.org/10.1038/35057062`.

[75] J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001. doi: 10.1126/science.1058040. URL `https://science.sciencemag.org/content/sci/291/5507/1304.full.pdf`.

[76] Jill Cheng, Philipp Kapranov, Jorg Drenkow, Sujit Dike, Shane Brubaker, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, 308 (5725):1149–1154, 2005. doi: 10.1126/science.1108625. URL `https://science.sciencemag.org/content/sci/308/5725/1149.full.pdf`.

[77] Alexander Hüttenhofer, Peter Schattner, and Norbert Polacek. Non-coding rnas: hope or hype? *Trends in Genetics*, 21(5):289–297, 2005. ISSN 0168-9525. doi: 10.1016/j.tig.2005.03.007. URL `https://doi.org/10.1016/j.tig.2005.03.007`.

[78] Alexander F. Palazzo and Eliza S. Lee. Non-coding rna: what is functional and what is junk? *Frontiers in Genetics*, 6(2), 2015. ISSN 1664-8021. doi: 10.3389/fgene.2015.00002. URL `https://www.frontiersin.org/article/10.3389/fgene.2015.00002`.

[79] Kevin V. Morris and John S. Mattick. The rise of regulatory rna. *Nature Reviews Genetics*, 15(6):423–437, 2014. ISSN 1471-0064. doi: 10.1038/nrg3722. URL `https://doi.org/10.1038/nrg3722`.

[80] Ian Dunham, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012. ISSN 1476-4687. doi: 10.1038/nature11247. URL `https://doi.org/10.1038/nature11247`.

[81] Allison Piovesan, Francesca Antonaros, Lorenza Vitale, Pierluigi Strippoli, Maria Chiara Pelleri, and Maria Caracausi. Human protein-coding genes and gene feature statistics in 2019. *BMC Research Notes*, 12(1):315, 2019. ISSN 1756-0500. doi:

10.1186/s13104-019-4343-8. URL https://doi.org/10.1186/s13104-019-4343-8.

[82] Neville E. Sanjana, Jason Wright, Kaijie Zheng, Ophir Shalem, Pierre Fontanillas, Julia Joung, Christine Cheng, Aviv Regev, and Feng Zhang. High-resolution interrogation of functional elements in the noncoding genome. *Science*, 353(6307):1545–1549, 2016. doi: 10.1126/science.aaf7613. URL https://science.sciencemag.org/content/sci/353/6307/1545.full.pdf.

[83] Mihaela Pertea, Alaina Shumate, Geo Pertea, Ales Varabyou, Florian P. Breitwieser, Yu-Chi Chang, Anil K. Madugundu, Akhilesh Pandey, and Steven L. Salzberg. Chess: a new human gene catalog curated from thousands of large-scale rna sequencing experiments reveals extensive transcriptional noise. *Genome Biology*, 19(1):208, 2018. ISSN 1474-760X. doi: 10.1186/s13059-018-1590-2. URL https://doi.org/10.1186/s13059-018-1590-2.

[84] Jingshan Huang, Karen Eilbeck, Barry Smith, Judith A. Blake, Dejing Dou, et al. The non-coding rna ontology (ncro): a comprehensive resource for the unification of non-coding rna biology. *Journal of Biomedical Semantics*, 7(1):24, 2016. ISSN 2041-1480. doi: 10.1186/s13326-016-0066-0. URL https://doi.org/10.1186/s13326-016-0066-0.

[85] Jingshan Huang, Karen Eilbeck, Barry Smith, Judith A. Blake, Dejing Dou, et al. The development of non-coding rna ontology. *International journal of data mining and bioinformatics*, 15(3):214–232, 2016. ISSN 1748-5673 1748-5681. doi: 10.1504/IJDMB.2016.077072. URL https://pubmed.ncbi.nlm.nih.gov/27990175https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5156483/.

[86] Thomas Derrien, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, et al. The gencode v7 catalog of human long noncoding rnas: Analysis of their gene structure, evolution, and expression. *Genome Research*, 22(9):1775–1789, 2012. doi: 10.1101/gr.132159.111. URL http://genome.cshlp.org/content/22/9/1775.abstract.

[87] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, et al. Gencode: The reference human genome annotation for the encode project. *Genome Research*, 22(9):1760–1774, 2012. doi: 10.1101/gr.135350.111. URL http://genome.cshlp.org/content/22/9/1760.abstract.

[88] Barbara Uszczynska-Ratajczak, Julien Lagarde, Adam Frankish, Roderic Guigó, and Rory Johnson. Towards a complete map of the human long non-coding rna transcriptome. *Nature Reviews Genetics*, 19(9):535–548, 2018. ISSN 1471-0064. doi:

10.1038/s41576-018-0017-y. URL https://doi.org/10.1038/s41576-018-0017-y.

[89] Luisa Statello, Chun-Jie Guo, Ling-Ling Chen, and Maite Huarte. Gene regulation by long non-coding rnas and its biological functions. *Nature Reviews Molecular Cell Biology*, 22(2): 96–118, 2021. ISSN 1471-0080. doi: 10.1038/s41580-020-00315-9. URL https://doi.org/10.1038/s41580-020-00315-9.

[90] Jeffrey J. Quinn and Howard Y. Chang. Unique features of long non-coding rna biogenesis and function. *Nature Reviews Genetics*, 17(1):47–62, 2016. ISSN 1471-0064. doi: 10.1038/nrg.2015.10. URL https://doi.org/10.1038/nrg.2015.10.

[91] T. Fehlmann, C. Backes, J. Alles, U. Fischer, M. Hart, et al. A high-resolution map of the human small non-coding transcriptome. *Bioinformatics*, 34(10):1621–1628, 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx814. URL https://doi.org/10.1093/bioinformatics/btx814.

[92] Gunter Meister. Argonaute proteins: functional insights and emerging roles. *Nature Reviews Genetics*, 14(7):447–459, 2013. ISSN 1471-0064. doi: 10.1038/nrg3462. URL https://doi.org/10.1038/nrg3462.

[93] Manel Esteller. Non-coding rnas in human disease. *Nature Reviews Genetics*, 12(12):861–874, 2011. ISSN 1471-0064. doi: 10.1038/nrg3074. URL https://doi.org/10.1038/nrg3074.

[94] R. C. Lee, R. L. Feinbaum, and V. Ambros. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell*, 75(5):843–54, 1993. ISSN 0092-8674 (Print) 0092-8674. doi: 10.1016/0092-8674(93)90529-y. URL https://doi.org/10.1016/0092-8674(93)90529-y.

[95] Amy E. Pasquinelli, Brenda J. Reinhart, Frank Slack, Mark Q. Martindale, Mitzi I. Kuroda, et al. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory rna. *Nature*, 408(6808):86–89, 2000. ISSN 1476-4687. doi: 10.1038/35040556. URL https://doi.org/10.1038/35040556.

[96] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun. The 21-nucleotide let-7 rna regulates developmental timing in caenorhabditis elegans. *Nature*, 403(6772):901–6, 2000. ISSN 0028-0836 (Print) 0028-0836. doi: 10.1038/35002607. URL https://doi.org/10.1038/35002607.

[97] Rosalind C. Lee and Victor Ambros. An extensive class of small rnas in caenorhabditis elegans. *Science*, 294 (5543):862–864, 2001. doi: 10.1126/science.1065329. URL https://science.sciencemag.org/content/sci/294/5543/862.full.pdf.

[98] J. Brennecke, D. R. Hipfner, A. Stark, R. B. Russell, and S. M. Cohen. bantam encodes a developmentally regulated microrna that controls cell proliferation and regulates the proapoptotic gene hid in drosophila. *Cell*, 113(1):25–36, 2003. ISSN 0092-8674 (Print) 0092-8674. doi: 10.1016/s0092-8674(03)00231-9. URL https://doi.org/10.1016/s0092-8674(03)00231-9.

[99] Alexander Stark, Michael F. Lin, Pouya Kheradpour, Jakob S. Pedersen, Leopold Parts, et al. Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature*, 450(7167):219–232, 2007. ISSN 1476-4687. doi: 10.1038/nature06340. URL https://doi.org/10.1038/nature06340.

[100] VICTOR AMBROS, BONNIE BARTEL, DAVID P. BARTEL, CHRISTOPHER B. BURGE, JAMES C. CARRINGTON, et al. A uniform system for microrna annotation. *RNA*, 9(3): 277–279, 2003. doi: 10.1261/rna.2183803. URL http://rnajournal.cshlp.org/content/9/3/277.abstract.

[101] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. mirbase: from microrna sequences to function. *Nucleic Acids Research*, 47(D1):D155–D162, 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1141. URL https://doi.org/10.1093/nar/gky1141.

[102] B. Fromm, A. Keller, X. Yang, M. R. Friedlander, K. J. Peterson, and S. Griffiths-Jones. Quo vadis micrornas? *Trends Genet*, 36(7):461–463, 2020. ISSN 0168-9525 (Print) 0168-9525. doi: 10.1016/j.tig.2020.03.007. URL https://doi.org/10.1016/j.tig.2020.03.007.

[103] Nicole Ludwig, Meike Becker, Timo Schumann, Timo Speer, Tobias Fehlmann, Andreas Keller, and Eckart Meese. Bias in recent mirbase annotations potentially associated with rna quality issues. *Scientific Reports*, 7(1):5162, 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-05070-0. URL https://doi.org/10.1038/s41598-017-05070-0.

[104] C. Ender, A. Krek, M. R. Friedländer, M. Beitzinger, L. Weinmann, W. Chen, S. Pfeffer, N. Rajewsky, and G. Meister. A human snorna with microrna-like functions. *Mol Cell*, 32(4): 519–28, 2008. ISSN 1097-2765. doi: 10.1016/j.molcel.2008.10.017. URL https://doi.org/10.1016/j.molcel.2008.10.017.

[105] T. Fehlmann, T. Laufer, C. Backes, M. Kahramann, J. Alles, et al. Large-scale validation of mirnas by disease association, evolutionary conservation and pathway activity. *RNA Biol*, 16(1):93–103, 2019. ISSN 1547-6286 (Print) 1547-6286. doi: 10.1080/15476286.2018.1559689. URL https://doi.org/10.1080/15476286.2018.1559689.

[106] Ahmet M. Denli, Bastiaan B. J. Tops, Ronald H. A. Plasterk, René F. Ketting, and Gregory J. Hannon. Processing of primary micrornas by the microprocessor complex. *Nature*, 432(7014):

231–235, 2004. ISSN 1476-4687. doi: 10.1038/nature03049. URL https://doi.org/10.1038/nature03049.

[107] Yangming Wang, Rostislav Medvid, Collin Melton, Rudolf Jaenisch, and Robert Blelloch. Dgcr8 is essential for microrna biogenesis and silencing of embryonic stem cell self-renewal. *Nature Genetics*, 39(3):380–385, 2007. ISSN 1546-1718. doi: 10.1038/ng1969. URL https://doi.org/10.1038/ng1969.

[108] David P. Bartel. Metazoan micrornas. *Cell*, 173(1):20–51, 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.03.006. URL https://doi.org/10.1016/j.cell.2018.03.006.

[109] Erno Wienholds, Marco J. Koudijs, Freek J. M. van Eeden, Edwin Cuppen, and Ronald H. A. Plasterk. The microrna-producing enzyme dicer1 is essential for zebrafish development. *Nature Genetics*, 35(3):217–218, 2003. ISSN 1546-1718. doi: 10.1038/ng1251. URL https://doi.org/10.1038/ng1251.

[110] G. Meister, M. Landthaler, A. Patkaniowska, Y. Dorsett, G. Teng, and T. Tuschl. Human argonaute2 mediates rna cleavage targeted by mirnas and sirnas. *Mol Cell*, 15(2):185–97, 2004. ISSN 1097-2765 (Print) 1097-2765. doi: 10.1016/j.molcel.2004.07.007. URL https://doi.org/10.1016/j.molcel.2004.07.007.

[111] F. Kern, J. Amand, I. Senatorov, A. Isakova, C. Backes, E. Meese, A. Keller, and T. Fehlmann. mirswitch: detecting microrna arm shift and switch events. *Nucleic Acids Res*, 48(W1):W268–w274, 2020. ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gkaa323. URL https://doi.org/10.1093/nar/gkaa323.

[112] Ligang Wu, Jihua Fan, and Joel G. Belasco. Micrornas direct rapid deadenylation of mrna. *Proceedings of the National Academy of Sciences of the United States of America*, 103(11):4034–4039, 2006. doi: 10.1073/pnas.0510928103. URL https://www.pnas.org/content/pnas/103/11/4034.full.pdf.

[113] J. Robin Lytle, Therese A. Yario, and Joan A. Steitz. Target mrnas are repressed as efficiently by microrna-binding sites in the 5′ utr as in the 3′ utr. *Proceedings of the National Academy of Sciences*, 104(23):9667–9672, 2007. doi: 10.1073/pnas.0703820104. URL https://www.pnas.org/content/pnas/104/23/9667.full.pdf.

[114] Stefanie Jonas and Elisa Izaurralde. Towards a molecular understanding of microrna-mediated gene silencing. *Nature Reviews Genetics*, 16(7):421–433, 2015. ISSN 1471-0064. doi: 10.1038/nrg3965. URL https://doi.org/10.1038/nrg3965.

[115] Shobha Vasudevan, Yingchun Tong, and Joan A. Steitz. Switching from repression to activation: Micrornas can up-regulate translation. *Science*, 318(5858):1931–1934, 2007. doi: 10.1126/science.1149460. URL https://science.sciencemag.org/content/sci/318/5858/1931.full.pdf.

[116] Julius Brennecke, Alexander Stark, Robert B. Russell, and Stephen M. Cohen. Principles of microrna–target recognition. *PLOS Biology*, 3(3):e85, 2005. doi: 10.1371/journal.pbio.0030085. URL https://doi.org/10.1371/journal.pbio.0030085.

[117] Nicole T. Schirle, Jessica Sheu-Gruttadauria, and Ian J. MacRae. Structural basis for microrna targeting. *Science*, 346 (6209):608–613, 2014. doi: 10.1126/science.1258040. URL https://science.sciencemag.org/content/sci/346/6209/608.full.pdf.

[118] M. Hart, F. Kern, C. Backes, S. Rheinheimer, T. Fehlmann, A. Keller, and E. Meese. The deterministic role of 5-mers in microrna-gene targeting. *RNA Biol*, 15(6):819–825, 2018. ISSN 1547-6286 (Print) 1547-6286. doi: 10.1080/15476286.2018.1462652. URL https://doi.org/10.1080/15476286.2018.1462652.

[119] David M. Garcia, Daehyun Baek, Chanseok Shin, George W. Bell, Andrew Grimson, and David P. Bartel. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other micrornas. *Nature Structural & Molecular Biology*, 18(10):1139–1146, 2011. ISSN 1545-9985. doi: 10.1038/nsmb.2115. URL https://doi.org/10.1038/nsmb.2115.

[120] B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120(1):15–20, 2005. ISSN 0092-8674 (Print) 0092-8674. doi: 10.1016/j.cell.2004.12.035. URL https://doi.org/10.1016/j.cell.2004.12.035.

[121] Andrew Grimson, Kyle Kai-How Farh, Wendy K. Johnston, Philip Garrett-Engele, Lee P. Lim, and David P. Bartel. Microrna targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell*, 27(1):91–105, 2007. ISSN 1097-2765 1097-4164. doi: 10.1016/j.molcel.2007.06.017. URL https://pubmed.ncbi.nlm.nih.gov/17612493https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3800283/.

[122] Jean Hausser and Mihaela Zavolan. Identification and consequences of mirna–target interactions — beyond repression of gene expression. *Nature Reviews Genetics*, 15(9):599–612, 2014. ISSN 1471-0064. doi: 10.1038/nrg3765. URL https://doi.org/10.1038/nrg3765.

[123] Lee P. Lim, Nelson C. Lau, Philip Garrett-Engele, Andrew Grimson, Janell M. Schelter, John Castle, David P. Bartel, Peter S. Linsley, and Jason M. Johnson. Microarray analysis shows that some micrornas downregulate large numbers of target mrnas. *Nature*, 433(7027):769–773, 2005. ISSN 1476-4687. doi: 10.1038/nature03315. URL https://doi.org/10.1038/nature03315.

