TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Genomorientierte Bioinformatik

Computational modeling of miRNA-mediated gene regulation in consideration of miRNP binding information from AGO-bound CLIP-Seq data analysis

Daniel Christian Ellwanger

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. H. R. Fries
Prüfer der Dissertation:	1. Univ.-Prof. Dr. H.-W. Mewes
	2. Univ.-Prof. Dr. Dr. F. J. Theis
	3. Univ.-Prof. Dr. L. Kaderali Technische Universität Dresden

Die Dissertation wurde am 17. März 2015 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 9. Juni 2015 angenommen.

Acknowledgements

This dissertation would not have been succeeded without the support of several great people that accompanied me during the last years.

First of all, I would like to express my appreciation and thanks to my doctoral adviser Prof. Dr. Hans-Werner Mewes. For many years, you have been a tremendous mentor for me. I am deeply grateful that you encouraged my research and promoted my professional growth not only as a scientist, but also as a lecturer. I would also like to thank my thesis committee members, Prof. Dr. Juliane Winkelmann, Prof. Dr. Dmitrij Frishman, and Dr. Volker Stümpflen, for their valuable comments and suggestions on my research.

Thanks to all members of the Institute of Bioinformatics and Systems Biology at the Helmholtz Zentrum München for many successful collaborations. I am especially grateful for the various highly imaginative discussion with my longtime room mate and good friend Jörn Leonhardt. In the same breath, I would also like to thank Dr. Florian Büttner and Matthias Arnold. Our technical debates were always highly constructive and insightful. Further, I would like to thank Barbara Berschneider and Dr. Dr. Melanie Königshoff as well as Dr. Eva Schulte and Prof. Dr. Juliane Winkelmann for our great cross-functional collaboration elucidating highly interesting biomedical questions. In this context, I would like to express my thanks to Prof. Dr. Oliver Eickelberg and Prof. Dr. Hans-Werner Mewes for giving me the opportunity to visit the research group of Prof. Dr. Naftali Kaminski at the University of Pittsburgh. Thank you Naftali for the paramount professional and personal experience during my stay. Especially, I would like to thank Dr. Jada Milosevic, Dr. Kusum Pandit, and Prof. Dr. Takis Benos for constructive and fruitful discussions.

Thanks to my colleagues at the chair of Genome-oriented Bioinformatics for the great working atmosphere. Here, I was involved in several teaching activities at which I met many talented students. Most notably, it has been a honor for me to work with Gregor Sturm and André Seitz as well as to supervise Alice Meier and Göksel Kaya.

A special thanks to my parents and my sister for all your support, not only during the last years. Above all I would like to express my deep appreciation to my beloved wife Julia. Words cannot express adequately how grateful I am for all of the sacrifices that you made on my behalf. Your constant support and encouragement during all of the ups and downs of my research was of inestimable value. $\star$

"It is now clear an extensive miRNA world was flying almost unseen by our genetic radar. As much as geneticists like to think that nothing can escape genetic analysis, the miRNA genes are so small that they almost escaped our notice. [...] The flowering of the diverse and numerous miRNA genes in animals and plants may turn out to mediate much of the gene regulation that generates cell diversity and developmental patterning, as well as the gene regulation underlying other recent inventions in animals such as synaptic signaling and its modulation."

Gary Ruvkun, Bruce Witghman and Ilho Ha. *Cell*, 2004.

"Over the past few years, remarkable progress has been made in our understanding of miRNA biogenesis and function; however, the mechanisms that miRNAs use to regulate gene expression remain unclear and several controversies surround the topic."

Eric Huntzinger, and Elisa Izaurralde. *Nat Rev Genet*, 2011.

Abstract

As our knowledge of the eukaryotic genome structure and organization has evolved, a new family of small non-coding regulatory RNAs, called microRNAs (miRNAs), has emerged. 20 years after their first discovery, our comprehension of these molecules and their interactions is still limited yet. Bound to a ribonucleoprotein complex (miRNP), miRNAs were suggested to pair with messenger RNAs (mRNAs) inducing target decay or translational repression. By forming a post-transcriptional regulation layer, they allow the adjustment or amplification of target gene expression regulation conducted by transcription factors. The miRNA targetome is extensive and constitutes highly wired networks of regulatory interactions. At this, several highly-connected miRNAs were identified playing key roles in the control of crucial cellular processes. In this context, the perturbation of miRNA-mediated regulation was associated to the outcome of malignant diseases, such as cancer. To date, there are more than 1 800 miRNA genes and 2 500 mature transcripts known in human − for the major fraction, the function remains to be revealed.

Recent technologies contributed significant progress in the field of miRNA research. By Argonaute (AGO):RNA cross-linking (CL), immunoprecipitation (IP) and subsequent high-throughput sequencing (CLIP-Seq) of bound RNAs, it became feasible, for the first time, to determine miRNP target sites of a whole transcriptome with high specificity. Until then, target information was sparse and incomplete. Thus, the emergence of the AGO-bound CLIP-Seq protocols afforded a wealth of data for the analysis of miRNA-mediated regulation. This doctoral thesis aimed to quantitatively elucidate basic miRNP:target interaction paradigms, to examine how these are impacted by genetic variance, and finally, to develop a novel computational approach to qualitatively model global miRNA-mediated regulation by means of novel information extracted from available AGO-bound CLIP-Seq libraries.

First, the most important feature for target prediction, the pairing between the target sequence and the miRNA 5'-end, the miRNA seed, was revisited. A set of canonical seed types was identified by a sequence pattern mining strategy in AGO-bound CLIP-Seq data. Quantitative seed type analysis confirmed the proposed specificity of long seeds, but also revealed that the majority of *bona fide* sites are formed by less specific and non-conserved seeds holding a minor impact on target expression. Their potential important role in the

miRNA regulome was discussed. An evaluation of current computational target prediction models showed that the majority of functional sites remain uncovered.

Next, sequence-, structure-, and homology-based attributes of miRNP target sites were extracted and analyzed. Here, known features were confirmed. Also a novel characteristic was detected and its potential relevance for target site determination was discussed. A generic machine learning approach was implemented which was shown to improve the precision of existing methods. Further, the novel information on miRNP target sites was applied to a biological case study. Here, a miRNA regulation was identified and experimentally verified which may contribute to the pathogenic phenotype of idiopathic pulmonary fibrosis.

Human genetic variation has been associated to complex traits and diseases. Here, the genomic diversity arises from 1% of variation, mostly induced by single nucleotide polymorphisms (SNPs). Of these, reported SNPs affecting the miRNA regulation pathway are rare. By utilizing the AGO-bound CLIP-Seq library, the question was examined whether genetic variance interferes with miRNP bindig site features. A set of trait-associated index SNPs and proximal SNPs in linkage disequilibrium were computed using data from genome-wide association studies. The analysis of their localization revealed an enrichment in 3'-untranslated regions (3'-UTRs) of protein-coding genes − the predominant region embedding miRNA binding sites. Here, several potential mechanisms were investigated affecting miRNA-mediated regulation. In the end, 53 *cis*-miR-SNPs were found altering the canonical miRNA seed pairing, the 3'-UTR folding and/or the 3'-UTR splicing. It was observed that *cis*-miR-SNPs induce an allelic expression imbalance and induce a noticeable target expression variation.

Finally, the computational modeling of globally miRNA-mediated regulation was addressed. Understanding how regulatory networks globally coordinate the response of a cell to changing conditions, such as perturbations by shifting environments, is an elementary challenge in systems biology which has yet to be met. Genome-wide gene expression measurements are high dimensional as these are reflecting the condition-specific interplay of thousands of cellular components. The integration of prior biological knowledge, such as AGO-bound CLIP-Seq libraries, into the modeling process of systems-wide gene regulation enables the large-scale interpretation of gene expression signals in the context of known regulatory relations. Within this thesis a novel approach called COGERE was developed. It denotes a method for the inference of condition-specific gene regulatory networks in

human and mouse. A framework to integrate existing knowledge of regulatory interactions from multiple sources to a comprehensive model of prior information is presented. Further, an algorithm was developed for the inference of condition-specific regulation by evaluating the mutual dependency between regulator (transcription factor or miRNA) and target gene expression using prior information. This dependency is scored by the non-parametric, nonlinear correlation coefficient η^2 (eta squared) that is derived by a two-way analysis of variance. In this thesis, it is shown that COGERE significantly outperforms alternative methods in scoring prior information as well as in predicting condition-specific gene regulatory networks on simulated data sets. Furthermore, by inferring the cancer gene regulatory network, the value of COGERE to promote hypothesis-driven clinical research is demonstrated.

Zusammenfassung

Über die letzten Jahrzehnte expandierte unser Wissen über die Topologie des eukaryotischen Genoms in einem beachtlichen Ausmaß. Hierbei sind viele neue Elemente beschrieben worden – unter anderem auch eine Familie kleiner nicht-kodierender RNAs, den sogenannten microRNAs (miRNAs). Noch heute, 20 Jahre nach ihrer ersten Entdeckung, sind unsere Erkenntnisse über diese Moleküle und deren Interaktionen limitiert. Integriert in einem Ribonukleoproteinkomplex (miRNP) binden miRNAs komplementär an Boten-RNA und erzwingen so deren Abbau oder unterbinden die Translation. Somit bilden miRNAs eine post-transkriptionelle Regulationsebene, welche eine Adjustierung oder Amplifikation der von Transkriptionsfatkoren gesteuerten Genexpression ermöglicht. Die Gesamtheit der miRNA Zielgene (Regulom) ist beachtlich groß. Es ist somit nicht erstaunlich, dass die Menge an Interaktionen ein dichtes post-transkriptionelles genregulatorisches Netzwerk impliziert. Hierbei wurden bereits verschiedenen hoch-verbundenen miRNAs entscheidende Rollen in der Kontrolle von essenziellen Zellprozessen zugewiesen. In diesem Zusammenhang wurde auch die Perturbation der miRNA-mediierten Genregulation mit schweren Erkrankungen, wie zum Beispiel Krebs, in Verbindung gebracht. Nach heutigem Stand sind mehr als 1 800 miRNA Gene und 2 500 reife Genprodukte im menschlichen Genom bekannt – für die Mehrzahl ist die Funktion noch gänzlich unbekannt.

Moderne Technologien haben einen signifikanten Anteil zum aktuellen Fortschritt in der miRNA Forschung beigetragen. Ein kürzlich veröffentlichtes Protokoll erzwingt kovalente Bindungen zwischen dem miRNP und der gebundenen Boten-RNA (cross-linking Verfahren, CL). Anschließend wird das RNA-bindende Protein, in diesem Fall Argonaute (AGO), anhand von Immunopräzipitation (IP) isoliert. Das anschließende Sequenzieren der gebundenen RNAs (Seq) ermöglicht, zum ersten Mal, eine hoch-spezfische, transkriptomweite Identifikation von miRNP Bindestellen. Bis zu diesem Zeitpunkt waren die Informationen über miRNA Zielgene lückenhaft und ungenau. Entsprechend enthalten AGO CLIP-Seq Daten neuen, hoch-relevante Informationen. Diese Doktorarbeit hatte das Ziel anhand der quantitativen Exploration von verfügbaren AGO CLIP-Seq Bibliotheken grundlegende miRNP:Zielgen Interaktionsparadigmen aufzuklären, zu erörtern wie diese durch genetische Variation beeinflusst werden und letztlich einen Ansatz zu entwickeln,

um globale qualitative Modelle zellulärer miRNA-mediierter Regulation berechnen zu können.

Zuerst wurde die wichtigste Charakteristik für die Vorhersage von miRNA Zielgenen, die Paarung zwischen der Boten-RNA und der 5'-terminale miRNA Seedsequenz, erneut untersucht. Hierbei wurde eine Menge an kanonischen Seedtypen anhand einer statistischen Mustersuche in der AGO CLIP-Seq Sequenzbibliothek identifiziert. Die quantitative Analyse der Seedtypen bestätigte die bisher angenommene Spezifität langer Seeds, jedoch zeigte auch, dass die Mehrheit der *bona fide* Bindestellen durch weniger spezifische und schwach konservierte Typen gestellt wird. Diese weisen zudem eine geringere regulatorische Effektivität auf. Ihre dennoch potentiell sehr wichtige Rolle im miRNA-Regulom wurde in diesen Rahmen diskutiert. Eine Evaluierung von gängigen Vorhersagealgorithmen zeigte, dass die Majorität der funktionellen Bindestellen von diesen Methoden nicht aufgefunden wird.

Als Nächstes wurden sequenz-, struktur-, und homologie-basierte Attribute von miRNP Bindestellen extrahiert und ausgewertet. Hierbei wurden bestehende Erkenntnisse bestätigt, sowie eine neue Charakteristik entdeckt und ihre Relevanz für die Bestimmung von Bindestellen diskutiert. Anhand von Maschinellen Lernen wurde ein generischer Ansatz entwickelt, um genomweit miRNP Bindestellen klassifizieren zu können. Es wurde gezeigt, dass dieser Ansatz die Präzision bestehender Vorhersagemethoden verbessern kann. Schließlich wurden die neuen Erkenntnisse im Rahmen einer biologischen Fallstudie angewandt. Hierbei wurde eine miRNA Regulation identifiziert und experimentell verifiziert, welche vermutlich bei der Ätiologie von idiopathischer pulmonaler Fibrose eine Rolle spielt.

Die humane genetische Variation wurde mit komplexen phänotypischen Merkmalen und Krankheiten assoziiert. Diese wird meistens durch Einzelnukleotidpolymorphismen (Single Nucleotide Polymorphisms, SNPs) erzeugt. Bis heute wurden diese kaum im Zusammenhang mit miRNA Regulation beschrieben. Mehrere Indizien deuten jedoch auch auf eine mögliche Modifikation der post-transkriptionellen Regulation von betroffenen Genen hin. Demnach wurde unter Verwendung der AGO CLIP-Seq Daten die Frage erörtert, ob genetische Varianz mit Eigenschaften der miRNP Bindestellen interferiert. Eine Menge an Index-SNPs und proximalen SNPs in Kopplungsungleichgewicht wurde mit Hilfe von Daten aus genomweiten Assoziationsstudien erstellt. Die Betrachtung der genomischen Positionen dieser SNPs brachte hervor, dass eine Anreicherung in den

3‘-untranslatierten Regionen (3’-UTRs) von protein-kodierenden Genen besteht. Diese Segemente sind prominent für das enkodieren von miRNA Bindestellen. Verschiedene Mechanismen wurden eruiert, welche eine mögliche Störung der miRNA Regulation erwirken könnten. Schließlich wurden 53 *cis*-miR-SNPs gefunden, welche kanonische miRNA Seed-komplementäre Stellen verändern, die lokale 3‘-UTR Faltung modifizieren und/oder alternativ-gespleißte 3’-UTR Transkripte erzeugen. Unter Verwendung von Expressionsanalysen wurde gezeigt, dass diese *cis*-miR-SNPs eine allel-abhängige Transkriptkonzentration, sowie eine erhöhte Konzentrationsvarianz aufweisen.