[124] A. Stark, J. Brennecke, N. Bushati, R. B. Russell, and S. M. Cohen. Animal micrornas confer robustness to gene expression and have a significant impact on 3'utr evolution. *Cell*, 123(6):1133–46, 2005. ISSN 0092-8674 (Print) 0092-8674. doi: 10.1016/j.cell.2005.11.023. URL https://doi.org/10.1016/j.cell.2005.11.023.

[125] Matthias Selbach, Björn Schwanhäusser, Nadine Thierfelder, Zhuo Fang, Raya Khanin, and Nikolaus Rajewsky. Widespread changes in protein synthesis induced by micrornas. *Nature*, 455 (7209):58–63, 2008. ISSN 1476-4687. doi: 10.1038/nature07228. URL https://doi.org/10.1038/nature07228.

[126] Robin C. Friedman, Kyle Kai-How Farh, Christopher B. Burge, and David P. Bartel. Most mammalian mrnas are conserved targets of micrornas. *Genome research*, 19(1):92–105, 2009. ISSN 1088-9051 1549-5477. doi: 10.1101/gr.082701.108. URL https://pubmed.ncbi.nlm.nih.gov/18955434https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2612969/.

[127] Yong Zhao and Deepak Srivastava. A developmental view of microrna function. *Trends in Biochemical Sciences*, 32(4):189–197, 2007. ISSN 0968-0004. doi: 10.1016/j.tibs.2007.02.006. URL https://doi.org/10.1016/j.tibs.2007.02.006.

[128] Bünyamin Akgül and İpek Erdoğan. Intracytoplasmic re-localization of mirisc complexes. *Frontiers in Genetics*, 9(403), 2018. ISSN 1664-8021. doi: 10.3389/fgene.2018.00403. URL https://www.frontiersin.org/article/10.3389/fgene.2018.00403.

[129] Luca F. R. Gebert and Ian J. MacRae. Regulation of microrna function in animals. *Nature Reviews Molecular Cell Biology*, 20(1): 21–37, 2019. ISSN 1471-0080. doi: 10.1038/s41580-018-0045-7. URL https://doi.org/10.1038/s41580-018-0045-7.

[130] Munekazu Yamakuchi, Marcella Ferlito, and Charles J. Lowenstein. mir-34a repression of sirt1 regulates apoptosis. *Proceedings of the National Academy of Sciences*, 105(36):13421–13426, 2008. doi: 10.1073/pnas.0801613105. URL https://www.pnas.org/content/pnas/105/36/13421.full.pdf.

[131] Lorenzo Baronti, Ileana Guzzetti, Parisa Ebrahimi, Sarah Friebe Sandoz, Emilie Steiner, et al. Base-pair conformational switch modulates mir-34a targeting of sirt1 mrna. *Nature*, 583 (7814):139–144, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2336-3. URL https://doi.org/10.1038/s41586-020-2336-3.

[132] T. E. Swingler, L. T. Le, S. Gardner, M. J. Barter, D. A. Young, T. Dalmay, and I. M. Clark. Microrna-455 targets multiple genes in the wnt signalling pathway. *Osteoarthritis and Cartilage*, 24: S346, 2016. ISSN 1063-4584. doi: 10.1016/j.joca.2016.01.621. URL https://doi.org/10.1016/j.joca.2016.01.621.

[133] Dominic Henn, Masood Abu-Halima, Dominik Wermke, Florian Falkner, Benjamin Thomas, et al. Microrna-regulated pathways of flow-stimulated angiogenesis and vascular remodeling in vivo. *Journal of Translational Medicine*, 17(1):22, 2019. ISSN 1479-5876. doi: 10.1186/s12967-019-1767-9. URL https://doi.org/10.1186/s12967-019-1767-9.

[134] Stefania Rosano, Davide Corà, Sushant Parab, Serena Zaffuto, Claudio Isella, et al. A regulatory microrna network controls endothelial cell phenotypic switch during sprouting angiogenesis. *eLife*, 9:e48095, 2020. ISSN 2050-084X. doi: 10.7554/eLife.48095. URL https://doi.org/10.7554/eLife.48095.

[135] Clara Benna, Senthilkumar Rajendran, Marco Rastrelli, and Simone Mocellin. mirna deregulation targets specific pathways in leiomyosarcoma development: an in silico analysis. *Journal of Translational Medicine*, 17(1):153, 2019. ISSN 1479-5876. doi: 10.1186/s12967-019-1907-2. URL https://doi.org/10.1186/s12967-019-1907-2.

[136] Morten T. Venø, Cristina R. Reschke, Gareth Morris, Niamh M. C. Connolly, Junyi Su, et al. A systems approach delivers a functional microrna catalog and expanded targets for seizure suppression in temporal lobe epilepsy. *Proceedings of the National Academy of Sciences*, 117(27):15977–15988, 2020. doi: 10.1073/pnas.1919313117. URL https://www.pnas.org/content/pnas/117/27/15977.full.pdf.

[137] Francesca M. Buffa, Carme Camps, Laura Winchester, Cameron E. Snell, Harriet E. Gee, Helen Sheldon, Marian Taylor, Adrian L. Harris, and Jiannis Ragoussis. microrna-associated progression pathways and potential therapeutic targets identified by integrated mrna and microrna expression profiling in breast cancer. *Cancer Research*, 71(17):5635–5645, 2011. doi: 10.1158/0008-5472.Can-11-0489. URL https://cancerres.aacrjournals.org/content/canres/71/17/5635.full.pdf.

[138] Claire Josse, Nassim Bouznad, Pierre Geurts, Alexandre Irrthum, Vân Anh Huynh-Thu, Laurence Servais, Alexandre Hego, Philippe Delvenne, Vincent Bours, and Cécile Oury. Identification of a microrna landscape targeting the pi3k/akt signaling pathway in inflammation-induced colorectal carcinogenesis. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 306(3):G229–G243, 2014. doi: 10.1152/ajpgi.00484.2012. URL https://journals.physiology.org/doi/abs/10.1152/ajpgi.00484.2012.

[139] Vinod Kumar, Virender Kumar, Amit Kumar Chaudhary, Donald W. Coulter, Timothy McGuire, and Ram I. Mahato. Impact of mirna-mrna profiling and their correlation on medulloblastoma tumorigenesis. *Molecular Therapy - Nucleic*

Acids, 12:490–503, 2018. ISSN 2162-2531. doi: 10.1016/j.omtn.2018.06.004. URL https://www.sciencedirect.com/science/article/pii/S2162253118301318.

[140] Minh T. N. Le, Huangming Xie, Beiyan Zhou, Poh Hui Chia, Pamela Rizk, Moonkyoung Um, Gerald Udolph, Henry Yang, Bing Lim, and Harvey F. Lodish. Microrna-125b promotes neuronal differentiation in human cells by repressing multiple targets. *Molecular and Cellular Biology*, 29(19):5290–5305, 2009. doi: 10.1128/mcb.01694-08. URL https://mcb.asm.org/content/mcb/29/19/5290.full.pdf.

[141] Sana Mujahid, Tanya Logvinenko, MaryAnn V. Volpe, and Heber C. Nielsen. mirna regulated pathways in late stage murine lung development. *BMC Developmental Biology*, 13(1):13, 2013. ISSN 1471-213X. doi: 10.1186/1471-213X-13-13. URL https://doi.org/10.1186/1471-213X-13-13.

[142] Antony Rodriguez, Elena Vigorito, Simon Clare, Madhuri V. Warren, Philippe Couttet, et al. Requirement of bic/microrna-155 for normal immune function. *Science*, 316(5824):608–611, 2007. doi: 10.1126/science.1139253. URL https://science.sciencemag.org/content/sci/316/5824/608.full.pdf.

[143] Claire E. Gustafson, Mary M. Cavanagh, Jun Jin, Cornelia M. Weyand, and Jörg J. Goronzy. Functional pathways regulated by microrna networks in cd8 t-cell aging. *Aging Cell*, 18(1):e12879, 2019. ISSN 1474-9718. doi: 10.1111/acel.12879. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/acel.12879.

[144] Rotem Ben-Hamo and Sol Efroni. Microrna regulation of molecular pathways as a generic mechanism and as a core disease phenotype. *Oncotarget*, 6(3), 2015. ISSN 1949-2553. doi: 10.18632/oncotarget.2734. URL https://www.oncotarget.com/article/2734/text/.

[145] Tim Kehl, Christina Backes, Fabian Kern, Tobias Fehlmann, Nicole Ludwig, Eckart Meese, Hans-Peter Lenhof, and Andreas Keller. About mirnas, mirna seeds, target genes and target pathways. *Oncotarget*, 8(63):107167–107175, 2017. ISSN 1949-2553. doi: 10.18632/oncotarget.22363. URL https://pubmed.ncbi.nlm.nih.gov/29291020https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5739805/.

[146] Lu Zhang, Yi Liao, and Liling Tang. Microrna-34 family: a potential tumor suppressor and therapeutic candidate in cancer. *Journal of Experimental & Clinical Cancer Research*, 38(1):53, 2019. ISSN 1756-9966. doi: 10.1186/s13046-019-1059-5. URL https://doi.org/10.1186/s13046-019-1059-5.

[147] Ioannis S. Vlachos, Konstantinos Zagganas, Maria D. Paraskevopoulou, Georgios Georgakilas, Dimitra Karagkouni,

Thanasis Vergoulis, Theodore Dalamagas, and Artemis G. Hatzigeorgiou. Diana-mirpath v3.0: deciphering microrna function with experimental support. *Nucleic Acids Research*, 43(W1): W460–W466, 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv403. URL https://doi.org/10.1093/nar/gkv403.

[148] Colin C. Pritchard, Heather H. Cheng, and Muneesh Tewari. Microrna profiling: approaches and considerations. *Nature Reviews Genetics*, 13(5):358–369, 2012. ISSN 1471-0064. doi: 10.1038/nrg3198. URL https://doi.org/10.1038/nrg3198.

[149] Caifu Chen, Ruoying Tan, Linda Wong, Richard Fekete, and Jason Halsey. *Quantitation of MicroRNAs by Real-Time RT-qPCR*, pages 113–134. Humana Press, Totowa, NJ, 2011. ISBN 978-1-60761-944-4. doi: 10.1007/978-1-60761-944-4_8. URL https://doi.org/10.1007/978-1-60761-944-4_8.

[150] Caifu Chen, Dana A. Ridzon, Adam J. Broomer, Zhaohui Zhou, Danny H. Lee, et al. Real-time quantification of micrornas by stem–loop rt–pcr. *Nucleic Acids Research*, 33(20):e179–e179, 2005. ISSN 0305-1048. doi: 10.1093/nar/gni178. URL https://doi.org/10.1093/nar/gni178.

[151] Diego A Forero, Yeimy González-Giraldo, Luis J Castro-Vega, and George E Barreto. qpcr-based methods for expression analysis of mirnas. *BioTechniques*, 67(4):192–199, 2019. doi: 10.2144/btn-2019-0065. URL https://www.future-science.com/doi/abs/10.2144/btn-2019-0065.

[152] Chang-Gong Liu, George Adrian Calin, Stefano Volinia, and Carlo M. Croce. Microrna expression profiling using microarrays. *Nature Protocols*, 3(4):563–578, 2008. ISSN 1750-2799. doi: 10.1038/nprot.2008.14. URL https://doi.org/10.1038/nprot.2008.14.

[153] J. Michael Thomson, Joel S. Parker, and Scott M. Hammond. *Microarray Analysis of miRNA Gene Expression*, volume 427, pages 107–122. Academic Press, 2007. ISBN 0076-6879. doi: 10.1016/S0076-6879(07)27006-5. URL https://www.sciencedirect.com/science/article/pii/S0076687907270065.

[154] Y. Li, T. Fehlmann, A. Borcherding, S. Drmanac, S. Liu, et al. Coolmps: evaluation of antibody labeling based massively parallel non-coding rna sequencing. *Nucleic Acids Res*, 49(2): e10, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa1122. URL https://doi.org/10.1093/nar/gkaa1122.

[155] Eric A. Miska, Ezequiel Alvarez-Saavedra, Matthew Townsend, Akira Yoshii, Nenad Šestan, Pasko Rakic, Martha Constantine-Paton, and H. Robert Horvitz. Microarray analysis of microrna expression in the developing mammalian brain. *Genome Biology*,

5(9):R68, 2004. ISSN 1474-760X. doi: 10.1186/gb-2004-5-9-r68. URL https://doi.org/10.1186/gb-2004-5-9-r68.

[156] Andreas Keller, Trine Rounge, Christina Backes, Nicole Ludwig, Randi Gislefoss, Petra Leidinger, Hilde Langseth, and Eckart Meese. Sources to variability in circulating human mirna signatures. *RNA Biology*, 14(12):1791–1798, 2017. ISSN 1547-6286. doi: 10.1080/15476286.2017.1367888. URL https://doi.org/10.1080/15476286.2017.1367888.

[157] Hanni Willenbrock, Jesper Salomon, Rolf Søkilde, Kim Bundvig Barken, Thomas Nøhr Hansen, Finn Cilius Nielsen, Søren Møller, and Thomas Litman. Quantitative mirna expression analysis: Comparing microarrays with next-generation sequencing. *RNA*, 15(11):2028–2034, 2009. doi: 10.1261/rna.1699809. URL http://rnajournal.cshlp.org/content/15/11/2028.abstract.

[158] Anna Git, Heidi Dvinge, Mali Salmon-Divon, Michelle Osborne, Claudia Kutter, James Hadfield, Paul Bertone, and Carlos Caldas. Systematic comparison of microarray profiling, real-time pcr, and next-generation sequencing technologies for measuring differential microrna expression. *RNA*, 16(5):991–1006, 2010. doi: 10.1261/rna.1947110. URL http://rnajournal.cshlp.org/content/16/5/991.abstract.

[159] Dena Leshkowitz, Shirley Horn-Saban, Yisrael Parmet, and Ester Feldmesser. Differences in microrna detection levels are technology and sequence dependent. *RNA*, 19(4):527–538, 2013. doi: 10.1261/rna.036475.112. URL http://rnajournal.cshlp.org/content/19/4/527.abstract.

[160] Kenneth W. Witwer and Marc K. Halushka. Toward the promise of micrornas – enhancing reproducibility and rigor in microrna research. *RNA Biology*, 13(11):1103–1116, 2016. ISSN 1547-6286. doi: 10.1080/15476286.2016.1236172. URL https://doi.org/10.1080/15476286.2016.1236172.

[161] Marc R. Friedländer, Wei Chen, Catherine Adamidi, Jonas Maaskola, Ralf Einspanier, Signe Knespel, and Nikolaus Rajewsky. Discovering micrornas from deep sequencing data using mirdeep. *Nature Biotechnology*, 26(4):407–415, 2008. ISSN 1546-1696. doi: 10.1038/nbt1394. URL https://doi.org/10.1038/nbt1394.

[162] Christian Wake, Adam Labadorf, Alexandra Dumitriu, Andrew G. Hoss, Joli Bregu, Kenneth H. Albrecht, Anita L. DeStefano, and Richard H. Myers. Novel microrna discovery using small rna sequencing in post-mortem human brain. *BMC Genomics*, 17(1):776, 2016. ISSN 1471-2164. doi: 10.1186/s12864-016-3114-3. URL https://doi.org/10.1186/s12864-016-3114-3.

[163] Michael Hagemann-Jensen, Ilgar Abdullayev, Rickard Sandberg, and Omid R. Faridani. Small-seq for single-cell small-rna sequencing. *Nature Protocols*, 13(10):2407–2424, 2018. ISSN 1750-2799. doi: 10.1038/s41596-018-0049-y. URL https://doi.org/10.1038/s41596-018-0049-y.

[164] Ryan K. Y. Wong, Meabh MacMahon, Jayne V. Woodside, and David A. Simpson. A comparison of rna extraction and sequencing protocols for detection of small rnas in plasma. *BMC Genomics*, 20(1):446, 2019. ISSN 1471-2164. doi: 10.1186/s12864-019-5826-7. URL https://doi.org/10.1186/s12864-019-5826-7.