Im letzten Teil dieser Arbeit wurde die Problematik der Modellierung von globalen miRNA-mediierten Regulationsnetzwerken adressiert. Das Verständnis wie diese Netzwerke die Antwort einer Zelle auf veränderte (Umwelt-)Bedingungen koordinieren ist eine elementare, sowie hochaktuelle Fragestellung der Systembiologie. Genomweite Genexpressionsmessungen sind hochdimensional, da diese das konditionsspezifische Zusammenspiel von Tausenden zellulären Komponenten wiedergeben. Die Integration von Vorwissen, wie zum Beispiel Informationen aus den AGO CLIP-Seq Bibliotheken, in die Modellierung von systemweiter Genregulation, ermöglicht die Interpretation von Expressionssignalen im Kontext bekannter regulatorischer Relationen in großem Umfang. Im Rahmen dieser Doktorarbeit wurde der neue Ansatz COGERE entwickelt. Dies ist eine Methode zur Inferenz konditionsspezifischer genetischer Netzwerke in Mensch und Maus. Zunächst wird ein Framework beschrieben, welches die Integration von aktuellem Wissen aus einer Vielzahl an semantisch unterschiedlichen Ressourcen zu einem einheitlichen Modell ermöglicht (Prior-Modell). Desweiteren wird ein Ansatz präsentiert, welcher die Inferenz von konditionsspezifischen regulatorischen Interaktionen anhand des Prior-Modells durchführt. Hierfür wurde die gegenseitige Abhängigkeit zwischen einem Regulator (Transkriptionsfaktor oder miRNA) und dem Zielgen anhand von Expressionsprofilen bewertet. Das Maß, welches hierfür verwendet wurde ist der nicht-parametrischen, nicht-linearen Korrelationskoeffizienten η^2 (Eta-Quadrat). Dieser wurde aus einer zweifaktoriellen Varianzanalyse abgeleitet. Es wird gezeigt, dass COGERE, sowohl die Integration von Vorwissen, als auch die Bewertung konditionsspezifischer Regulationen aktueller Methoden signifikant verbessert. Desweiteren wird anhand einer Genexpressionstudie von Krebsgeweben gezeigt, dass COGERE eine wertvolle Ressource für die hypothesen-getriebene klinische Forschung ist.

Scientific contributions

The main scientific contributions of this thesis have been published in peer-reviewed scientific journals (•) and presented at conferences (⋆) as listed below.

Chapter 2

- **Ellwanger DC**, Büttner FA, Mewes HW, and Stümpflen V. The sufficient minimal set of miRNA seed types. *Bioinformatics*, 27(10):1346-50, 2011.

⋆ **Ellwanger DC**, Büttner FA, Mewes HW, and Stümpflen V. The sufficient minimal set of miRNA seed types. *German Conference on Bioinformatics* (Freising, Germany), 2011.

⋆ Büttner FA, **Ellwanger DC**, Mewes HW, and Stümpflen V. Large scale analysis reveals novel insights into the characteristics of miRNA targeting. *Lecture Notes in Informatics Edts.*, Schomburg D & Grote A (Braunschweig, Germany), 2010.

Chapter 3

- Berschneider B, **Ellwanger DC**, Shimbori C, White ES, Kolb M, Neth P, and Königshoff M. miR-92a regulates TGF-β1 induced WISP1 expression in pulmonary fibrosis. *Int J Biochem Cell Biol.*, 53:432-41, 2014.

⋆ Berschneider B, **Ellwanger DC**, Mewes HW, Neth P, Königshoff M. Regulation of Wnt1-inducible signaling pathway protein 1 by miRNAs in pulmonary fibroblasts. *Am J Respir Crit Care Med*, 187 A6060 (Philadelphia, USA), 2013.

⋆ Berschneider B, **Ellwanger DC**, Thiel C, Stümpflen V, and Königshoff M. microRNA regulation of WISP1 in pulmonary fibrosis: an *in silico* approach. *Pneumologie*, 65 - A6 (Homburg/Saar, Germany), 2011.

Chapter 4

- Arnold M[†], **Ellwanger DC**[†], Hartsperger ML, Pfeufer A, and Stümpflen V. *Cis*-acting polymorphisms affect complex traits through modifications of microRNA regulation pathways. *PLoS One*, 7(5):e36694, 2012.
  [†] equal contributors

Chapter 5

- **Ellwanger DC**, Leonhardt JF, and Mewes HW. Large-scale modeling of condition-specific gene regulatory networks by information integration and inference. *Nucleic Acids Res.*, Oct 7, 2014.

⋆ **Ellwanger DC**, Leonhardt JF, and Mewes HW. Large-scale modeling of condition-specific gene regulatory networks by information integration and inference. *Workshop Computational Biology @ Bayer* (Boston, USA), 2014.

⋆ **Ellwanger DC**, Leonhardt JF, and Mewes HW. COGERE: modeling of condition-specific gene regulation and regulator gene centrality by information integration and inference. *International Conference on Systems biology* (Copenhagen, Denmark), 2013.

Further contributions

In addition, there are several peer-reviewed scientific articles (•) and presentations (⋆) which are not discussed in this thesis, but were results of relevant collaborations during my PhD course.

- Schulte EC, **Ellwanger DC**, Dihanich S, Manzoni C, Stangl K, Schormair B, Graf E, Eck S, Mollenhauer B, Haubenberger D, Pirker W, Zimprich A, Brücke T, Lichtner P, Peters A, Gieger C, Trenkwalder C, Mewes HW, Meitinger T, Lewis PA, Klünemann HH, Winkelmann J. Rare variants in LRRK1 and Parkinson's disease. *Neurogenetics*, 15(1):49-57, 2014.

- Schulte EC, Stahl I, Czamara D, **Ellwanger DC**, Eck S, Graf E, Mollenhauer B, Zimprich A, Lichtner P, Haubenberger D, Pirker W, Brücke T, Bereznai B, Molnar MJ, Peters A, Gieger C, Müller-Myhsok B, Trenkwalder C, Winkelmann J. Rare variants in PLXNA4 and Parkinson's disease. *PLoS One*, 8(11):e79145, 2013.

- Milosevic J, Pandit K, Magister M, Rabinovich E, **Ellwanger DC**, Yu G, Vuga LJ, Weksler B, Benos PV, Gibson KF, McMillan M, Kahn M, and Kaminski N. Profibrotic role of miR-154 in pulmonary fibrosis. *Am J Respir Cell Mol Biol*, 47(6):879-87, 2012.

⋆ Oehrle B, Irlmer M, Burgstaller G, Beckers J, **Ellwanger DC**, Leonhardt JF, Eickelberg O. The role of aberrant fibroblast invasion in fibrotic lung disease. *Am J Respir Crit Care Med*, 187 A3995 (Philadelphia, USA), 2013.