[165] Anna M. L. Coenen-Stass, Iddo Magen, Tony Brooks, Iddo Z. Ben-Dov, Linda Greensmith, Eran Hornstein, and Pietro Fratta. Evaluation of methodologies for microrna biomarker detection by next generation sequencing. *RNA Biology*, 15(8):1133–1145, 2018. ISSN 1547-6286. doi: 10.1080/15476286.2018.1514236. URL https://www.tandfonline.com/doi/abs/10.1080/15476286.2018.1514236.

[166] Fatima Heinicke, Xiangfu Zhong, Manuela Zucknick, Johannes Breidenbach, Arvind Y. M. Sundaram, et al. Systematic assessment of commercially available low-input mirna library preparation kits. *RNA Biology*, 17(1):75–86, 2020. ISSN 1547-6286. doi: 10.1080/15476286.2019.1667741. URL https://doi.org/10.1080/15476286.2019.1667741.

[167] Hsi-Yuan Huang, Yang-Chi-Dung Lin, Jing Li, Kai-Yao Huang, Sirjana Shrestha, et al. mirtarbase 2020: updates to the experimentally validated microrna–target interaction database. *Nucleic Acids Research*, 48(D1):D148–D154, 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz896. URL https://doi.org/10.1093/nar/gkz896.

[168] Thomas Clément, Véronique Salone, and Mathieu Rederstorff. *Dual Luciferase Gene Reporter Assays to Study miRNA Function*, pages 187–198. Springer New York, New York, NY, 2015. ISBN 978-1-4939-2547-6. doi: 10.1007/978-1-4939-2547-6_17. URL https://doi.org/10.1007/978-1-4939-2547-6_17.

[169] Bruce R. Branchini, Tara L. Southworth, Danielle M. Fontaine, Dawn Kohrt, Catherine M. Florentine, and Martha J. Grossel. A firefly luciferase dual color bioluminescence reporter assay using two substrates to simultaneously monitor two gene expression events. *Scientific Reports*, 8(1):5990, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-24278-2. URL https://doi.org/10.1038/s41598-018-24278-2.

[170] David S. McNabb, Robin Reed, and Robert A. Marciniak. Dual luciferase assay system for rapid assessment of gene expression in saccharomyces cerevisiae. *Eukaryotic Cell*, 4(9):

1539–1549, 2005. doi: 10.1128/ec.4.9.1539-1549.2005. URL https://ec.asm.org/content/eukcell/4/9/1539.full.pdf.

[171] Daniel W. Thomson, Cameron P. Bracken, and Gregory J. Goodall. Experimental strategies for microrna target identification. *Nucleic Acids Research*, 39(16):6845–6853, 2011. ISSN 0305-1048. doi: 10.1093/nar/gkr330. URL https://doi.org/10.1093/nar/gkr330.

[172] Donald E. Kuhn, Mickey M. Martin, David S. Feldman, Alvin V. Terry, Gerard J. Nuovo, and Terry S. Elton. Experimental validation of mirna targets. *Methods*, 44(1):47–54, 2008. ISSN 1046-2023. doi: 10.1016/j.ymeth.2007.09.005. URL https://www.sciencedirect.com/science/article/pii/S1046202307001703.

[173] Yi Jin, Zujian Chen, Xiqiang Liu, and Xiaofeng Zhou. *Evaluating the MicroRNA Targeting Sites by Luciferase Reporter Gene Assay*, pages 117–127. Humana Press, Totowa, NJ, 2013. ISBN 978-1-62703-083-0. doi: 10.1007/978-1-62703-083-0_10. URL https://doi.org/10.1007/978-1-62703-083-0_10.

[174] Alejandro Sarrion-Perdigones, Lyra Chang, Yezabel Gonzalez, Tatiana Gallego-Flores, Damian W. Young, and Koen J. T. Venken. Examining multiple cellular pathways at once using multiplex hextuple luciferase assaying. *Nature Communications*, 10(1):5710, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-13651-y. URL https://doi.org/10.1038/s41467-019-13651-y.

[175] Richard L. Moyle, Lilia C. Carvalhais, Lara-Simone Pretorius, Ekaterina Nowak, Gayathery Subramaniam, Jessica Dalton-Morgan, and Peer M. Schenk. An optimized transient dual luciferase assay for quantifying microrna directed repression of targeted sequences. *Frontiers in Plant Science*, 8(1631), 2017. ISSN 1664-462X. doi: 10.3389/fpls.2017.01631. URL https://www.frontiersin.org/article/10.3389/fpls.2017.01631.

[176] Justin M. Wolter, Kasuen Kotagama, Alexandra C. Pierre-Bez, Mari Firago, and Marco Mangone. $3'$life: a functional assay to detect mirna targets in high-throughput. *Nucleic Acids Research*, 42(17):e132–e132, 2014. ISSN 0305-1048. doi: 10.1093/nar/gku626. URL https://doi.org/10.1093/nar/gku626.

[177] Francisca Alcaraz-Pérez, Victoriano Mulero, and María L. Cayuela. Application of the dual-luciferase reporter assay to the analysis of promoter activity in zebrafish embryos. *BMC Biotechnology*, 8(1):81, 2008. ISSN 1472-6750. doi: 10.1186/1472-6750-8-81. URL https://doi.org/10.1186/1472-6750-8-81.

[178] Jinbo Li and Yan Zhang. Current experimental strategies for intracellular target identification of microrna. *ExRNA*, 1(1):6,

2019. ISSN 2398-0060. doi: 10.1186/s41544-018-0002-9. URL https://doi.org/10.1186/s41544-018-0002-9.

[179] Weijun Liu and Xiaowei Wang. Prediction of functional microrna targets by integrative modeling of microrna binding and target expression data. *Genome Biology*, 20(1):18, 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1629-z. URL https://doi.org/10.1186/s13059-019-1629-z.

[180] Yu Xun, Yingxin Tang, Linmin Hu, Hui Xiao, Shengwen Long, Mengting Gong, Chenxi Wei, Ke Wei, and Shuanglin Xiang. Purification and identification of mirna target sites in genome using dna affinity precipitation. *Frontiers in Genetics*, 10(778), 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00778. URL https://www.frontiersin.org/article/10.3389/fgene.2019.00778.

[181] Aleksandra Helwak and David Tollervey. Mapping the mirna interactome by cross-linking ligation and sequencing of hybrids (clash). *Nature Protocols*, 9(3):711–728, 2014. ISSN 1750-2799. doi: 10.1038/nprot.2014.043. URL https://doi.org/10.1038/nprot.2014.043.

[182] Hans-Hermann Wessels, Svetlana Lebedeva, Antje Hirsekorn, Ricardo Wurmus, Altuna Akalin, Neelanjan Mukherjee, and Uwe Ohler. Global identification of functional microrna-mrna interactions in drosophila. *Nature Communications*, 10(1):1626, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09586-z. URL https://doi.org/10.1038/s41467-019-09586-z.

[183] Pablo Landgraf, Mirabela Rusu, Robert Sheridan, Alain Sewer, Nicola Iovino, et al. A mammalian microrna expression atlas based on small rna library sequencing. *Cell*, 129(7):1401–1414, 2007. ISSN 0092-8674 1097-4172. doi: 10.1016/j.cell.2007.04.040. URL https://pubmed.ncbi.nlm.nih.gov/17604727https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2681231/.

[184] Jessica A Weber, David H Baxter, Shile Zhang, David Y Huang, Kuo How Huang, Ming Jen Lee, David J Galas, and Kai Wang. The microrna spectrum in 12 body fluids. *Clinical Chemistry*, 56(11):1733–1741, 2010. ISSN 0009-9147. doi: 10.1373/clinchem.2010.147405. URL https://doi.org/10.1373/clinchem.2010.147405.

[185] Nicole Ludwig, Petra Leidinger, Kurt Becker, Christina Backes, Tobias Fehlmann, et al. Distribution of mirna expression across human tissues. *Nucleic Acids Research*, 44(8):3865–3877, 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw116. URL https://doi.org/10.1093/nar/gkw116.

[186] Lucia Lorenzi, Hua-Sheng Chiu, Francisco Avila Cobos, Stephen Gross, Pieter-Jan Volders, et al. The rna atlas expands the catalog of human non-coding rnas. *Nature Biotechnology*,

2021. ISSN 1546-1696. doi: 10.1038/s41587-021-00936-1. URL `https://doi.org/10.1038/s41587-021-00936-1`.

[187] Nicole Ludwig, Tobias Fehlmann, Valentina Galata, Andre Franke, Christina Backes, Eckart Meese, and Andreas Keller. Small ncrna-seq results of human tissues: Variations depending on sample integrity. *Clinical Chemistry*, 64(7):1074–1084, 2018. ISSN 0009-9147. doi: 10.1373/clinchem.2017.285767. URL `https://doi.org/10.1373/clinchem.2017.285767`.

[188] T. Fehlmann, C. Backes, M. Pirritano, T. Laufer, V. Galata, et al. The sncrna zoo: a repository for circulating small noncoding rnas in animals. *Nucleic Acids Res*, 47(9):4431–4441, 2019. ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gkz227. URL `https://doi.org/10.1093/nar/gkz227`.

[189] Alina Isakova, Tobias Fehlmann, Andreas Keller, and Stephen R. Quake. A mouse tissue atlas of small noncoding rna. *Proceedings of the National Academy of Sciences*, 117(41):25634–25645, 2020. doi: 10.1073/pnas.2002277117. URL `https://www.pnas.org/content/pnas/117/41/25634.full.pdf`.

[190] Eva C. Schwarz, Christina Backes, Arne Knörck, Nicole Ludwig, Petra Leidinger, et al. Deep characterization of blood cell mirnomes by ngs. *Cellular and Molecular Life Sciences*, 73(16): 3169–3181, 2016. ISSN 1420-9071. doi: 10.1007/s00018-016-2154-9. URL `https://doi.org/10.1007/s00018-016-2154-9`.

[191] Simonas Juzenas, Geetha Venkatesh, Matthias Hübenthal, Marc P. Hoeppner, Zhipei Gracie Du, et al. A comprehensive, cell specific microrna catalogue of human peripheral blood. *Nucleic Acids Research*, 45(16):9290–9301, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx706. URL `https://doi.org/10.1093/nar/gkx706`.

[192] Anne Hecksteden, Petra Leidinger, Christina Backes, Stefanie Rheinheimer, Mark Pfeiffer, et al. mirnas and sports: tracking training status and potentially confounding diagnoses. *Journal of Translational Medicine*, 14(1):219, 2016. ISSN 1479-5876. doi: 10.1186/s12967-016-0974-x. URL `https://doi.org/10.1186/s12967-016-0974-x`.

[193] F. Kern, N. Ludwig, C. Backes, E. Maldener, T. Fehlmann, A. Suleymanov, E. Meese, A. Hecksteden, A. Keller, and T. Meyer. Systematic assessment of blood-borne micrornas highlights molecular profiles of endurance sport and carbohydrate uptake. *Cells*, 8(9), 2019. ISSN 2073-4409. doi: 10.3390/cells8091045. URL `https://doi.org/10.3390/cells8091045`.

[194] Nicole Noren Hooten, Kotb Abdelmohsen, Myriam Gorospe, Ngozi Ejiogu, Alan B. Zonderman, and Michele K. Evans. microrna expression patterns reveal differential expression of target genes with age. *PLOS ONE*, 5(5):e10724, 2010.

doi: 10.1371/journal.pone.0010724. URL https://doi.org/10.1371/journal.pone.0010724.

[195] Thalyana Smith-Vikos and Frank J. Slack. Micrornas and their roles in aging. *Journal of Cell Science*, 125(1):7–17, 2012. ISSN 0021-9533. doi: 10.1242/jcs.099200. URL https://doi.org/10.1242/jcs.099200.

[196] Jung Hwa Jin and Suh Yousin. Microrna in aging: From discovery to biology. *Current Genomics*, 13(7):548–557, 2012. ISSN 1389-2029/1875-5488. doi: 10.2174/138920212803251436. URL http://www.eurekaselect.com/node/103215/article.

[197] Akiko Kogure, Masaharu Uno, Takako Ikeda, and Eisuke Nishida. The microrna machinery regulates fasting-induced changes in gene expression and longevity in caenorhabditis elegans. *Journal of Biological Chemistry*, 292(27):11300–11309, 2017. ISSN 0021-9258. doi: 10.1074/jbc.M116.765065. URL https://doi.org/10.1074/jbc.M116.765065.

[198] Jingjing Du, Peiwen Zhang, Mailin Gan, Xue Zhao, Yan Xu, et al. Microrna-204-5p regulates 3t3-l1 preadipocyte proliferation, apoptosis and differentiation. *Gene*, 668:1–7, 2018. ISSN 0378-1119. doi: 10.1016/j.gene.2018.05.036. URL https://www.sciencedirect.com/science/article/pii/S0378111918305250.

[199] Qiuzhong Zhou, Qianfen Wan, Yuxi Jiang, Jin Liu, Li Qiang, and Lei Sun. A landscape of murine long non-coding rnas reveals the leading transcriptome alterations in adipose tissue during aging. *Cell Reports*, 31(8), 2020. ISSN 2211-1247. doi: 10.1016/j.celrep.2020.107694. URL https://doi.org/10.1016/j.celrep.2020.107694.

[200] Ling Jin, Qirui Song, Weili Zhang, Bin Geng, and Jun Cai. Roles of long noncoding rnas in aging and aging complications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1865(7):1763–1771, 2019. ISSN 0925-4439. doi: 10.1016/j.bbadis.2018.09.021. URL https://www.sciencedirect.com/science/article/pii/S0925443918303569.

[201] Owen R. Jones, Alexander Scheuerlein, Roberto Salguero-Gómez, Carlo Giovanni Camarda, Ralf Schaible, et al. Diversity of ageing across the tree of life. *Nature*, 505(7482):169–173, 2014. ISSN 1476-4687. doi: 10.1038/nature12789. URL https://doi.org/10.1038/nature12789.

[202] Xiao Dong, Brandon Milholland, and Jan Vijg. Evidence for a limit to human lifespan. *Nature*, 538(7624):257–259, 2016. ISSN 1476-4687. doi: 10.1038/nature19793. URL https://doi.org/10.1038/nature19793.

[203] David Melzer, Luke C. Pilling, and Luigi Ferrucci. The genetics of human ageing. *Nature Reviews Genetics*, 21(2):88–101, 2020. ISSN 1471-0064. doi: 10.1038/s41576-019-0183-6. URL https://doi.org/10.1038/s41576-019-0183-6.

[204] Weiqi Zhang, Jing Qu, Guang-Hui Liu, and Juan Carlos Izpisua Belmonte. The ageing epigenome and its rejuvenation. *Nature Reviews Molecular Cell Biology*, 21(3):137–150, 2020. ISSN 1471-0080. doi: 10.1038/s41580-019-0204-5. URL https://doi.org/10.1038/s41580-019-0204-5.

[205] Arnold B. Mitnitski, Alexander J. Mogilner, and Kenneth Rockwood. Accumulation of deficits as a proxy measure of aging. *TheScientificWorldJOURNAL*, 1:321027, 2001. ISSN 2356-6140. doi: 10.1100/tsw.2001.58. URL https://doi.org/10.1100/tsw.2001.58.

[206] Angela Y. Chang, Vegard F. Skirbekk, Stefanos Tyrovolas, Nicholas J. Kassebaum, and Joseph L. Dieleman. Measuring population ageing: an analysis of the global burden of disease study 2017. *The Lancet Public Health*, 4(3):e159–e167, 2019. ISSN 2468-2667. doi: 10.1016/S2468-2667(19)30019-2. URL https://doi.org/10.1016/S2468-2667(19)30019-2.

[207] Handan Melike Dönertaş, Daniel K. Fabian, Matías Fuentealba, Linda Partridge, and Janet M. Thornton. Common genetic associations between age-related diseases. *Nature Aging*, 1(4):400–412, 2021. ISSN 2662-8465. doi: 10.1038/s43587-021-00051-5. URL https://doi.org/10.1038/s43587-021-00051-5.

[208] Efraim Jaul and Jeremy Barron. Age-related diseases and clinical and public health implications for the 85 years old and over population. *Frontiers in Public Health*, 5(335), 2017. ISSN 2296-2565. doi: 10.3389/fpubh.2017.00335. URL https://www.frontiersin.org/article/10.3389/fpubh.2017.00335.