⋆ **Ellwanger DC**, Büttner FA, Mewes HW, Stümpflen V. Novel miRNA-mediated acute myeloid leukemia progression systems. *International Conference on Systems biology* (Edinburgh, Scotland), 2010.

Contents

List of Abbreviations

AEI	allelic expression imbalance
AGO	Argonaute
ANOVA	analysis of variance
CCR	cross-link centered region
CDS	coding sequence
eQTL	expression quantitative trait loci
FDR	false discovery rate
GRN	gene regulatory network
GWA	genome-wide association
ID	identifier
IPF	idiopathic pulmonary fibrosis
kb	kilobases
LD	linkage disequilibrium
MAF	minor allele frequency
mRNA	messenger RNA
miRNA	microRNA

miRNP miRNA ribonucleoprotein complex

MRE miRNA response element

ncRNA non-coding RNA

nt nucleotide(s)

P *P-value*

pFB primary lung fibroblast

RBP RNA-binding protein

RG regulator gene

RISC RNA-induced silencing complex

RNAi RNA interference

RNP ribonucleoprotein complex

SNP single-nucleotide polymorphism

rRNA ribosomal RNA

SVM support vector machine

TF transcription factor

TG target gene

tRNA transfer RNA

UTR untranslated region

List of Algorithms

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 The elucidation of ncRNA

RNA has long been thought to be the primordial molecule of life and, as such, was the central subject of molecular biology research. To date, this assumption is known as the RNA world hypothesis, proposing that self-replicating RNA molecules devolved their information storage function to the more stable DNA and their catalytic functions to the more chemically versatile polypeptides[1].

In the early 1940s, the prevailing role of RNA was suggested to be an intermediary, a messenger, between DNA and the only functional components of the cell, the enzymes. This view implied that each gene produces an enzyme, a hypothesis which is known as the 'one gene, one enzyme' concept. Since the simple methods applied for gene detection at that time, such as expressed sequence tag sequencing of polyadenylated messenger RNAs (mRNAs) and computational predictions using extrinsic information from comparative genome analysis, were working best for highly expressed, evolutionary conserved protein-coding genes, this idea was broadly accepted[2].

A few years after the elucidation of the DNA structure, in the late-1950s, Francis Crick stated an explanation of the flow of genetic information within a biological system, the celebrated central dogma of molecular biology: genetic information is transcribed from DNA and translated from RNA to proteins. Further, he replaced the existent concept by the almost forgotten 'one gene, one ribosome, one protein' theory. Here, it was assumed that each gene encodes a specific mRNA and an accordant protein synthesis machinery, a

gene-specific ribosome. However, the later finding of a class of stable RNAs comprising polyribonucleotides with low variation in size and base composition in the ribosome challenged this theory. These ribosomal RNAs (rRNAs) aggregate with a variety of proteins forming the translational apparatus which is, on the other hand, programmed by unstable mRNAs. Apparently, rRNA seemed to be functional without being translated and as such the existence of RNA molecules without protein-coding potential, so-called non-coding RNA (ncRNA), was predicted for the first time[2].

The subsequent 'adaptor' hypothesis of Francis Crick described a second class of functional ncRNA which was assigned again a key role in the translation process. He suggested that an RNA molecule mediates between the codon on the mRNA and the corresponding amino acid attached on the encoded polypeptide during the protein synthesis. Since the triplet recognition can be basically conducted by simple Watson-Crick base pairing, RNA seemed to be the evolutionary preferred molecule over proteins[2]. Indeed, the existence of this transfer RNA (tRNA) was experimentally verified by Mahlon Hoagland *et al.* in 1958[3]. Therefore, the defined capacity of RNA was extended from being a pure information-carrying intermediate by additional catalytic and structural roles in the translation process.

Since the fraction of RNA beside rRNA and tRNA was complex, non-abundant and mostly unstable, it was highly unattractive to perform further investigations and, in the end, this fraction was generally assumed to be entirely represented by mRNA. Additionally, the main focus of the molecular biological research was broadly focused on solving the genetic code during this time and thus, there was limited commitment to address the question whether there are more RNAs than the already known ones[2].

Already in 1961 Jacob and Monod[4] speculated in their famous work on the lac operon of *Escherichia coli* that gene expression is controlled through transcriptional regulation conducted by polyribonucleotides. However, a later experiment showed that the locus encoding the lac repressor is translated to a polypeptide which allosterical inhibits the lactose substrate. Thus, this visionary idea faded. However, studies of such kind led to the established transcription factor (TF) paradigm of gene regulation, i.e. gene expression is controlled by proteins binding to *cis*-regulatory elements on the DNA. This concept also emphasized that not all proteins have to be enzymes and that combinatorial interactions of these *trans*-acting regulators may result in a regulatory landscape of high intrinsic complexity that is sufficient to control cell diversity[1].

In the following years, several classes of abundant small RNAs were deliberately isolated, albeit some of them were discovered unexpectedly. Biochemical fractionation led to the finding of heterogeneous protein complexes containing small nuclear RNAs, so-called ribonucleoprotein complexs (RNPs). Later research revealed that these small nuclear RNAs (snRNAs), namely U1, U2, U4, U5 and U6, play a crucial role in the RNA splicing process as part of the spliceosomes by RNA-RNA and RNA-protein interactions[1]. Further functional RNAs were detected such as the small nucleolar RNAs (snoRNAs) conducting the methylation and pseudouridylation of rRNAs, tRNAs and snRNAs. The surprising finding that the signal recognition protein guiding the protein translocation processes is not a protein complex rather a ribonucleoprotein (protein-RNA complex), led to its renaming to signal recognition particle[2]. Yet, the function of all of these small RNAs seemed to be restricted to protein synthesis.

In 1969 Britten and Davidson[5,6] published their unconventional theory on gene regulation in higher cells which attracted a great deal of attention. Based on the observation that the diversity of heterogeneous nuclear RNA was much greater in the nucleus than in the cytoplasm and plant as well as animal DNA contains a large amount of repetitive non-coding sequences, they supposed that gene expression is controlled by extensive RNA-based regulatory networks. However, research in this field focused on gene regulation by TFs rather than RNAs and therefore this idea faded quickly[1]. Later, in 1972, the public opinion was mirrored by Susumi Ohno's article in which he originated the term 'junk DNA'[7] for repetitive non-coding DNA. Therefore, it is not surprising that even after the discovery of introns in 1977, one of the most unexpected findings in molecular biology, or the demonstration of the existence of RNAs with enzymatic capabilities (ribozymes) in the early 1980s, the theory of regulatory ncRNA was not revisited[1].

1.2 Discovery of small regulatory ncRNAs

In 1993, a crucial finding was made by Bruce Wightman *et al.*[8] and Rosalind Lee *et al.*[9] during their investigation of the regulatory processes in the development of *Caenorhabditis elegans*. Until then, it was already known that the lin-14 gene product controls stage specific lineages and is abundant in late-stage embryos and the first larval stage (L1) but was only barley found in the subsequent L2 stage. Further, the gain-of-function

mutation at the lin-14 3'-untranslated region (UTR) locus caused the reiteration of the L1 stage resulting in a the same retarded *Caenorhabditis elegans* phenotype which has been observed for the loss-of-function mutation at the lin-4 locus. Further, lin-4 has been suggested to temporal decrease lin-14 protein levels[8].