[209] Yujun Hou, Xiuli Dan, Mansi Babbar, Yong Wei, Steen G. Hasselbalch, Deborah L. Croteau, and Vilhelm A. Bohr. Ageing as a risk factor for neurodegenerative disease. *Nature Reviews Neurology*, 15(10):565–581, 2019. ISSN 1759-4766. doi: 10.1038/s41582-019-0244-7. URL https://doi.org/10.1038/s41582-019-0244-7.

[210] Maxwell L. Elliott, Avshalom Caspi, Renate M. Houts, Antony Ambler, Jonathan M. Broadbent, et al. Disparities in the pace of biological aging among midlife adults of the same chronological age have implications for future frailty risk and policy. *Nature Aging*, 1(3):295–308, 2021. ISSN 2662-8465. doi: 10.1038/s43587-021-00044-4. URL https://doi.org/10.1038/s43587-021-00044-4.

[211] Thomas A. Rando and Tony Wyss-Coray. Asynchronous, contagious and digital aging. *Nature Aging*, 1(1):29–35, 2021.

ISSN 2662-8465. doi: 10.1038/s43587-020-00015-1. URL https://doi.org/10.1038/s43587-020-00015-1.

[212] Claudio Franceschi, Paolo Garagnani, Cristina Morsiani, Maria Conte, Aurelia Santoro, Andrea Grignolio, Daniela Monti, Miriam Capri, and Stefano Salvioli. The continuum of aging and age-related diseases: Common mechanisms but different rates. *Frontiers in Medicine*, 5(61), 2018. ISSN 2296-858X. doi: 10.3389/fmed.2018.00061. URL https://www.frontiersin.org/article/10.3389/fmed.2018.00061.

[213] L. Partridge and N. H. Barton. Optimally, mutation and the evolution of ageing. *Nature*, 362(6418):305–311, 1993. ISSN 1476-4687. doi: 10.1038/362305a0. URL https://doi.org/10.1038/362305a0.

[214] Shuling Song, Eric W. F. Lam, Tamara Tchkonia, James L. Kirkland, and Yu Sun. Senescent cells: Emerging targets for human aging and age-related diseases. *Trends in Biochemical Sciences*, 45(7):578–592, 2020. ISSN 0968-0004. doi: 10.1016/j.tibs.2020.03.008. URL https://doi.org/10.1016/j.tibs.2020.03.008.

[215] Mariela Jaskelioff, Florian L. Muller, Ji-Hye Paik, Emily Thomas, Shan Jiang, et al. Telomerase reactivation reverses tissue degeneration in aged telomerase-deficient mice. *Nature*, 469(7328): 102–106, 2011. ISSN 1476-4687. doi: 10.1038/nature09603. URL https://doi.org/10.1038/nature09603.

[216] Simon C. Johnson, Peter S. Rabinovitch, and Matt Kaeberlein. mtor is a key modulator of ageing and age-related disease. *Nature*, 493(7432):338–345, 2013. ISSN 1476-4687. doi: 10.1038/nature11861. URL https://doi.org/10.1038/nature11861.

[217] Ulaş Işıldak, Mehmet Somel, Janet M. Thornton, and Handan Melike Dönertaş. Temporal changes in the gene expression heterogeneity during brain development and aging. *Scientific Reports*, 10(1):4080, 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-60998-0. URL https://doi.org/10.1038/s41598-020-60998-0.

[218] Daniel Glass, Ana Viñuela, Matthew N. Davies, Adaikalavan Ramasamy, Leopold Parts, et al. Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome Biology*, 14 (7):R75, 2013. ISSN 1474-760X. doi: 10.1186/gb-2013-14-7-r75. URL https://doi.org/10.1186/gb-2013-14-7-r75.

[219] Sarah E. Harris, Valentina Riggio, Louise Evenden, Tamara Gilchrist, Sarah McCafferty, et al. Age-related gene expression changes, and transcriptome wide association study of physical and cognitive aging traits, in the lothian birth cohort 1936. *Aging*, 9(12):2489–2503, 2017. ISSN 1945-4589. doi: 10.18632/aging.101333. URL https://doi.org/10.18632/aging.101333.

[220] Stephen Frenk and Jonathan Houseley. Gene expression hallmarks of cellular ageing. *Biogerontology*, 19(6):547–566, 2018. ISSN 1573-6768. doi: 10.1007/s10522-018-9750-z. URL https://doi.org/10.1007/s10522-018-9750-z.

[221] Timothy V. Pyrkov, Konstantin Avchaciov, Andrei E. Tarkhov, Leonid I. Menshikov, Andrei V. Gudkov, and Peter O. Fedichev. Longitudinal analysis of blood markers reveals progressive loss of resilience and predicts human lifespan limit. *Nature Communications*, 12(1):2765, 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23014-1. URL https://doi.org/10.1038/s41467-021-23014-1.

[222] Nicholas Schaum, Benoit Lehallier, Oliver Hahn, Róbert Pálovics, Shayan Hosseinzadeh, et al. Ageing hallmarks exhibit organ-specific temporal signatures. *Nature*, 583(7817):596–602, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2499-y. URL https://doi.org/10.1038/s41586-020-2499-y.

[223] Irina Shchukina, Juhi Bagaitkar, Oleg Shpynov, Ekaterina Loginicheva, Sofia Porter, et al. Enhanced epigenetic profiling of classical human monocytes reveals a specific signature of healthy aging in the dna methylome. *Nature Aging*, 1(1):124–141, 2021. ISSN 2662-8465. doi: 10.1038/s43587-020-00002-6. URL https://doi.org/10.1038/s43587-020-00002-6.

[224] Ilias Angelidis, Lukas M. Simon, Isis E. Fernandez, Maximilian Strunz, Christoph H. Mayr, et al. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nature Communications*, 10(1):963, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08831-9. URL https://doi.org/10.1038/s41467-019-08831-9.

[225] Brunilda Balliu, Matthew Durrant, Olivia de Goede, Nathan Abell, Xin Li, et al. Genetic regulation of gene expression and splicing during a 10-year period of human aging. *Genome Biology*, 20(1):230, 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1840-y. URL https://doi.org/10.1186/s13059-019-1840-y.

[226] Sara Ahadi, Wenyu Zhou, Sophia Miryam Schüssler-Fiorenza Rose, M. Reza Sailani, Kévin Contrepois, Monika Avina, Melanie Ashland, Anne Brunet, and Michael Snyder. Personal aging markers and ageotypes revealed by deep longitudinal profiling. *Nature Medicine*, 26(1):83–90, 2020. ISSN 1546-170X. doi: 10.1038/s41591-019-0719-5. URL https://doi.org/10.1038/s41591-019-0719-5.

[227] Benoit Lehallier, David Gate, Nicholas Schaum, Tibor Nanasi, Song Eun Lee, et al. Undulating changes in human plasma proteome profiles across the lifespan. *Nature Medicine*, 25(12):1843–1850, 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0673-2. URL https://doi.org/10.1038/s41591-019-0673-2.

[228] Nicole Almanzar, Jane Antony, Ankit S. Baghel, Isaac Bakerman, Ishita Bansal, et al. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, 583(7817): 590–595, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2496-1. URL https://doi.org/10.1038/s41586-020-2496-1.

[229] Emma S. Chambers, Milica Vukmanovic-Stejic, Barbara B. Shih, Hugh Trahair, Priya Subramanian, et al. Recruitment of inflammatory monocytes by senescent fibroblasts inhibits antigen-specific tissue immunity during human aging. *Nature Aging*, 1(1):101–113, 2021. ISSN 2662-8465. doi: 10.1038/s43587-020-00010-6. URL https://doi.org/10.1038/s43587-020-00010-6.

[230] Lindsay M. Reynolds, Jingzhong Ding, Jackson R. Taylor, Kurt Lohman, Nicola Soranzo, et al. Transcriptomic profiles of aging in purified human immune cells. *BMC Genomics*, 16(1):333, 2015. ISSN 1471-2164. doi: 10.1186/s12864-015-1522-4. URL https://doi.org/10.1186/s12864-015-1522-4.

[231] Yehezqel Elyahu, Idan Hekselman, Inbal Eizenberg-Magar, Omer Berner, Itai Strominger, et al. Aging promotes reorganization of the cd4 t cell landscape toward extreme regulatory and effector phenotypes. *Science Advances*, 5(8):eaaw8330, 2019. doi: 10.1126/sciadv.aaw8330. URL https://advances.sciencemag.org/content/advances/5/8/eaaw8330.full.pdf.

[232] Akira Terao, Anjali Apte-Deshpande, Linda Dousman, Stephen Morairty, Barrett P. Eynon, Thomas S. Kilduff, and Yvonne R. Freund. Immune response gene expression increases in the aging murine hippocampus. *Journal of Neuroimmunology*, 132(1):99–112, 2002. ISSN 0165-5728. doi: 10.1016/S0165-5728(02)00317-X. URL https://doi.org/10.1016/S0165-5728(02)00317-X.

[233] Akash Srivastava, Emanuel Barth, Maria A. Ermolaeva, Madlen Guenther, Christiane Frahm, Manja Marz, and Otto W. Witte. Tissue-specific gene expression changes are associated with aging in mice. *Genomics, Proteomics & Bioinformatics*, 2020. ISSN 1672-0229. doi: 10.1016/j.gpb.2020.12.001. URL https://www.sciencedirect.com/science/article/pii/S1672022920301339.

[234] Kailiang Sun and Eric C. Lai. Adult-specific functions of animal micrornas. *Nature Reviews Genetics*, 14(8):535–548, 2013. ISSN 1471-0064. doi: 10.1038/nrg3471. URL https://doi.org/10.1038/nrg3471.

[235] Donghong Cai and Jing-Dong J. Han. Aging-associated lncrnas are evolutionarily conserved and participate in nfκb signaling. *Nature Aging*, 1(5):438–453, 2021. ISSN 2662-8465. doi: 10.1038/s43587-021-00056-0. URL https://doi.org/10.1038/s43587-021-00056-0.

[236] Stephen M. Eacker, Ted M. Dawson, and Valina L. Dawson. Understanding micrornas in neurodegeneration. *Nature Reviews Neuroscience*, 10(12):837–841, 2009. ISSN 1471-0048. doi: 10.1038/nrn2726. URL https://doi.org/10.1038/nrn2726.

[237] Ana Gámez-Valero, Jaume Campdelacreu, Dolores Vilas, Lourdes Ispierto, Ramón Reñé, Ramiro Álvarez, M. Pilar Armengol, Francesc E. Borràs, and Katrin Beyer. Exploratory study on microrna profiles from plasma-derived extracellular vesicles in alzheimer's disease and dementia with lewy bodies. *Translational Neurodegeneration*, 8(1):31, 2019. ISSN 2047-9158. doi: 10.1186/s40035-019-0169-5. URL https://doi.org/10.1186/s40035-019-0169-5.

[238] Simona Maciotta Rolandin, Mirella Meregalli, and Yvan Torrente. The involvement of micrornas in neurodegenerative diseases. *Frontiers in Cellular Neuroscience*, 7(265), 2013. ISSN 1662-5102. doi: 10.3389/fncel.2013.00265. URL https://www.frontiersin.org/article/10.3389/fncel.2013.00265.

[239] Masashi Abe and Nancy M. Bonini. Micrornas and neurodegeneration: role and impact. *Trends in Cell Biology*, 23(1):30–36, 2013. ISSN 0962-8924. doi: 10.1016/j.tcb.2012.08.013. URL https://doi.org/10.1016/j.tcb.2012.08.013.

[240] Sébastien S. Hébert and Bart De Strooper. mirnas in neurodegeneration. *Science*, 317(5842):1179–1180, 2007. doi: 10.1126/science.1148530. URL https://science.sciencemag.org/content/sci/317/5842/1179.full.pdf.

[241] Salil Sharma and Hui-Chen Lu. micrornas in neurodegeneration: Current findings and potential impacts. *Journal of Alzheimer's disease & Parkinsonism*, 8(1):420, 2018. ISSN 2161-0460. doi: 10.4172/2161-0460.1000420. URL https://pubmed.ncbi.nlm.nih.gov/29862137https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5976447/.

[242] Camille A. Juźwik, Sienna S. Drake, Yang Zhang, Nicolas Paradis-Isler, Alexandra Sylvester, Alexandre Amar-Zifkin, Chelsea Douglas, Barbara Morquette, Craig S. Moore, and Alyson E. Fournier. microrna dysregulation in neurodegenerative diseases: A systematic review. *Progress in Neurobiology*, 182:101664, 2019. ISSN 0301-0082. doi: 10.1016/j.pneurobio.2019.101664. URL https://www.sciencedirect.com/science/article/pii/S030100821830203X.

[243] Dipen Rajgor. Macro roles for micrornas in neurodegenerative diseases. *Non-coding RNA Research*, 3(3):154–159, 2018. ISSN 2468-0540. doi: 10.1016/j.ncrna.2018.07.001. URL https://www.sciencedirect.com/science/article/pii/S2468054018300374.

[244] Valery L. Feigin, Emma Nichols, Tahiya Alam, Marlena S. Bannick, Ettore Beghi, et al. Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Neurology*, 18(5):459–480, 2019. ISSN 1474-4422. doi: 10.1016/S1474-4422(18)30499-X. URL https://doi.org/10.1016/S1474-4422(18)30499-X.

[245] Günther Deuschl, Ettore Beghi, Franz Fazekas, Timea Varga, Kalliopi A. Christoforidi, Eveline Sipido, Claudio L. Bassetti, Theo Vos, and Valery L. Feigin. The burden of neurological diseases in europe: an analysis for the global burden of disease study 2017. *The Lancet Public Health*, 5(10):e551–e567, 2020. ISSN 2468-2667. doi: 10.1016/S2468-2667(20)30190-0. URL https://doi.org/10.1016/S2468-2667(20)30190-0.

[246] Brittany N. Dugger and Dennis W. Dickson. Pathology of neurodegenerative diseases. *Cold Spring Harbor Perspectives in Biology*, 9(7), 2017. doi: 10.1101/cshperspect.a028035. URL http://cshperspectives.cshlp.org/content/9/7/a028035.abstract.

[247] Rebekah M. Ahmed, Yazi D. Ke, Steve Vucic, Lars M. Ittner, William Seeley, John R. Hodges, Olivier Piguet, Glenda Halliday, and Matthew C. Kiernan. Physiological changes in neurodegeneration — mechanistic insights and clinical utility. *Nature Reviews Neurology*, 14(5):259–271, 2018. ISSN 1759-4766. doi: 10.1038/nrneurol.2018.23. URL https://doi.org/10.1038/nrneurol.2018.23.

[248] Werner Poewe, Klaus Seppi, Caroline M. Tanner, Glenda M. Halliday, Patrik Brundin, Jens Volkmann, Anette-Eleonore Schrag, and Anthony E. Lang. Parkinson disease. *Nature Reviews Disease Primers*, 3(1):17013, 2017. ISSN 2056-676X. doi: 10.1038/nrdp.2017.13. URL https://doi.org/10.1038/nrdp.2017.13.

[249] David S. Knopman, Helene Amieva, Ronald C. Petersen, Gäel Chételat, David M. Holtzman, Bradley T. Hyman, Ralph A. Nixon, and David T. Jones. Alzheimer disease. *Nature Reviews Disease Primers*, 7(1):33, 2021. ISSN 2056-676X. doi: 10.1038/s41572-021-00269-y. URL https://doi.org/10.1038/s41572-021-00269-y.

[250] Michael A. DeTure and Dennis W. Dickson. The neuropathological diagnosis of alzheimer's disease. *Molecular Neurodegeneration*, 14(1):32, 2019. ISSN 1750-1326. doi: 10.1186/s13024-019-0333-5. URL https://doi.org/10.1186/s13024-019-0333-5.

[251] Li Gan, Mark R. Cookson, Leonard Petrucelli, and Albert R. La Spada. Converging pathways in neurodegeneration, from genetics to mechanisms. *Nature Neuroscience*, 21(10):1300–1309, 2018. ISSN 1546-1726. doi: 10.1038/s41593-018-0237-7. URL https://doi.org/10.1038/s41593-018-0237-7.

[252] Leonardo Guzman-Martinez, Ricardo B. Maccioni, Víctor Andrade, Leonardo Patricio Navarrete, María Gabriela Pastor, and Nicolas Ramos-Escobar. Neuroinflammation as a common feature of neurodegenerative disorders. *Frontiers in Pharmacology*, 10(1008), 2019. ISSN 1663-9812. doi: 10.3389/fphar.2019.01008. URL https://www.frontiersin.org/article/10.3389/fphar.2019.01008.