Based on this starting position, Lee *et al.*[9] delineated the regulator lin-4 and Wightman *et al.*[8] characterized its mediated regulation. The latter group validated that the protein concentration of lin-14 decreased by a factor of 10 between early and late larval stages, but in this context, they also observed that the RNA level of lin-14 was stable. In addition, they used reporter assays to identify the lin-14 3'-UTR sequence as sufficient component for its temporal regulation. At this, they found conserved sequences of 10 nucleotide(s) (nt) or more which exhibited high complementarity to the lin-4 RNAs. On the whole, they argued that lin-14 is regulated by lin-4 post-transcriptionally via *cis*-regulatory elements located on the 3'-UTR. This was novel and different to the popular concept of transcriptional regulation by proteins[8]. Lee *et al.*[9] gradually realized that they were dealing with a small ncRNA instead of a protein-coding gene. They were neither able to determine a conserved protein sequence nor to identify a conventional position of a start or stop codon in the lin-4 open reading frame. Additional *in vitro* mutagenesis, such as reading frame disruption and non-sense mutations, had no effect on the regulatory function of lin-4. Thus, they suggested that the lin-4 gene did not encode a functional protein. By Northern blot analysis they detected two small lin-4 transcripts, namely lin-4L with 61 nt length and lin-4S with 22 nt length. Both could be mapped on the same region and are transcribed in the same orientation. Therefore, they concluded that the lin-4 gene product is a small RNA, processed from a three-times longer RNA precursor with a putative stem loop structure.

The lin-4 gene remained a single idiosyncrasy of *Caenorhabditis elegans* until the turn of the millennium, when a second ncRNA gene encoding a small post-transcriptional regulator, lethal-7 (let-7), was identified[10]. This ncRNA exhibited the same characteristics as lin-14. But importantly, Amy Pasquinelli *et al.*[11] demonstrated that let-7 was completely conserved and expressed in nematode, fly and humans. Since lin-4 and let-7 showed temporal regulation, they were classified as 'small temporal RNAs' (stRNAs)[12].

As recently as double-stranded RNA interference (RNAi) was discovered[1], targeted

1 For a thoroughly review on RNAi, please refer to the article of Sen and Blau[13].

cloning following size selection of RNAs was performed. In this line, Thomas Tuschl's laboratory[14] detected endogenous small regulatory polyribonucleotides of the RNAi pathway of length $21-22$ nt in *Drosophila melanogaster*. Subsequent works recovering this observation in other organisms led to the perception that RNAi is a more general control mechanism. In consequence, it has been assumed that lin-4 and let-7 are members of a very large family of small RNAs which were termed 'microRNAs (miRNAs)'[2].

Since the first release of the central miRNA repository, miRBase[15], at the end of 2002, the number of registered genes strongly increased from 56 human and 41 murine to 1 881 and 1 193 entities cataloged in the latest release (June 2014), respectively (Figure 1.1) There are multiple gene loci spread over the whole genome whereat several genes are related by a duplication event[16]. For instance, 11 paralogous loci in *Homo sapiens* belong to the let-7 family (e.g. let-7a, let-7b). In general, 31% of human miRNA genes belong to a multimember family which is suggested to have a common ancestor based on hairpin similarity. Further, several human genomic loci exhibit distinct precursor sequences, but express identical mature sequences (e.g. mir-1-1 and mir-1-2). The nomenclature of miRNA genes detected by cloning or sequencing is simply sequential (e.g. mir-1, mir-2). This holds not for genes found in early genetic studies such as let-7 and lin-4, e.g. posterior homologs of the latter are named mir-125. In the literature, genes and the predicted stem-

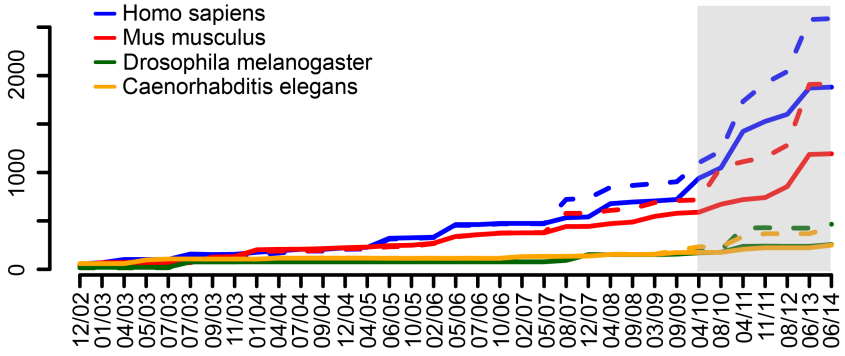


Figure 1.1 | **Cumulative number of registered miRNA genes and mature transcripts.** Shown is the count of listed miRNA genes (solid lines) and mature transcripts (dashed lines) for human (blue), house mouse (red), fruit fly (green), and a roundworm (orange) in the central miRNA repository, miRBase[15], from December 2002 (release 1.0) to June 2014 (release 21.0). In the last four years, the number of registered human miRNA genes grew by a factor of two (mature transcripts by a factor of 2.4).

loop portion of the primary transcript are signified by the term 'mir', the mature product is designated with 'miR'.

Currently, there are 196 families of mature miRNAs known to be conserved among mammals; 34 miRNAs are phylogenetically conserved from *Caenorhabditis elegans* to *Homo sapiens*. At this, some miRNAs have a last common ancestor and are, as such, evolutionary related. However, the 5'-end of their mature sequences diverge implying a distinct targeting pattern (e.g. miR-141 and miR-220c)[16].

It is of note that the majority of listed miRNAs in miRBase have been annotated by high-throughput sequencing. This technology is very sensitive leading to false positives which may be rather decay intermediates of other RNA species. A non negligible fraction of entries are only supported by small numbers of sequencing reads and some of them exhibit varying 5'-ends − a region which is in fact under high selective pressure (but may vary due to authentic isoforms[17]). A detailed verification of miRBase (release 14, 09/2009) had revealed that 173 of 564 (31%) tested loci lacked convincing evidence that they produce genuine mature miRNAs[16].

Since this thesis focused on animal miRNAs, the following sections will exclusively describe features of these species. The miRNA pathway in animals emerged independently from the pathway in plants resulting in different primary modes of actions. However, core components are conserved between both kingdoms[1].

1.3 miRNA biogenesis

miRNAs are encoded by gene sequences with an average length greater than 1 000 nt located within inter- and intragenic, i.e. exonic or intronic, genomic contexts. Here, a good deal of human miRNAs ($\sim 40\%$) are encoded within introns of coding or non-coding transcripts[19,20]. In the following, the canonical and alternative maturation pathways of functional miRNA transcripts are described. An illustration is given in Figure 1.2.

1 For information on the miRNA pathway in plants, please refer to the article of Rogers and Chen[18].

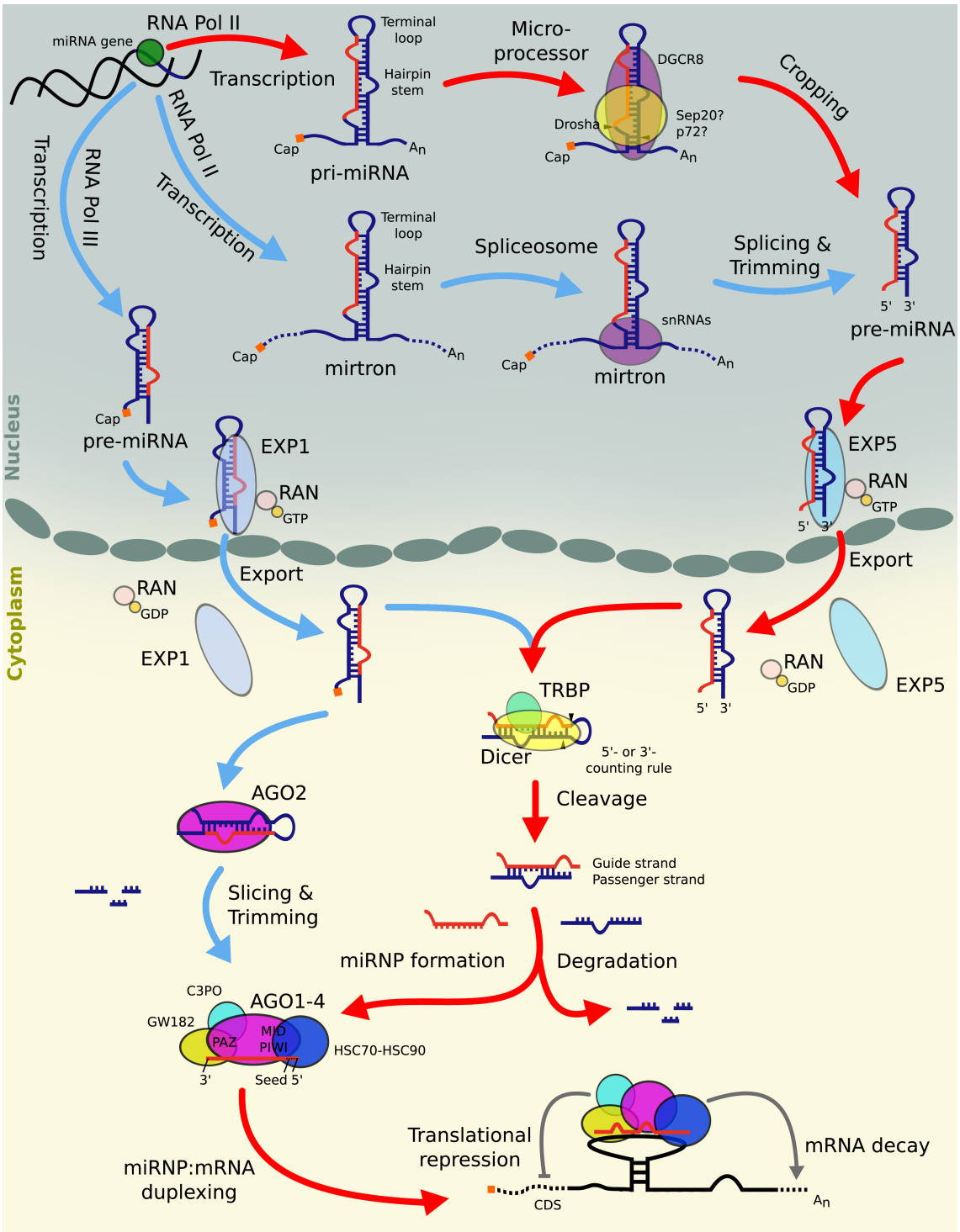


Figure 1.2 | **Pathways of miRNA biogenesis.** *(continued on next page)*

Figure 1.2 *(previous page)* | Pri-miRNAs are produced by RNA Pol II/III. The transcripts are processed by the 'Microprocessor' complex or the spliceosome and then exported by EXP5 to the cytoplasm. Alternatively, EXP1 conducts directly their subsequent nuclear export. The pre-miRNA is processed by AGO2 or Dicer resulting in double stranded RNAs. The guide strand and/or the passenger strand are loaded onto the AGO protein following the miRNP assembly. The canonical pathway is highlighted by red arrows; alternative maturation steps are highlighted by blue arrows.

1.3.1 The canonical maturation pathway

miRNA genes are mainly transcribed by RNA polymerase II and, thus, controlled by correspondent TFs and epigenetic regulations, such as DNA methylation and histone modifications[16]. Since the primary transcripts of miRNA genes are rapidly processed, these are very transient impeding the global mapping of transcription start sites. While these sites have not yet been characterized for most miRNA genes, some promoter characteristics have been inferred from collective analysis of CpG islands, RNA sequencing data, ChIP-Seq[16] and nucleosome positioning analyses and ChIP–chip screens[21]. The transcription initiation site has been found from hundred bases to 20 kilobases (kb) upstream and farther of the miRNA coding region[21,22]. General features of miRNA promoters have been described similar to those of coding genes: conservation, CG content, comprising a TATA element, a TFIIB recognition element, an initiator, a motif 10 element, and a downstream promoter element[21]. Promotors of intronic genes located on the same strand as their host are coincident with the host promoter[23]. However, about one third of intronic miRNA genes have multiple transcription initiation sites and, as such, exhibit independent promoter regions[21]. Furthermore, miRNA loci located in close proximity (up to 50 kb) have been observed to form polycistronic transcriptional units sharing a single promoter[19]. Indeed, these miRNAs are generally co-transcribed, but not necessarily produce active mature transcripts simultaneously[16,24,25]. It has been suggested that genomic clustered miRNAs are encoded by a common primary transcript[20]. Interestingly, miRNAs encoded by the largest human polycistronic locus, C19MC, are transcribed by RNA polymerase III[24].

The primary transcript of miRNA genes (pri-miRNA) is about 500 to 3 000 nt long, often polyadenylated and capped[24], and folds into a typical distinctive structure containing a local hairpin stem of $33-35$ base pairs, a terminal loop, and single-stranded flanking regions on both sides[16]. The enzyme Drosha, an RNase III, processes this stem-loop

by complexing with its co-factor DGCR8 (also known as Pasha or PASH-1), a protein with RNA-binding domains determining the precise cleavage site. This 'Microprocessor' complex endonucleolytically cleaves the 5' and 3' strand of the stem of the pri-miRNA liberating a small hairpin-shaped RNA of about 65 nt length (pre-miRNA)[16]. Recent reports suggested that additional specificity factors, such as the splicing factor SRp20 or the DEAD-box RNA helicase p72, may contribute to pri-miRNA processing[16,24]. Drosha-mediated cleavage occurs co-transcriptionally and does not affect splicing of host pre-mRNAs of intronic miRNA genes, but destabilizes mRNAs hosting exonic miRNAs[16]. It is of note that the 'Microprocessor' complex defines the 5'-terminus of the mature miRNA, the most important region defining its specificity, the so-called miRNA seed[26–28]. Thus, during this processing step, a precise recognition and cleavage is fundamental for a subsequent accurate miRNA target recognition.