[253] Janos Groh, Konrad Knöpper, Panagiota Arampatzi, Xidi Yuan, Lena Lößlein, Antoine-Emmanuel Saliba, Wolfgang Kastenmüller, and Rudolf Martini. Accumulation of cytotoxic t cells in the aged cns leads to axon degeneration and contributes to cognitive and motor decline. *Nature Aging*, 1(4):357–367, 2021. ISSN 2662-8465. doi: 10.1038/s43587-021-00049-z. URL https://doi.org/10.1038/s43587-021-00049-z.

[254] Kelvin K. Leung, Jonathan W. Bartlett, Josephine Barnes, Emily N. Manning, Sebastien Ourselin, and Nick C. Fox. Cerebral atrophy in mild cognitive impairment and alzheimer disease. *Rates and acceleration*, 80(7):648–654, 2013. doi: 10.1212/WNL.0b013e318281ccd3. URL https://n.neurology.org/content/neurology/80/7/648.full.pdf.

[255] Theresa M. Harrison, Renaud La Joie, Anne Maass, Suzanne L. Baker, Kaitlin Swinnerton, et al. Longitudinal tau accumulation and atrophy in aging and alzheimer disease. *Annals of Neurology*, 85(2):229–240, 2019. ISSN 0364-5134. doi: 10.1002/ana.25406. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.25406.

[256] H. Braak and E. Braak. Neuropathological stageing of alzheimer-related changes. *Acta Neuropathologica*, 82(4):239–259, 1991. ISSN 1432-0533. doi: 10.1007/BF00308809. URL https://doi.org/10.1007/BF00308809.

[257] Alberto Serrano-Pozo, Sudeshna Das, and Bradley T. Hyman. Apoe and alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches. *The Lancet Neurology*, 20(1): 68–80, 2021. ISSN 1474-4422. doi: 10.1016/S1474-4422(20)30412-9. URL https://doi.org/10.1016/S1474-4422(20)30412-9.

[258] Julia Marschallinger, Tal Iram, Macy Zardeneta, Song E. Lee, Benoit Lehallier, et al. Lipid-droplet-accumulating microglia represent a dysfunctional and proinflammatory state in the aging brain. *Nature Neuroscience*, 23(2):194–208, 2020. ISSN 1546-1726. doi: 10.1038/s41593-019-0566-1. URL https://doi.org/10.1038/s41593-019-0566-1.

[259] Tyler K. Ulland and Marco Colonna. Trem2 — a key player in microglial biology and alzheimer disease. *Nature Reviews Neurology*, 14(11):667–675, 2018. ISSN 1759-4766. doi: 10.1038/s41582-018-0072-1. URL https://doi.org/10.1038/s41582-018-0072-1.

[260] Willa D. Brenowitz, Peter T. Nelson, Lilah M. Besser, Katherine B. Heller, and Walter A. Kukull. Cerebral amyloid angiopathy and its co-occurrence with alzheimer's disease and other cerebrovascular neuropathologic changes. *Neurobiology of Aging*, 36(10):2702–2708, 2015. ISSN 0197-4580. doi: 10.1016/j.neurobiolaging.2015.06.028. URL https://www.sciencedirect.com/science/article/pii/S0197458015003401.

[261] Axel Montagne, Daniel A. Nation, Abhay P. Sagare, Giuseppe Barisano, Melanie D. Sweeney, et al. Apoe4 leads to blood–brain barrier dysfunction predicting cognitive decline. *Nature*, 581 (7806):71–76, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2247-3. URL https://doi.org/10.1038/s41586-020-2247-3.

[262] William A. Banks, May J. Reed, Aric F. Logsdon, Elizabeth M. Rhea, and Michelle A. Erickson. Healthy aging and the blood–brain barrier. *Nature Aging*, 1(3):243–254, 2021. ISSN 2662-8465. doi: 10.1038/s43587-021-00043-5. URL https://doi.org/10.1038/s43587-021-00043-5.

[263] Melanie D. Sweeney, Abhay P. Sagare, and Berislav V. Zlokovic. Blood–brain barrier breakdown in alzheimer disease and other neurodegenerative disorders. *Nature Reviews Neurology*, 14(3): 133–150, 2018. ISSN 1759-4766. doi: 10.1038/nrneurol.2017.188. URL https://doi.org/10.1038/nrneurol.2017.188.

[264] E. Ray Dorsey, Alexis Elbaz, Emma Nichols, Foad Abd-Allah, Ahmed Abdelalim, et al. Global, regional, and national burden of parkinson's disease, 1990-2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Neurology*, 17(11):939–953, 2018. ISSN 1474-4422. doi: 10.1016/S1474-4422(18)30295-3. URL https://doi.org/10.1016/S1474-4422(18)30295-3.

[265] Emma Nichols, Cassandra E. I. Szoeke, Stein Emil Vollset, Nooshin Abbasi, Foad Abd-Allah, et al. Global, regional, and national burden of alzheimer's disease and other dementias, 1990-2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Neurology*, 18(1):88–106, 2019. ISSN 1474-4422. doi: 10.1016/S1474-4422(18)30403-4. URL https://doi.org/10.1016/S1474-4422(18)30403-4.

[266] Michael L. Kramer and Walter J. Schulz-Schaeffer. Presynaptic α-synuclein aggregates, not lewy bodies, cause neurodegeneration in dementia with lewy bodies. *The Journal of Neuroscience*, 27(6):1405–1410, 2007. doi: 10.1523/jneurosci.4564-06.2007. URL https://www.jneurosci.org/content/jneuro/27/6/1405.full.pdf.

[267] Heiko Braak, Kelly Del Tredici, Udo Rüb, Rob A. I. de Vos, Ernst N. H. Jansen Steur, and Eva Braak. Staging of brain pathology related to sporadic parkinson's disease. *Neurobiology of*

Aging, 24(2):197–211, 2003. ISSN 0197-4580. doi: 10.1016/S0197-4580(02)00065-9. URL https://www.sciencedirect.com/science/article/pii/S0197458002000659.

[268] Walter J. Schulz-Schaeffer. The synaptic pathology of α-synuclein aggregation in dementia with lewy bodies, parkinson's disease and parkinson's disease dementia. *Acta Neuropathologica*, 120(2):131–143, 2010. ISSN 1432-0533. doi: 10.1007/s00401-010-0711-0. URL https://doi.org/10.1007/s00401-010-0711-0.

[269] Walter J. Schulz-Schaeffer. Is cell death primary or secondary in the pathophysiology of idiopathic parkinson's disease? *Biomolecules*, 5(3):1467–1479, 2015. ISSN 2218-273X. doi: 10.3390/biom5031467. URL https://www.mdpi.com/2218-273X/5/3/1467.

[270] Walter J. Schulz-Schaeffer. Neurodegeneration in parkinson disease. *Neurology*, 79(24):2298, 2012. doi: 10.1212/WNL.0b013e318278b6a7. URL http://n.neurology.org/content/79/24/2298.abstract.

[271] Dag Aarsland, Byron Creese, Marios Politis, K. Ray Chaudhuri, Dominic H. ffytche, Daniel Weintraub, and Clive Ballard. Cognitive decline in parkinson disease. *Nature Reviews Neurology*, 13(4):217–231, 2017. ISSN 1759-4766. doi: 10.1038/nrneurol.2017.27. URL https://doi.org/10.1038/nrneurol.2017.27.

[272] K. J. Billingsley, S. Bandres-Ciga, S. Saez-Atienzar, and A. B. Singleton. Genetic risk factors in parkinson's disease. *Cell and Tissue Research*, 373(1):9–20, 2018. ISSN 1432-0878. doi: 10.1007/s00441-018-2817-y. URL https://doi.org/10.1007/s00441-018-2817-y.

[273] Hirotaka Iwaki, Cornelis Blauwendraat, Hampton L. Leonard, Ganqiang Liu, Jodi Maple-Grødem, et al. Genetic risk of parkinson disease and progression. *An analysis of 13 longitudinal cohorts*, 5(4):e348, 2019. doi: 10.1212/nxg.0000000000000348. URL https://ng.neurology.org/content/nng/5/4/e348.full.pdf.

[274] Paul M. A. Antony, Nico J. Diederich, Rejko Krüger, and Rudi Balling. The hallmarks of parkinson's disease. *The FEBS Journal*, 280(23):5981–5993, 2013. ISSN 1742-464X. doi: 10.1111/febs.12335. URL https://febs.onlinelibrary.wiley.com/doi/abs/10.1111/febs.12335.

[275] Callum N. Watson, Antonio Belli, and Valentina Di Pietro. Small non-coding rnas: New class of biomarkers and potential therapeutic targets in neurodegenerative disease. *Frontiers in Genetics*, 10(364), 2019. ISSN 1664-

8021. doi: 10.3389/fgene.2019.00364. URL https://www.frontiersin.org/article/10.3389/fgene.2019.00364.

[276] Claudia Ramaker, Johan Marinus, Anne Margarethe Stiggelbout, and Bob Johannes van Hilten. Systematic evaluation of rating scales for impairment and disability in parkinson's disease. *Movement Disorders*, 17(5):867–876, 2002. ISSN 0885-3185. doi: 10.1002/mds.10248. URL https://movementdisorders.onlinelibrary.wiley.com/doi/abs/10.1002/mds.10248.

[277] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. The unified parkinson's disease rating scale (updrs): Status and recommendations. *Movement Disorders*, 18(7):738–750, 2003. ISSN 0885-3185. doi: 10.1002/mds.10473. URL https://movementdisorders.onlinelibrary.wiley.com/doi/abs/10.1002/mds.10473.

[278] Philippe Robert, Steven Ferris, Serge Gauthier, Ralf Ihl, Bengt Winblad, and Frank Tennigkeit. Review of alzheimer's disease scales: is there a need for a new multi-domain scale for therapy evaluation in medical practice? *Alzheimer's Research & Therapy*, 2(4):24, 2010. ISSN 1758-9193. doi: 10.1186/alzrt48. URL https://doi.org/10.1186/alzrt48.

[279] Oskar Hansson. Biomarkers for neurodegenerative diseases. *Nature Medicine*, 2021. ISSN 1546-170X. doi: 10.1038/s41591-021-01382-x. URL https://doi.org/10.1038/s41591-021-01382-x.

[280] Shorena Janelidze, Niklas Mattsson, Sebastian Palmqvist, Ruben Smith, Thomas G. Beach, et al. Plasma p-tau181 in alzheimer's disease: relationship to other biomarkers, differential diagnosis, neuropathology and longitudinal progression to alzheimer's dementia. *Nature Medicine*, 26(3):379–386, 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0755-1. URL https://doi.org/10.1038/s41591-020-0755-1.

[281] Nicholas C. Cullen, Antoine Leuzy, Sebastian Palmqvist, Shorena Janelidze, Erik Stomrud, et al. Individualized prognosis of cognitive decline and dementia in mild cognitive impairment based on plasma biomarker combinations. *Nature Aging*, 1(1):114–123, 2021. ISSN 2662-8465. doi: 10.1038/s43587-020-00003-5. URL https://doi.org/10.1038/s43587-020-00003-5.

[282] Harald Hampel, Sid E. O'Bryant, José L. Molinuevo, Henrik Zetterberg, Colin L. Masters, Simone Lista, Steven J. Kiddle, Richard Batrla, and Kaj Blennow. Blood-based biomarkers for alzheimer disease: mapping the road to the clinic. *Nature Reviews Neurology*, 14(11):639–652, 2018. ISSN 1759-4766. doi:

10.1038/s41582-018-0079-7. URL https://doi.org/10.1038/s41582-018-0079-7.

[283] Xiaochen Xi, Tianxiao Li, Yiming Huang, Jiahui Sun, Yumin Zhu, Yang Yang, and Zhi J. Lu. Rna biomarkers: Frontier of precision medicine for cancer. *Non-Coding RNA*, 3(1), 2017. ISSN 2311-553X. doi: 10.3390/ncrna3010009. URL https://doi.org/10.3390/ncrna3010009.

[284] Andreas Keller, Petra Leidinger, Andrea Bauer, Abdou ElSharawy, Jan Haas, et al. Toward the blood-borne mirnome of human diseases. *Nature Methods*, 8(10):841–843, 2011. ISSN 1548-7105. doi: 10.1038/nmeth.1682. URL https://doi.org/10.1038/nmeth.1682.

[285] Patrick S. Mitchell, Rachael K. Parkin, Evan M. Kroh, Brian R. Fritz, Stacia K. Wyman, et al. Circulating micrornas as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences*, 105(30):10513–10518, 2008. doi: 10.1073/pnas.0804549105. URL https://www.pnas.org/content/pnas/105/30/10513.full.pdf.

[286] Stępień E, Marina C. Costa, Szczepan Kurc, Anna Drożdż, Nuno Cortez-Dias, and Francisco J. Enguita. The circulating non-coding rna landscape for biomarker research: lessons and prospects from cardiovascular diseases. *Acta Pharmacologica Sinica*, 39(7):1085–1099, 2018. ISSN 1745-7254. doi: 10.1038/aps.2018.35. URL https://doi.org/10.1038/aps.2018.35.

[287] Tobias Fehlmann, Nicole Ludwig, Christina Backes, Eckart Meese, and Andreas Keller. Distribution of microrna biomarker candidates in solid tissues and body fluids. *RNA Biology*, 13(11):1084–1088, 2016. ISSN 1547-6286. doi: 10.1080/15476286.2016.1234658. URL https://doi.org/10.1080/15476286.2016.1234658.

[288] Christine Happel, Aniruddha Ganguly, and Danilo A. Tagle. Extracellular rnas as potential biomarkers for cancer. *Journal of Cancer Metastasis and Treatment*, 6:32, 2020. ISSN 2454-2857. doi: 10.20517/2394-4722.2020.71. URL http://dx.doi.org/10.20517/2394-4722.2020.71.

[289] Colin C. Pritchard, Evan Kroh, Brent Wood, Jason D. Arroyo, Katy J. Dougherty, Melanie M. Miyaji, Jonathan F. Tait, and Muneesh Tewari. Blood cell origin of circulating micrornas: A cautionary note for cancer biomarker studies. *Cancer Prevention Research*, 5(3):492–497, 2012. doi: 10.1158/1940-6207.Capr-11-0370. URL https://cancerpreventionresearch.aacrjournals.org/content/canprevres/5/3/492.full.pdf.

[290] Christina Backes, Eckart Meese, and Andreas Keller. Specific mirna disease biomarkers in blood, serum and plasma: Challenges and prospects. *Molecular Diagnosis & Therapy*, 20(6): 509–518, 2016. ISSN 1179-2000. doi: 10.1007/s40291-016-0221-4. URL https://doi.org/10.1007/s40291-016-0221-4.

[291] Ana Mompeón, Luis Ortega-Paz, Xavier Vidal-Gómez, Tiago Januario Costa, Daniel Pérez-Cremades, et al. Disparate mirna expression in serum and plasma of patients with acute myocardial infarction: a systematic and paired comparative analysis. *Scientific Reports*, 10(1):5373, 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-61507-z. URL https://doi.org/10.1038/s41598-020-61507-z.

[292] Ana E. Jenike and Marc K. Halushka. mir-21: a non-specific biomarker of all maladies. *Biomarker Research*, 9(1):18, 2021. ISSN 2050-7771. doi: 10.1186/s40364-021-00272-1. URL https://doi.org/10.1186/s40364-021-00272-1.

[293] Baqer A. Haider, Alexander S. Baras, Matthew N. McCall, Joshua A. Hertel, Toby C. Cornish, and Marc K. Halushka. A critical evaluation of microrna biomarkers in non-neoplastic disease. *PLOS ONE*, 9(2):e89565, 2014. doi: 10.1371/journal.pone.0089565. URL https://doi.org/10.1371/journal.pone.0089565.

[294] Shirin Moradifard, Moslem Hoseinbeyki, Shahla Mohammad Ganji, and Zarrin Minuchehr. Analysis of microrna and gene expression profiles in alzheimer's disease: A meta-analysis approach. *Scientific Reports*, 8(1):4767, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-20959-0. URL https://doi.org/10.1038/s41598-018-20959-0.

[295] Tobias Fehlmann, Mustafa Kahraman, Nicole Ludwig, Christina Backes, Valentina Galata, et al. Evaluating the use of circulating microrna profiles for lung cancer detection in symptomatic patients. *JAMA Oncology*, 6(5):714–723, 2020. ISSN 2374-2437. doi: 10.1001/jamaoncol.2020.0001. URL https://doi.org/10.1001/jamaoncol.2020.0001.

[296] Petra Leidinger, Christina Backes, Stephanie Deutscher, Katja Schmitt, Sabine C. Mueller, et al. A blood based 12-mirna signature of alzheimer disease patients. *Genome Biology*, 14(7): R78, 2013. ISSN 1474-760X. doi: 10.1186/gb-2013-14-7-r78. URL https://doi.org/10.1186/gb-2013-14-7-r78.

[297] Nicolai A. Schultz, Christian Dehlendorff, Benny V. Jensen, Jon K. Bjerregaard, Kaspar R. Nielsen, et al. Microrna biomarkers in whole blood for detection of pancreatic cancer. *JAMA*, 311(4):392–404, 2014. ISSN 0098-7484. doi: 10.1001/jama.2013.284664. URL https://doi.org/10.1001/jama.2013.284664.

[298] Bill Qi, Laura M Fiori, Gustavo Turecki, and Yannis J Trakadis. Machine learning analysis of blood microrna data in major depression: A case-control study for biomarker discovery. *International Journal of Neuropsychopharmacology*, 23(8):505–510, 2020. ISSN 1461-1457. doi: 10.1093/ijnp/pyaa029. URL https://doi.org/10.1093/ijnp/pyaa029.

[299] Francisco Azuaje, Yvan Devaux, and Daniel Wagner. Challenges and standards in reporting diagnostic and prognostic biomarker studies. *Clinical and Translational Science*, 2(2):156–161, 2009. ISSN 1752-8054. doi: 10.1111/j.1752-8062.2008.00075.x. URL https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1111/j.1752-8062.2008.00075.x.

[300] Ravi Dhingra and Ramachandran S. Vasan. Biomarkers in cardiovascular disease: Statistical assessment and section on key novel heart failure biomarkers. *Trends in Cardiovascular Medicine*, 27(2):123–133, 2017. ISSN 1050-1738. doi: 10.1016/j.tcm.2016.07.005. URL https://www.sciencedirect.com/science/article/pii/S1050173816301050.

[301] Leon Tribolet, Emily Kerr, Christopher Cowled, Andrew G. D. Bean, Cameron R. Stewart, Megan Dearnley, and Ryan J. Farr. Microrna biomarkers for infectious diseases: From basic research to biosensing. *Frontiers in Microbiology*, 11(1197), 2020. ISSN 1664-302X. doi: 10.3389/fmicb.2020.01197. URL https://www.frontiersin.org/article/10.3389/fmicb.2020.01197.

[302] Kenta Hyeon Tae Cho, Bing Xu, Cherie Blenkiron, and Mhoyra Fraser. Emerging roles of mirnas in brain development and perinatal brain injury. *Frontiers in Physiology*, 10(227), 2019. ISSN 1664-042X. doi: 10.3389/fphys.2019.00227. URL https://www.frontiersin.org/article/10.3389/fphys.2019.00227.

[303] Chiranjib Chakraborty, Ashish Ranjan Sharma, Garima Sharma, Manojit Bhattacharya, and Sang-Soo Lee. Micrornas: Possible regulatory molecular switch controlling the bbb microenvironment. *Molecular Therapy - Nucleic Acids*, 19:933–936, 2020. ISSN 2162-2531. doi: 10.1016/j.omtn.2019.12.024. URL https://doi.org/10.1016/j.omtn.2019.12.024.

[304] Qiuhong Ji, Yuhua Ji, Jingwen Peng, Xin Zhou, Xinya Chen, Heng Zhao, Tian Xu, Ling Chen, and Yun Xu. Increased brain-specific mir-9 and mir-124 in the serum exosomes of acute ischemic stroke patients. *PLOS ONE*, 11(9):e0163645, 2016. doi: 10.1371/journal.pone.0163645. URL https://doi.org/10.1371/journal.pone.0163645.

[305] S. Swarbrick, N. Wragg, S. Ghosh, and Alexandra Stolzing. Systematic review of mirna as biomarkers in alzheimer's disease. *Molecular Neurobiology*, 56(9):6156–6167, 2019. ISSN 1559-1182. doi: 10.1007/s12035-019-1500-y. URL https://doi.org/10.1007/s12035-019-1500-y.

[306] Ian Fyfe. Rna biomarkers of parkinson disease. *Nature Reviews Neurology*, 17(3):132–132, 2021. ISSN 1759-4766. doi: 10.1038/s41582-021-00470-3. URL `https://doi.org/10.1038/s41582-021-00470-3`.

[307] Ronald B. Postuma and Daniela Berg. Advances in markers of prodromal parkinson disease. *Nature Reviews Neurology*, 12(11): 622–634, 2016. ISSN 1759-4766. doi: 10.1038/nrneurol.2016.152. URL `https://doi.org/10.1038/nrneurol.2016.152`.

[308] Johora Hanna, Gazi S. Hossain, and Jannet Kocerha. The potential for microrna therapeutics and clinical research. *Frontiers in Genetics*, 10(478), 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00478. URL `https://www.frontiersin.org/article/10.3389/fgene.2019.00478`.

[309] Kioomars Saliminejad, Hamid Reza Khorram Khorshid, and Seyed Hamidollah Ghaffari. Why have microrna biomarkers not been translated from bench to clinic? *Future Oncology*, 15 (8):801–803, 2019. doi: 10.2217/fon-2018-0812. URL `https://www.futuremedicine.com/doi/abs/10.2217/fon-2018-0812`.

[310] Jeff Gauthier, Antony T Vincent, Steve J Charette, and Nicolas Derome. A brief history of bioinformatics. *Briefings in Bioinformatics*, 20(6):1981–1996, 2018. ISSN 1477-4054. doi: 10.1093/bib/bby063. URL `https://doi.org/10.1093/bib/bby063`.

[311] Joel B. Hagen. The origins of bioinformatics. *Nature Reviews Genetics*, 1(3):231–236, 2000. ISSN 1471-0064. doi: 10.1038/35042090. URL `https://doi.org/10.1038/35042090`.

[312] Arcady Mushegian. Grand challenges in bioinformatics and computational biology. *Frontiers in Genetics*, 2(60), 2011. ISSN 1664-8021. doi: 10.3389/fgene.2011.00060. URL `https://www.frontiersin.org/article/10.3389/fgene.2011.00060`.

[313] Jonathan C Fuller, Pierre Khoueiry, Holger Dinkel, Kristoffer Forslund, Alexandros Stamatakis, Joseph Barry, Aidan Budd, Theodoros G Soldatos, Katja Linssen, and Abdul Mateen Rajput. Biggest challenges in bioinformatics. *EMBO reports*, 14(4):302–304, 2013. ISSN 1469-221X. doi: 10.1038/embor.2013.34. URL `https://www.embopress.org/doi/abs/10.1038/embor.2013.34`.

[314] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016. ISSN 1476-4687. doi: 10.1038/533452a. URL `https://doi.org/10.1038/533452a`.

[315] Neha Kulkarni, Luca Alessandrì, Riccardo Panero, Maddalena Arigoni, Martina Olivero, Giulio Ferrero, Francesca Cordero, Marco Beccuti, and Raffaele A. Calogero. Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines. *BMC Bioinformatics*, 19(10):349, 2018.

ISSN 1471-2105. doi: 10.1186/s12859-018-2296-x. URL https://doi.org/10.1186/s12859-018-2296-x.

[316] Jason A. Papin, Feilim Mac Gabhann, Herbert M. Sauro, David Nickerson, and Anand Rampadarath. Improving reproducibility in computational biology research. *PLOS Computational Biology*, 16(5):e1007881, 2020. doi: 10.1371/journal.pcbi.1007881. URL https://doi.org/10.1371/journal.pcbi.1007881.

[317] Bartholomeus van den Bogert, Jos Boekhorst, Walter Pirovano, and Ali May. On the role of bioinformatics and data science in industrial microbiome applications. *Frontiers in Genetics*, 10(721), 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00721. URL https://www.frontiersin.org/article/10.3389/fgene.2019.00721.

[318] Teresa K Attwood, Sarah Blackford, Michelle D Brazas, Angela Davies, and Maria Victoria Schneider. A global perspective on evolving bioinformatics and data science training needs. *Briefings in Bioinformatics*, 20(2):398–404, 2017. ISSN 1477-4054. doi: 10.1093/bib/bbx100. URL https://doi.org/10.1093/bib/bbx100.

[319] Daniel J Rigden and Xosé M Fernández. The 2021 nucleic acids research database issue and the online molecular biology database collection. *Nucleic Acids Research*, 49(D1):D1–D9, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1216. URL https://doi.org/10.1093/nar/gkaa1216.

[320] Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, et al. Ensembl 2021. *Nucleic Acids Research*, 49(D1):D884–D891, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa942. URL https://doi.org/10.1093/nar/gkaa942.

[321] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv1189. URL https://doi.org/10.1093/nar/gkv1189.

[322] Jairo Navarro Gonzalez, Ann S Zweig, Matthew L Speir, Daniel Schmelter, Kate R Rosenbloom, et al. The ucsc genome browser database: 2021 update. *Nucleic Acids Research*, 49(D1):D1046–D1057, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1070. URL https://doi.org/10.1093/nar/gkaa1070.

[323] Carrie A Davis, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, et al. The encyclopedia of dna elements (encode): data portal update. *Nucleic Acids Research*, 46(D1):D794–D801, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx1081. URL https://doi.org/10.1093/nar/gkx1081.

[324] The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1100. URL https://doi.org/10.1093/nar/gkaa1100.

[325] Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, et al. Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, 49(D1):D437–D451, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1038. URL https://doi.org/10.1093/nar/gkaa1038.

[326] RNAcentral Consortium. Rnacentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Research*, 49(D1):D212–D220, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa921. URL https://doi.org/10.1093/nar/gkaa921.

[327] C. Backes, T. Fehlmann, F. Kern, T. Kehl, H. P. Lenhof, E. Meese, and A. Keller. mircarta: a central repository for collecting mirna candidates. *Nucleic Acids Res*, 46(D1):D160–d167, 2018. ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gkx851. URL https://doi.org/10.1093/nar/gkx851.

[328] Bastian Fromm, Diana Domanska, Eirik Høye, Vladimir Ovchinnikov, Wenjing Kang, et al. Mirgenedb 2.0: the metazoan microrna complement. *Nucleic Acids Research*, 48(D1):D132–D141, 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz885. URL https://doi.org/10.1093/nar/gkz885.

[329] Editorial: the 18th annual nucleic acids research web server issue 2020. *Nucleic Acids Research*, 48(W1):W1–W4, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa528. URL https://doi.org/10.1093/nar/gkaa528.

[330] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2. URL https://www.sciencedirect.com/science/article/pii/S0022283605803602.

[331] Yoon Byung-Jun. Hidden markov models and their applications in biological sequence analysis. *Current Genomics*, 10 (6):402–415, 2009. ISSN 1389-2029/1875-5488. doi: 10.2174/138920209789177575. URL http://www.eurekaselect.com/node/69904/article.

[332] Jacob Schreiber, Ritambhara Singh, Jeffrey Bilmes, and William Stafford Noble. A pitfall for machine learning methods aiming to predict across cell types. *Genome Biology*, 21(1):282,

2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02177-y. URL https://doi.org/10.1186/s13059-020-02177-y.

[333] Raquel Dias and Ali Torkamani. Artificial intelligence in clinical and genomic diagnostics. *Genome Medicine*, 11(1):70, 2019. ISSN 1756-994X. doi: 10.1186/s13073-019-0689-8. URL https://doi.org/10.1186/s13073-019-0689-8.

[334] Lieyang Chen, Anthony Cruz, Steven Ramsey, Callum J. Dickson, Jose S. Duca, Viktor Hornak, David R. Koes, and Tom Kurtzman. Hidden bias in the dud-e dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLOS ONE*, 14(8):e0220113, 2019. doi: 10.1371/journal.pone.0220113. URL https://doi.org/10.1371/journal.pone.0220113.

[335] Polina Mamoshina, Marina Volosnikova, Ivan V. Ozerov, Evgeny Putin, Ekaterina Skibina, Franco Cortese, and Alex Zhavoronkov. Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Frontiers in Genetics*, 9(242), 2018. ISSN 1664-8021. doi: 10.3389/fgene.2018.00242. URL https://www.frontiersin.org/article/10.3389/fgene.2018.00242.

[336] Aidan R O'Brien, Gaetan Burgio, and Denis C Bauer. Domain-specific introduction to machine learning terminology, pitfalls and opportunities in crispr-based gene editing. *Briefings in Bioinformatics*, 22(1):308–314, 2020. ISSN 1477-4054. doi: 10.1093/bib/bbz145. URL https://doi.org/10.1093/bib/bbz145.

[337] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16): 2078–2079, 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp352. URL https://doi.org/10.1093/bioinformatics/btp352.

[338] Aaron R. Quinlan and Ira M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6): 841–842, 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq033. URL https://doi.org/10.1093/bioinformatics/btq033.

[339] Knut Reinert, Temesgen Hailemariam Dadi, Marcel Ehrhardt, Hannes Hauswedell, Svenja Mehringer, et al. The seqan c++ template library for efficient sequence analysis: A resource for programmers. *Journal of Biotechnology*, 261:157–168, 2017. ISSN 0168-1656. doi: 10.1016/j.jbiotec.2017.07.017. URL https://www.sciencedirect.com/science/article/pii/S0168165617315420.

[340] Amstutz Peter, Crusoe Michael R., Tijanić Nebojša, Chapman Brad, Chilton John, et al. *Common Workflow Language, v1.0*. figshare, 2016. doi: 10.6084/m9.figshare.3115156.v2. URL https://figshare.com/articles/dataset/Common_Workflow_Language_draft_3/3115156.

[341] Paolo Di Tommaso, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, 2017. ISSN 1546-1696. doi: 10.1038/nbt.3820. URL https://doi.org/10.1038/nbt.3820.

[342] F M^lder, KP Jablonski, B Letcher, MB Hall, CH Tomkins-Tinch, et al. Sustainable data analysis with snakemake [version 2; peer review: 2 approved]. *F1000Research*, 10(33), 2021. doi: 10.12688/f1000research.29032.2. URL http://openr.es/1770.

[343] Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, 46(W1):W537–W544, 2018. ISSN 0305-1048. doi: 10.1093/nar/gky379. URL https://doi.org/10.1093/nar/gky379.

[344] Dirk Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux J.*, 2014 (239):Article 2, 2014. ISSN 1075-3583. URL https://www.linuxjournal.com/content/docker-lightweight-linux-containers-consistent-development-and-deployment.

[345] Gregory M. Kurtzer, Vanessa Sochat, and Michael W. Bauer. Singularity: Scientific containers for mobility of compute. *PLOS ONE*, 12(5):e0177459, 2017. doi: 10.1371/journal.pone.0177459. URL https://doi.org/10.1371/journal.pone.0177459.

[346] Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris, Johannes Köster, and Team The Bioconda. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15(7):475–476, 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0046-7. URL https://doi.org/10.1038/s41592-018-0046-7.

[347] Liang Chen, Liisa Heikkinen, Changliang Wang, Yang Yang, Huiyan Sun, and Garry Wong. Trends in the development of mirna bioinformatics tools. *Briefings in Bioinformatics*, 20(5):1836–1852, 2019. ISSN 1477-4054. doi: 10.1093/bib/bby054. URL https://doi.org/10.1093/bib/bby054.

[348] Tobias Fehlmann, Christina Backes, Mustafa Kahraman, Jan Haas, Nicole Ludwig, et al. Web-based ngs data analysis using mirmaster: a large-scale meta-analysis of human mirnas. *Nucleic Acids Research*, 45(15):8731–8744, 2017. ISSN 0305-1048.

doi: 10.1093/nar/gkx595. URL https://doi.org/10.1093/nar/gkx595.