Following nuclear processing, the double-stranded pre-miRNA stem with its 2 nt long 3'-overhang is recognized by the shuttle protein Exportin-5 (EXP5)[16]. This protein not only protects the pre-miRNA against nuclear digestion, but also complexes with Ran-GTP to translocate the pre-mRNA through the nuclear pore complex into the cytoplasm[24]. Here, hydrolysis of GTP causes the disassembly of the complex and releases the pre-mRNA to the cytosol[16]. Subsequently, another RNase III endonuclease, Dicer, cleaves the terminal loop of the pre-miRNA inducing a double-stranded RNA of about 22 nt length. The cleavage takes place along with TRBP which binds the double-stranded RNA and activates Dicer through an induced structural rearrangement[24]. The Dicer-TRBP complex follows two rules in determining the cleavage site: i) by recognizing the 3'-overhang generated by Drosha; in this case the cleavage site is located at a typical distance of $21-25$ nt from the 3'-terminus of the pre-miRNA (3'-counting rule) and ii) by binding to the 5'-phosphorylated end of the pre-miRNA; in this case the pre-miRNA is cleaved 22 nt distal from the 5'-end (5'-counting rule)[16]. Notably, the Dicer protein is essential for cell viability as has been shown in knock-out studies which led to lethal phenotypes in mouse[24]. In the end, this processing step yields paired RNAs, termed the miRNA-3p/miRNA-5p duplex, featuring the RNase III characteristic 3'-overhangs at either end[16].

1.3.2 Alternative maturation pathways

As soon as operative miRNAs were observed during deep sequencing experiments of DGCR8, Drosha or Dicer deficient cells, it became clear, that there have to exist alternative functions of the miRNA biogenesis machinery bypassing the 'Microprocessor' or Dicer processing[29–31].

The most prominent alternative, the 'mirtron' pathway, substitutes Drosha cleavage with splicing. Here, loci within short introns produce pre-miRNA mimics. In general, introns of eukaryotic coding or ncRNA are spliced either shortly after or concurrent with transcription by a series of reactions catalyzed by a complex of small nuclear ribonucleoproteins, the spliceosome. Typically, the spliced intron product has an almost linear structure. In contrast, mirtrons exhibit a hairpin potential. During splicing they form the canonical lariat in which the 3'-branch point is ligated to the 5'-end of the intron[32]. Subsequently, the lariat is debranched and adopts the typical pre-miRNA stem-loop structure. This resembled miRNA precursor is qualified to join the remaining canonical pathway, i.e. export to the cytoplasm and processing by Dicer[16,32].

The 'Microprocessor' is also not required for i) endogenous small hairpin RNA genes which express directly transcripts making a tight hairpin turn, such as mir-320, and ii) small RNAs originating from other ncRNAs, such as tRNAs, snoRNAs or viral RNAs[16]. For the former scenario, pre-mir-320 has been observed to be transported unconventionally to the cytoplasm by Exportin-1 instead of EXP5[33].

In all these cases, biogenesis still depends on Dicer. But, it has also been reported that Dicer processing can be bypassed. In example, the precursor of miR-451 has a stem of about 18 nt in length that is too small to be processed by Dicer. In this case, an Argonaute (AGO) protein, in particular AGO2, slices the pre-mir-451 in the center of its 3'-strand liberating a 30 nt long intermediate, the AGO-cleaved pre-mir-451 (ac-pre-mir-451). This transcript has already the potential to regulate target gene expression. However, the ribonuclease PARN produces the intrinsic mature miR-451 transcript by trimming down the 3'-end of the precursor[16].

It should be noted that only about 1% of conserved miRNAs follow one of the alternative pathways. Thus, the vast majority of miRNAs are produced by canonical maturation[16].

1.3.3 The miRNA ribonucleoprotein complex

After the Dicer-TRBP complex disassociates from the miRNA-3p/miRNA-5 duplex, the double-stranded RNA is loaded onto an AGO protein to from a miRNA ribonucleoprotein complex (miRNP), called RNA-induced silencing complex (RISC). The miRNA duplex is unwinded and separated into the guide strand and the passenger strand (miRNA*). In principle, both strands are functional mature miRNAs. But, similar to other RNAi-related pathways, such as small interfering RNAs (siRNA)[34], mainly one strand determines in a context-specific manner the RISC target. Here, the guide strand is complementary to the mRNA target site and will reside in the miRNP; the passenger strand will be degraded[24]. The strand selection is mainly based on the thermodynamical stability of the two ends of the RNA duplex. At this, the guide strand usually exhibits a relatively unstable terminus at the 5'-side. A further criterion may be the first nucleotide of the mature miRNA 5'-end: AGO proteins preferentially select sequences starting with uracil[16]. However, strand selection is not completely deterministic, and for some pre-miRNAs both arms were measured in significant amounts[35]. Further, an event called 'arm switching' has been described. Here, each arm exhibits a tissue-specific thermodynamic stability. It has been suggested that the stability of the duplex ends is determined, at least partly, by alternative Drosha processing[36].

Eight AGO proteins are encoded on the human genome. These are classified into the AGO and PIWI subfamilies. The expression of PIWI proteins is mostly restricted to specific cell lines and primarily functions as repressor of transposons. The AGO subfamily comprises four members, AGO1 – 4, and was found to be ubiquitously expressed[37]. All four AGO proteins are capable of binding miRNA duplexes whereas only AGO2 has an additional slicing activity (Chapter 1.3.2). Consequently, all members of the AGO subfamily are capable of inducing post-transcriptional regulation. In contrast to *Drosophila melanogaster*, no obvious intrinsic feature exists determining the sorting of mature miRNAs to one of the four AGO proteins in humans[16].

Structural studies of AGO2 have revealed that its peptide chain folds to a bilobal architecture, i.e. two lobes composed by two domains each: the N-terminal[1] lobe with

1 N-terminus is the amino-end of the peptide chain.

the N-terminal domain and a PAZ domain, and the C-terminal[1] lobe with a MID domain and PIWI domain[16,37,38]. The interface between the MID and PIWI domain occupies a binding pocket for the 5'-terminal phosphate group of the guide strand, while the PAZ domain binds the miRNA 3'-end (Figure 1.3). The nuclotides $2-10$ of the miRNA 5'-end are located at an RNA binding groove and are pre-arranged in an A-form helix conformation[16]. This enables an effective scanning for mRNA target sites complementary to the miRNA seed sequence.

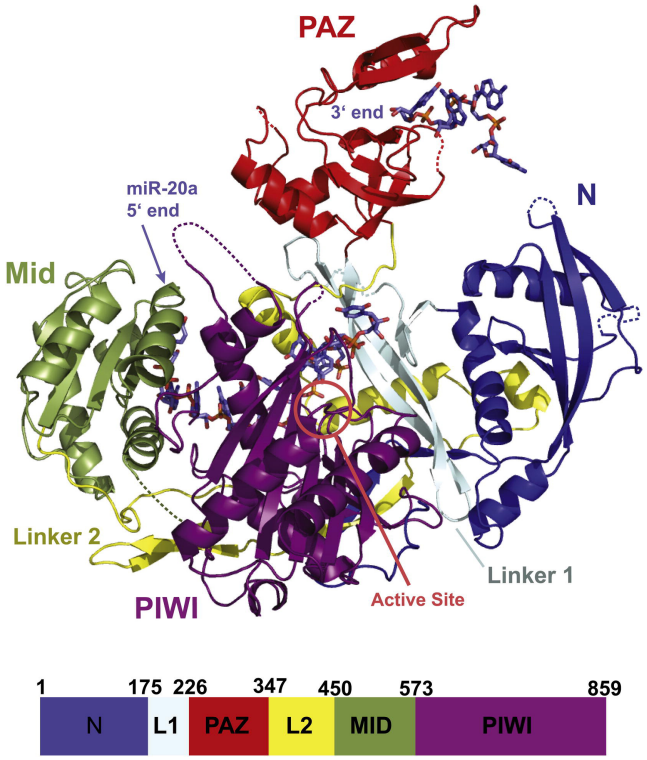


Figure 1.3 | **Structure of the miRNP.** Shown is the crystal structure of the AGO2 and miR-20a complex at 2.2 Å. Each domain and inter-domain linker of the AGO2 peptide is colored, respectively (bar diagram); the active site is highlighted. The miR-20a is displayed as stick model. The complex is formed as follows (starting from the miRNA 5'-end): the first nucleotide is bound to the MID domain (preferentially uracil or adenine[39]), nucleotides $2-10$ are located at the RNA binding groove, and nucleotides $17-20$ are bound to the PAZ domain. The figure is taken from Elkayam *et al.*, 2012[38] with permission of Elsevier (license number 3573240054196).