[349] Tobias Fehlmann, Fabian Kern, Omar Laham, Christina Backes, Jeffrey Solomon, Pascal Hirsch, Carsten Volz, Rolf Müller, and Andreas Keller. mirmaster 2.0: multi-species non-coding rna sequencing analyses at scale. *Nucleic Acids Research*, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab268. URL https://doi.org/10.1093/nar/gkab268.

[350] G. P. Schmartz, F. Kern, T. Fehlmann, V. Wagner, B. Fromm, and A. Keller. Encyclopedia of tools for the analysis of mirna isoforms. *Brief Bioinform*, 2020. ISSN 1467-5463. doi: 10.1093/bib/bbaa346. URL https://doi.org/10.1093/bib/bbaa346.

[351] Shirley Tam, Ming-Sound Tsao, and John D. McPherson. Optimization of mirna-seq data preprocessing. *Briefings in Bioinformatics*, 16(6):950–963, 2015. ISSN 1467-5463. doi: 10.1093/bib/bbv019. URL https://doi.org/10.1093/bib/bbv019.

[352] Sylvain Pradervand, Johann Weber, Jérôme Thomas, Manuel Bueno, Pratyaksha Wirapati, Karine Lefort, G. Paolo Dotto, and Keith Harshman. Impact of normalization on mirna microarray expression profiling. *RNA*, 15(3):493–501, 2009. doi: 10.1261/rna.1295509. URL http://rnajournal.cshlp.org/content/15/3/493.abstract.

[353] Albert Pla, Xiangfu Zhong, and Simon Rayner. miraw: A deep learning-based approach to predict microrna targets by analyzing whole microrna transcripts. *PLOS Computational Biology*, 14(7):e1006185, 2018. doi: 10.1371/journal.pcbi.1006185. URL https://doi.org/10.1371/journal.pcbi.1006185.

[354] Tomas Tokar, Chiara Pastrello, Andrea E M Rossos, Mark Abovsky, Anne-Christin Hauschild, Mike Tsay, Richard Lu, and Igor Jurisica. mirdip 4.1—integrative database of human microrna target predictions. *Nucleic Acids Research*, 46(D1):D360–D370, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx1144. URL https://doi.org/10.1093/nar/gkx1144.

[355] Aurélien Quillet, Chadi Saad, Gaëtan Ferry, Youssef Anouar, Nicolas Vergne, Thierry Lecroq, and Christophe Dubessy. Improving bioinformatics prediction of microrna targets by ranks aggregation. *Frontiers in Genetics*, 10(1330), 2020. ISSN 1664-8021. doi: 10.3389/fgene.2019.01330. URL https://www.frontiersin.org/article/10.3389/fgene.2019.01330.

[356] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. ISSN 1546-1718. doi: 10.1038/75556. URL https://doi.org/10.1038/75556.

[357] The Gene Ontology Consortium. The gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1113. URL https://doi.org/10.1093/nar/gkaa1113.

[358] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.27. URL https://doi.org/10.1093/nar/28.1.27.

[359] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Mari Ishiguro-Watanabe, and Mao Tanabe. Kegg: integrating viruses and cellular organisms. *Nucleic Acids Research*, 49(D1):D545–D551, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa970. URL https://doi.org/10.1093/nar/gkaa970.

[360] Adrian Garcia-Moreno and Pedro Carmona-Saez. Computational methods and software tools for functional analysis of mirna data. *Biomolecules*, 10(9):1252, 2020. ISSN 2218-273X. doi: 10.3390/biom10091252. URL https://www.mdpi.com/2218-273X/10/9/1252.

[361] Vinny Negi and Stephen Y. Chan. Discerning functional hierarchies of micrornas in pulmonary hypertension. *JCI Insight*, 2 (5), 2017. ISSN 2379-3708. doi: 10.1172/jci.insight.91327. URL https://doi.org/10.1172/jci.insight.91327.

[362] Bing Liu, Jiuyong Li, and Murray J. Cairns. Identifying mirnas, targets and functions. *Briefings in Bioinformatics*, 2014. ISSN 1477-4054 (Electronic)\i1467-5463 (Linking). doi: 10.1093/bib/bbs075. URL https://doi.org/10.1093/bib/bbs075.

[363] Katarzyna Rolle, Monika Piwecka, Agnieszka Belter, Dariusz Wawrzyniak, Jaroslaw Jeleniewicz, Miroslawa Z. Barciszewska, and Jan Barciszewski. The sequence and structure determine the function of mature human mirnas. *PLOS ONE*, 11 (3):e0151246, 2016. doi: 10.1371/journal.pone.0151246. URL https://doi.org/10.1371/journal.pone.0151246.

[364] Andrew R. Bassett, Ghows Azzam, Lucy Wheatley, Charlotte Tibbit, Timothy Rajakumar, et al. Understanding functional mirna–target interactions in vivo by site-specific genome engineering. *Nature Communications*, 5(1):4640, 2014. ISSN 2041-1723. doi: 10.1038/ncomms5640. URL https://doi.org/10.1038/ncomms5640.

[365] Joana A. Vidigal and Andrea Ventura. The biological functions of mirnas: lessons from in vivo studies. *Trends in Cell Biology*, 25 (3):137–147, 2015. ISSN 0962-8924. doi: 10.1016/j.tcb.2014.11.004. URL https://doi.org/10.1016/j.tcb.2014.11.004.

[366] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting

genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005. doi: 10.1073/pnas.0506580102. URL https://www.pnas.org/content/pnas/102/43/15545.full.pdf.

[367] Qinghua Jiang, Yadong Wang, Yangyang Hao, Liran Juan, Mingxiang Teng, Xinjun Zhang, Meimei Li, Guohua Wang, and Yunlong Liu. mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic Acids Research*, 37(suppl_1):D98–D104, 2008. ISSN 0305-1048. doi: 10.1093/nar/gkn714. URL https://doi.org/10.1093/nar/gkn714.

[368] Andreas Keller, Christina Backes, and Hans-Peter Lenhof. Computation of significance scores of unweighted gene set enrichment analyses. *BMC Bioinformatics*, 8(1):290, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-290. URL https://doi.org/10.1186/1471-2105-8-290.

[369] Ming Lu, Bing Shi, Juan Wang, Qun Cao, and Qinghua Cui. Tam: A method for enrichment and depletion analysis of a microrna category in a list of micrornas. *BMC Bioinformatics*, 11 (1):419, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-419. URL https://doi.org/10.1186/1471-2105-11-419.

[370] Christina Backes, Qurratulain T. Khaleeq, Eckart Meese, and Andreas Keller. mieaa: microrna enrichment analysis and annotation. *Nucleic Acids Research*, 44(W1):W110–W116, 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw345. URL https://doi.org/10.1093/nar/gkw345.

[371] Florian Schmidt, Markus List, Engin Cukuroglu, Sebastian Köhler, Jonathan Göke, and Marcel H Schulz. An ontology-based method for assessing batch effect adjustment approaches in heterogeneous datasets. *Bioinformatics*, 34(17):i908–i916, 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty553. URL https://doi.org/10.1093/bioinformatics/bty553.

[372] Patrice Godard and Jonathan van Eyll. Pathway analysis from lists of micrornas: common pitfalls and alternative strategy. *Nucleic Acids Research*, 43(7):3490–3497, 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv249. URL https://doi.org/10.1093/nar/gkv249.

[373] Thomas Bleazard, Janine A Lamb, and Sam Griffiths-Jones. Bias in microrna functional enrichment analysis. *Bioinformatics*, 31(10):1592–1598, 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv023. URL https://doi.org/10.1093/bioinformatics/btv023.

[374] Keegan Korthauer, Patrick K. Kimes, Claire Duvallet, Alejandro Reyes, Ayshwarya Subramanian, Mingxiang Teng, Chinmay Shukla, Eric J. Alm, and Stephanie C. Hicks. A practical guide to methods controlling false discoveries in com-

putational biology. *Genome Biology*, 20(1):118, 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1716-1. URL https://doi.org/10.1186/s13059-019-1716-1.

[375] Julio R. Banga. Optimization in computational systems biology. *BMC Systems Biology*, 2(1):47, 2008. ISSN 1752-0509. doi: 10.1186/1752-0509-2-47. URL https://doi.org/10.1186/1752-0509-2-47.

[376] Pedro Larrañaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, et al. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112, 2006. ISSN 1467-5463. doi: 10.1093/bib/bbk007. URL https://doi.org/10.1093/bib/bbk007.

[377] Wolfgang Huber, Vincent J. Carey, Li Long, Seth Falcon, and Robert Gentleman. Graphs in molecular biology. *BMC Bioinformatics*, 8(6):S8, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-S6-S8. URL https://doi.org/10.1186/1471-2105-8-S6-S8.

[378] Georgios A. Pavlopoulos, Maria Secrier, Charalampos N. Moschopoulos, Theodoros G. Soldatos, Sophia Kossida, Jan Aerts, Reinhard Schneider, and Pantelis G. Bagos. Using graph theory to analyze biological networks. *BioData Mining*, 4(1): 10, 2011. ISSN 1756-0381. doi: 10.1186/1756-0381-4-10. URL https://doi.org/10.1186/1756-0381-4-10.

[379] Mikaela Koutrouli, Evangelos Karatzas, David Paez-Espino, and Georgios A. Pavlopoulos. A guide to conquer the biological network era using graph theory. *Frontiers in Bioengineering and Biotechnology*, 8(34), 2020. ISSN 2296-4185. doi: 10.3389/fbioe.2020.00034. URL https://www.frontiersin.org/article/10.3389/fbioe.2020.00034.

[380] Thomas G. Minchington, Sam Griffiths-Jones, and Nancy Papalopulu. Dynamical gene regulatory networks are tuned by transcriptional autoregulation with microrna feedback. *Scientific Reports*, 10(1):12960, 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-69791-5. URL https://doi.org/10.1038/s41598-020-69791-5.

[381] Nicolás Peláez and Richard W. Carthew. *Chapter nine - Biological Robustness and the Role of MicroRNAs: A Network Perspective*, volume 99, pages 237–255. Academic Press, 2012. ISBN 0070-2153. doi: 10.1016/B978-0-12-387038-4.00009-4. URL https://www.sciencedirect.com/science/article/pii/B9780123870384000094.

[382] Xing Chen, Jun-Yan Cheng, and Jun Yin. Predicting microrna-disease associations using bipartite local models and hubness-aware regression. *RNA Biology*, 15(9):1192–1205, 2018. ISSN 1547-6286. doi: 10.1080/15476286.2018.1517010. URL https://doi.org/10.1080/15476286.2018.1517010.

[383] Xin Lai, Olaf Wolkenhauer, and Julio Vera. Understanding microrna-mediated gene regulatory networks through mathematical modelling. *Nucleic Acids Research*, 44(13):6019–6035, 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw550. URL `https://doi.org/10.1093/nar/gkw550`.

[384] Weiyang Lou, Jingxing Liu, Bisha Ding, Danni Chen, Liang Xu, Jun Ding, Donghai Jiang, Lin Zhou, Shusen Zheng, and Weimin Fan. Identification of potential mirna–mrna regulatory network contributing to pathogenesis of hbv-related hcc. *Journal of Translational Medicine*, 17(1):7, 2019. ISSN 1479-5876. doi: 10.1186/s12967-018-1761-7. URL `https://doi.org/10.1186/s12967-018-1761-7`.

[385] Shuting Jin, Xiangxiang Zeng, Jiansong Fang, Jiawei Lin, Stephen Y. Chan, Serpil C. Erzurum, and Feixiong Cheng. A network-based approach to uncover microrna-mediated disease comorbidities and potential pathobiological implications. *npj Systems Biology and Applications*, 5(1):41, 2019. ISSN 2056-7189. doi: 10.1038/s41540-019-0115-2. URL `https://doi.org/10.1038/s41540-019-0115-2`.

[386] Marissa Sumathipala and Scott T. Weiss. Predicting mirna-based disease-disease relationships through network diffusion on multi-omics biological data. *Scientific Reports*, 10(1):8705, 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-65633-6. URL `https://doi.org/10.1038/s41598-020-65633-6`.

[387] Haneul Noh, Charny Park, Soojun Park, Young Seek Lee, Soo Young Cho, and Hyemyung Seo. Prediction of mirna-mrna associations in alzheimer's disease mice using network topology. *BMC Genomics*, 15(1):644, 2014. ISSN 1471-2164. doi: 10.1186/1471-2164-15-644. URL `https://doi.org/10.1186/1471-2164-15-644`.

[388] Simon Fischer, René Handrick, Armaz Aschrafi, and Kerstin Otte. Unveiling the principle of microrna-mediated redundancy in cellular pathway regulation. *RNA Biology*, 12(3):238–247, 2015. ISSN 1547-6286. doi: 10.1080/15476286.2015.1017238. URL `https://doi.org/10.1080/15476286.2015.1017238`.

[389] Zhaoxu Gao, Jun Li, Li Li, Yanzhi Yang, Jian Li, Chunxiang Fu, Danmeng Zhu, Hang He, Huaqing Cai, and Lei Li. Structural and functional analyses of hub micrornas in an integrated gene regulatory network of arabidopsis. *Genomics, Proteomics & Bioinformatics*, 2021. ISSN 1672-0229. doi: 10.1016/j.gpb.2020.02.004. URL `https://www.sciencedirect.com/science/article/pii/S1672022921000425`.

[390] Fabian Kern, Ernesto Aparicio-Puerta, Yongping Li, Tobias Fehlmann, Tim Kehl, et al. mirtargetlink 2.0—interactive mirna target gene and target pathway networks. *Nucleic Acids Research*,

2021. ISSN 0305-1048. doi: 10.1093/nar/gkab297. URL https://doi.org/10.1093/nar/gkab297.

[391] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17 (1):195, 2019. ISSN 1741-7015. doi: 10.1186/s12916-019-1426-2. URL https://doi.org/10.1186/s12916-019-1426-2.

[392] Ian A Scott, David Cook, Enrico W Coiera, and Brent Richards. Machine learning in clinical practice: prospects and pitfalls. *Medical Journal of Australia*, 211(5):203–205.e1, 2019. ISSN 0025-729X. doi: 10.5694/mja2.50294. URL https://onlinelibrary.wiley.com/doi/abs/10.5694/mja2.50294.

[393] Jonathan G. Richens, Ciarán M. Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, 11(1):3923, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17419-7. URL https://doi.org/10.1038/s41467-020-17419-7.

[394] Riccardo Bellazzi, Marco Masseroli, Shawn Murphy, Amnon Shabo, and Paolo Romano. Clinical bioinformatics: challenges and opportunities. *BMC Bioinformatics*, 13(14):S1, 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-S14-S1. URL https://doi.org/10.1186/1471-2105-13-S14-S1.

[395] Xiangdong Wang and Lance Liotta. Clinical bioinformatics: a new emerging science. *Journal of Clinical Bioinformatics*, 1(1): 1, 2011. ISSN 2043-9113. doi: 10.1186/2043-9113-1-1. URL https://doi.org/10.1186/2043-9113-1-1.

[396] John W. Belmont and Chad A. Shaw. Clinical bioinformatics: emergence of a new laboratory discipline. *Expert Review of Molecular Diagnostics*, 16(11):1139–1141, 2016. ISSN 1473-7159. doi: 10.1080/14737159.2016.1246184. URL https://doi.org/10.1080/14737159.2016.1246184.

[397] Guy Haskin Fernald, Emidio Capriotti, Roxana Daneshjou, Konrad J. Karczewski, and Russ B. Altman. Bioinformatics challenges for personalized medicine. *Bioinformatics*, 27(13):1741–1748, 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr295. URL https://doi.org/10.1093/bioinformatics/btr295.

[398] Xiaoqin Liu, Xin Luo, Chunyang Jiang, and Hui Zhao. Difficulties and challenges in the development of precision medicine. *Clinical Genetics*, 95(5):569–574, 2019. ISSN 0009-9163. doi: 10.1111/cge.13511. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cge.13511.

[399] Ju Han Kim. Bioinformatics and genomic medicine. *Genetics in Medicine*, 4(6):62–65, 2002. ISSN 1530-0366. doi: 10.1097/00125817-200211001-00013. URL https://doi.org/10.1097/00125817-200211001-00013.

[400] Katia Grillone, Caterina Riillo, Francesca Scionti, Roberta Rocca, Giuseppe Tradigo, Pietro Hiram Guzzi, Stefano Alcaro, Maria Teresa Di Martino, Pierosandro Tagliaferri, and Pierfrancesco Tassone. Non-coding rnas in cancer: platforms and strategies for investigating the genomic "dark matter". *Journal of Experimental & Clinical Cancer Research*, 39(1):117, 2020. ISSN 1756-9966. doi: 10.1186/s13046-020-01622-x. URL `https://doi.org/10.1186/s13046-020-01622-x`.

[401] Frank J. Slack and Arul M. Chinnaiyan. The role of non-coding rnas in oncology. *Cell*, 179(5):1033–1055, 2019. ISSN 0092-8674. doi: 10.1016/j.cell.2019.10.017. URL `https://doi.org/10.1016/j.cell.2019.10.017`.

[402] Caterina Vicentini, Francesca Galuppini, Vincenzo Corbo, and Matteo Fassan. Current role of non-coding rnas in the clinical setting. *Non-coding RNA Research*, 4 (3):82–85, 2019. ISSN 2468-0540. doi: 10.1016/j.ncrna.2019.09.001. URL `https://www.sciencedirect.com/science/article/pii/S2468054019300265`.

[403] 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 Biol*, 16(8):1034–1043, 2019. ISSN 1547-6286 (Print) 1547-6286. doi: 10.1080/15476286.2019.1612217. URL `https://doi.org/10.1080/15476286.2019.1612217`.

[404] M. Kahraman, A. Röske, T. Laufer, T. Fehlmann, C. Backes, et al. Microrna in diagnosis and therapy monitoring of early-stage triple-negative breast cancer. *Sci Rep*, 8(1):11584, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-29917-2. URL `https://doi.org/10.1038/s41598-018-29917-2`.

[405] A. Keller, N. Ludwig, T. Fehlmann, M. Kahraman, C. Backes, et al. Low mir-150-5p and mir-320b expression predicts reduced survival of copd patients. *Cells*, 8(10), 2019. ISSN 2073-4409. doi: 10.3390/cells8101162. URL `https://doi.org/10.3390/cells8101162`.

[406] A. Keller, T. Fehlmann, C. Backes, F. Kern, R. Gislefoss, H. Langseth, T. B. Rounge, N. Ludwig, and E. Meese. Competitive learning suggests circulating mirna profiles for cancers decades prior to diagnosis. *RNA Biol*, 17 (10):1416–1426, 2020. ISSN 1547-6286 (Print) 1547-6286. doi: 10.1080/15476286.2020.1771945. URL `https://doi.org/10.1080/15476286.2020.1771945`.

[407] V. Palmieri, C. Backes, N. Ludwig, T. Fehlmann, F. Kern, E. Meese, and A. Keller. Imota: an interactive multi-omics tissue atlas for the analysis of human mirna-target interactions. *Nucleic Acids Res*, 46(D1):D770–d775, 2018. ISSN

0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gkx701. URL https://doi.org/10.1093/nar/gkx701.

[408] J. Solomon, F. Kern, T. Fehlmann, E. Meese, and A. Keller. Humir: Web services, tools and databases for exploring human microrna data. *Biomolecules*, 10(11), 2020. ISSN 2218-273x. doi: 10.3390/biom10111576. URL https://doi.org/10.3390/biom10111576.

[409] Tobias Fehlmann, Fabian Kern, Pascal Hirsch, Robin Steinhaus, Dominik Seelow, and Andreas Keller. Aviator: a web service for monitoring the availability of web services. *Nucleic Acids Research*, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab396. URL https://doi.org/10.1093/nar/gkab396.

[410] F. Schmidt, F. Kern, P. Ebert, N. Baumgarten, and M. H. Schulz. Tepic 2-an extended framework for transcription factor binding prediction and integrative epigenomic analysis. *Bioinformatics*, 35(9):1608–1609, 2019. ISSN 1367-4803 (Print) 1367-4803. doi: 10.1093/bioinformatics/bty856. URL https://doi.org/10.1093/bioinformatics/bty856.

[411] N. Grammes, D. Millenaar, T. Fehlmann, F. Kern, M. Böhm, F. Mahfoud, and A. Keller. Research output and international cooperation among countries during the covid-19 pandemic: Scientometric analysis. *J Med Internet Res*, 22(12):e24514, 2020. ISSN 1439-4456 (Print) 1438-8871. doi: 10.2196/24514. URL https://doi.org/10.2196/24514.

[412] Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03583-3. URL https://doi.org/10.1038/s41586-021-03583-3.

[413] Andrew C. Yang, Fabian Kern, Patricia M. Losada, Maayan R. Agam, Christina A. Maat, et al. Dysregulation of brain and choroid plexus cell types in severe covid-19. *Nature*, 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03710-0. URL https://doi.org/10.1038/s41586-021-03710-0.

[414] Federico Fuchs Wightman, Luciana E. Giono, Juan Pablo Fededa, and Manuel de la Mata. Target rnas strike back on micrornas. *Frontiers in Genetics*, 9(435), 2018. ISSN 1664-8021. doi: 10.3389/fgene.2018.00435. URL https://www.frontiersin.org/article/10.3389/fgene.2018.00435.

[415] Thomas B. Hansen, Trine I. Jensen, Bettina H. Clausen, Jesper B. Bramsen, Bente Finsen, Christian K. Damgaard, and Jørgen Kjems. Natural rna circles function as efficient microrna sponges. *Nature*, 495(7441):384–388, 2013. ISSN 1476-4687.

doi: 10.1038/nature11993. URL https://doi.org/10.1038/nature11993.

[416] Markus List, Azim Dehghani Amirabad, Dennis Kostka, and Marcel H Schulz. Large-scale inference of competing endogenous rna networks with sparse partial correlation. *Bioinformatics*, 35(14):i596–i604, 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz314. URL https://doi.org/10.1093/bioinformatics/btz314.

[417] Carrie Wright, Anandita Rajpurohit, Emily E. Burke, Courtney Williams, Leonardo Collado-Torres, et al. Comprehensive assessment of multiple biases in small rna sequencing reveals significant differences in the performance of widely used methods. *BMC Genomics*, 20(1):513, 2019. ISSN 1471-2164. doi: 10.1186/s12864-019-5870-3. URL https://doi.org/10.1186/s12864-019-5870-3.

[418] Jeanette Baran-Gale, C. Lisa Kurtz, Michael R. Erdos, Christina Sison, Alice Young, Emily E. Fannin, Peter S. Chines, and Praveen Sethupathy. Addressing bias in small rna library preparation for sequencing: A new protocol recovers micrornas that evade capture by current methods. *Frontiers in Genetics*, 6(352), 2015. ISSN 1664-8021. doi: 10.3389/fgene.2015.00352. URL https://www.frontiersin.org/article/10.3389/fgene.2015.00352.

[419] Anne Baroin-Tourancheau, Yan Jaszczyszyn, Xavier Benigni, and Laurence Amar. Evaluating and correcting inherent bias of microrna expression in illumina sequencing analysis. *Frontiers in molecular biosciences*, 6:17–17, 2019. ISSN 2296-889X. doi: 10.3389/fmolb.2019.00017. URL https://pubmed.ncbi.nlm.nih.gov/31069233https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6491513/.

[420] Hansruedi Mathys, Jose Davila-Velderrain, Zhuyu Peng, Fan Gao, Shahin Mohammadi, et al. Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, 570(7761):332–337, 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1195-2. URL https://doi.org/10.1038/s41586-019-1195-2.

[421] Alexandra Grubman, Gabriel Chew, John F. Ouyang, Guizhi Sun, Xin Yi Choo, et al. A single-cell atlas of entorhinal cortex from individuals with alzheimer's disease reveals cell-type-specific gene expression regulation. *Nature Neuroscience*, 22(12):2087–2097, 2019. ISSN 1546-1726. doi: 10.1038/s41593-019-0539-4. URL https://doi.org/10.1038/s41593-019-0539-4.

[422] Emily Stephenson, Gary Reynolds, Rachel A. Botting, Fernando J. Calero-Nieto, Michael D. Morgan, et al. Single-cell multi-omics analysis of the immune response in covid-19. *Nature Medicine*, 2021. ISSN 1546-170X. doi: 10.1038/s41591-021-

01329-2. URL https://doi.org/10.1038/s41591-021-01329-2.

[423] Xianwen Ren, Wen Wen, Xiaoying Fan, Wenhong Hou, Bin Su, et al. Covid-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell*, 184(7): 1895–1913.e19, 2021. ISSN 0092-8674. doi: 10.1016/j.cell.2021.01.053. URL https://www.sciencedirect.com/science/article/pii/S0092867421001483.

[424] Alina Isakova, Norma Neff, and Stephen R. Quake. Single cell profiling of total rna using smart-seq-total. *bioRxiv*, page 2020.06.02.131060, 2020. doi: 10.1101/2020.06.02.131060. URL https://www.biorxiv.org/content/biorxiv/early/2020/06/03/2020.06.02.131060.full.pdf.

[425] Yongjin Park, Liang He, Jose Davila-Velderrain, Lei Hou, Shahin Mohammadi, Hansruedi Mathys, Zhuyu Peng, David Bennett, Li-Huei Tsai, and Manolis Kellis. Single-cell deconvolution of 3,000 post-mortem brain samples for eqtl and gwas dissection in mental disorders. *bioRxiv*, page 2021.01.21.426000, 2021. doi: 10.1101/2021.01.21.426000. URL https://www.biorxiv.org/content/biorxiv/early/2021/01/21/2021.01.21.426000.full.pdf.

[426] S. Smajić, C. A. Prada-Medina, Z. Landoulsi, C. Dietrich, J. Jarazo, et al. Single-cell sequencing of the human midbrain reveals glial activation and a neuronal state specific to parkinson's disease. *medRxiv*, page 2020.09.28.20202812, 2020. doi: 10.1101/2020.09.28.20202812. URL https://www.medrxiv.org/content/medrxiv/early/2020/09/30/2020.09.28.20202812.full.pdf.

[427] Devika Agarwal, Cynthia Sandor, Viola Volpato, Tara M. Caffrey, Jimena Monzón-Sandoval, Rory Bowden, Javier Alegre-Abarrategui, Richard Wade-Martins, and Caleb Webber. A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders. *Nature Communications*, 11(1):4183, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17876-0. URL https://doi.org/10.1038/s41467-020-17876-0.

[428] Elisa Navarro, Evan Udine, Katia de Paiva Lopes, Madison Parks, Giulietta Riboldi, et al. Discordant transcriptional signatures of mitochondrial genes in parkinson's disease human myeloid cells. *bioRxiv*, page 2020.07.20.212407, 2020. doi: 10.1101/2020.07.20.212407. URL https://www.biorxiv.org/content/biorxiv/early/2020/07/22/2020.07.20.212407.full.pdf.

[429] Andrew C. Yang, Ryan T. Vest, Fabian Kern, Davis P. Lee, Christina A. Maat, et al. A human brain vascular atlas reveals diverse cell mediators of alzheimer's disease

risk. *bioRxiv*, page 2021.04.26.441262, 2021. doi: 10.1101/2021.04.26.441262. URL https://www.biorxiv.org/content/biorxiv/early/2021/04/27/2021.04.26.441262.full.pdf.

[430] Nayi Wang, Ji Zheng, Zhuo Chen, Yang Liu, Burak Dura, et al. Single-cell microrna-mrna co-sequencing reveals non-genetic heterogeneity and mechanisms of microrna regulation. *Nature Communications*, 10(1):95, 2019. ISSN 2041-1723. doi: 10.1038/s41467-018-07981-6. URL https://doi.org/10.1038/s41467-018-07981-6.

[431] Morten Muhlig Nielsen and Jakob Skou Pedersen. mirna activity inferred from single cell mrna expression. *Scientific Reports*, 11(1):9170, 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-88480-5. URL https://doi.org/10.1038/s41598-021-88480-5.

[432] Jakob Lewin Rukov, Roni Wilentzik, Ishai Jaffe, Jeppe Vinther, and Noam Shomron. Pharmaco-mir: linking micrornas and drug effects. *Briefings in Bioinformatics*, 15(4):648–659, 2013. ISSN 1467-5463. doi: 10.1093/bib/bbs082. URL https://doi.org/10.1093/bib/bbs082.

[433] Igor Koturbash, William H Tolleson, Lei Guo, Dianke Yu, Si Chen, Huixiao Hong, William Mattes, and Baitang Ning. micrornas as pharmacogenomic biomarkers for drug efficacy and drug safety assessment. *Biomarkers in Medicine*, 9 (11):1153–1176, 2015. doi: 10.2217/bmm.15.89. URL https://www.futuremedicine.com/doi/abs/10.2217/bmm.15.89.

[434] Jakob Lewin Rukov and Noam Shomron. Microrna pharmacogenomics: Post-transcriptional regulation of drug response. *Trends in Molecular Medicine*, 17(8):412–423, 2011. ISSN 1471-4914. doi: 10.1016/j.molmed.2011.04.003. URL https://www.sciencedirect.com/science/article/pii/S1471491411000773.

[435] Ali Akbar Jamali, Anthony Kusalik, and Fang-Xiang Wu. Mdipa: a microrna–drug interaction prediction approach based on non-negative matrix factorization. *Bioinformatics*, 36(20): 5061–5067, 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa577. URL https://doi.org/10.1093/bioinformatics/btaa577.

Acknowledgement

Without the gracious support of my love Kirsten, my parents Anja and Markus and my close friends Felix and Stefan this work would simply not have been possible.

I should express deepest gratitude to my supervisor Andreas Keller for giving me the opportunity to join his group at Saarland University and for his endless support during my time both as a master and PhD student. I greatly enjoyed the freedom I was given in making my own decisions.
Never ending appreciation to Eckart Meese who kept us all running, no matter how difficult things were seem to get.
Many thanks are also directed to the committee members for spending their precious time to review my thesis.

I would not have been at this point in my life without the infinite knowledge and wisdom of Christina Backes, who not only taught me how to be a good scientist but always showed me how to find a way out.

Below is a non-exhaustive, lexicographically sorted list of people who generously supported me at some point during the last eight years:

- Sebastian Alberternst
- Ernesto Aparicio-Puerta
- Christina Backes
- Nina Baumgarten
- Fatemeh Behjati Ardakani
- Nanna Dahlem
- Kathleen Dickey
- Dilip Durai
- Cornelia Erhardt
- Thomas Erhardt
- Alexander Fauß

- Tobias Fehlmann
- Anna Feldmann
- Jonas Fischer
- Matthias Flotho
- Bastian Fromm
- Valentina Galata
- Nico Gerstner
- Nadja L. Grammes
- Moritz Graus
- Malte Groß
- Pedro Guimaraes
- Immanuel Haffner
- Oliver Hahn
- Martin Hart
- Pascal Hirsch
- Markus Hollander
- Tal Iram
- Max Jakob
- Karin Jostock
- Anne Jungfleisch
- Mustafa Kahraman
- Tim Kehl
- Andreas Keller
- Anja Kern
- Markus Kern
- Lukas Koch
- Matthis Kruse
- Oliver Küchler
- Andrea Kupitz
- Lena Krammes
- Cedric Laczny
- Hans-Peter Lenhof

- Kerstin Lenhof
- Sabine Lessel
- Markus List
- Alexander Löffler
- Patricia M. Losada
- Nicole Ludwig
- Eckart Meese
- Björn Mohr
- Fabian Müller
- Jonas Müller
- Joris Nix
- Robert Palovics
- Luciano Pica
- Nicolas Schäfer
- Thorsten Schamper
- Michael Scherer
- Felix Scherzinger
- Georges P. Schmartz
- Florian Schmidt
- Nils Schmitt
- Lara Schneider
- Verena Schorr
- Johannes Schramm
- Marcel Schulz
- Walter Schulz-Schaeffer
- Carlo Seelinger
- Dominik Seelow
- Rebecca Serra Mari
- Shiva Sivajaran
- Jeffrey Solomon
- Fabian Spaniol
- Nora Speicher

- Daniel Stöckel
- Artur Suleymanov
- Lukas Tost
- Viktoria Wagner
- Alexander Wahls
- Kirsten Weber
- Thorsten Will
- Stefan Wingerter
- Tony Wyss-Coray
- Andrew C. Yang

Curriculum Vitae