1 C-terminus is the carboxyl-end of the peptide chain.

To improve its function, AGO recruits and interacts with a set of other proteins. Firstly, the HSC70-HSC90 chaperone complex mediates a conformational opening of AGO using ATP. This facilitates double-stranded RNA loading. Secondly, C3PO activates AGO2 by degrading the passenger strand[40]. Thirdly, and most notably, glycine-tryptophan (GW) proteins play a key role in miRNA-mediated translational repression and mRNA degradation[37]. Their N-terminal multiple gylcine-tryptophan repeats confer binding to the AGO proteins[41].

1.3.4 miRNA turnover

miRNA turnover is the balance between miRNA synthesis and its degradation. Both are vital factors for miRNA homeostasis. In comparison to the increasing number of studies elucidating miRNA transcription and maturation, miRNA half-life and degradation received less attention. Novel insights are just beginning to emerge. Studies silencing transcription or arresting processing enzymes from the miRNA pathway showed that miRNAs still persist for many hours or even days in affected cells[24]. In example, Bail *et al.*[42] found that the transcript level of 95% of measured miRNAs in human embryonic kidney (HEK293T) cells remained stable for at least 8 h following transcriptional shutoff. Thus, miRNAs appear to be generically stable molecules holding half-lifes similar to mRNAs. By comparison, the median half-life of mammalian proteins is about 48 h[43].

However, it has been observed that individual miRNAs also possess differential stability under varying conditions. While their mature transcripts rapidly decayed in specific environments, their precursor levels remained unaffected[44]. In example, human HeLa cells complete a cell cycle in less than one day. At this, miR-29a has a half-life of about 12 h in both phases, interphase and mitosis. The half-life of miR-29b, by contrast, is 4 h in cycling cells and about 12 h in mitotically arrested cells. Since transfection of a miR-29b-3p:miR-29b-5p-like duplex siRNA, lead to similar results, it has been suggested that regulation takes place after miR-29b maturation[44].

To date, only a few miRNA-degrading enzymes are known: 5'-to-3' and 3'-to-5' exoribonucleases, such as XRN1 and EXOSC4, but no endoribonucleases. These so-called 'miRNases' are basically RNases which are assumed to hold a substrate spectrum beyond miRNAs. Thus, it is not surprising that miRNases can be found widely conserved among eukaryotes. Since the number of studies in this field is small, substrate specificity of miR-

Nases remains largely elusive. Currently, little is known about the molecular mechanisms perturbing steady-state levels of miRNAs − even evidence for evolutionary conservation of miRNA turnover pathways is still missing[44].

In a recent publication, Winter and Diederichs[45] reported that miRNA binding in AGO pockets protects it from degradation by miRNases. They observed a longer half-life of the AGO-bound mature let-7a guide strands compared to the unbound passenger strands. While the guide strands were stable for more than 24 h, its passenger let-7a* strands exhibited a reduced half-life of less than 4 h following transcriptional shutoff. Further, after additional AGO2 knock-out they observed that the half-life of endogenous mature miRNAs (let-7a, miR-16, miR-20a, miR-21, and miR-93) dropped from more than 24 h to $9-12$ h. Further, ectopic AGO1-3 expression increased the stability of let-7a* which has been tested functional on its target genes (TGs).

1.4 miRNA-mediated regulation of gene expression

1.4.1 Recruitment of the miRNP to mRNA targets

After ATP-independent unwinding and degradation of the passenger strand, the miRNP effector complex will conduct the interaction with its target. In contrast to lin-14, the majority of metazoan mRNAs do not carry several regions of extensive complementarity to their regulatory miRNAs. Hence, several additional features are important for target recognition.

As described before, the seed region (nucleotides $2-10$) of the miRNA 5'-end is pre-arranged in an A-form helix conformation in the effector complex (Figure 1.3). In particular, bases $2-6$ are positioned for nucleating the interaction of the miRNA with the target mRNA. Consequently, the AGO protein reduces the entropic cost of unfolding the miRNA for target pairing. Studies have shown that the miRNP exhibits an about 10 times faster target detection and an about 300-fold higher affinity to its targets compared to unbound miRNAs[46,47]. Here, the higher the complementarity, the higher the rates of miRNA-guided binding by AGO2. However, only the seed region has a significant contribution whereas the miRNA 3'-terminal region plays only a secondary role. There is no experimental evidence that the extent of 3'-pairing correlates with the extend of gene regulation. Nonetheless, small loops in the seed:target hybrid may be tolerated if the

pairing in the 3'-region is extensive, albeit these kinds of targets are very rare. At this, the type of mismatch and its precise position within the duplex are important[43].

Further, seed complementary regions of length < 5 nt are unlikely to be bound by the miRNP[48]. Of note, miRNAs with almost identical seed sequences form families which target the same set of mRNAs; 64% of currently known human miRNAs (listed in miRBase[15]) are related by their seed sequence. Most of the family members are not co-expressed, but regulate their targets in different environments. In contrast to miRNA clusters, seed-related miRNAs were suggested to be under positive selection maintaining their specificity[43].

Since mRNAs are bound by translating ribosomes, functional miRNP interaction is infrequently conducted at the coding sequence (CDS) and up to approximately 15 nt downstream of the stop codon. The 3'-UTR has the highest density of miRNA target sites. Here, sites are preferentially located at the beginning or end of long 3'-UTRs. Since 3'-UTRs expanded during evolution, the sequences around miRNA target sites might have emerged early. Consistently, these target sites were found to be under strong evolutionary selection. Another spatial constraint is given by interactions of other RNA binding complexes. Although, sites positioned in the CDS hold minor regulatory effects, their interplay with 3'-UTR sites contribute significantly to the target regulation[43].

Due to its integration in the miRNP, miRNAs exhibit an optimal conformation for duplexing with the mRNA. On the contrary, mRNAs do not have an unpaired native structure in thermodynamic equilibrium and, thus, energy is required to unfold the target region. Kertesz *et al.*[49] suggested that the miRNA complementary site as well as flanking regions up- and downstream have to be opened. An increased local site accessibility, i.e. a lowered required energy for unfolding, raises the miRNP binding affinity[49]. Consequently, miRNP binding is also affected by the local adenine or uracil content around the target sites[50]; in contrast to guanine and cytosine which interact via three hydrogen bonds, adenine binds to uracil only via two hydrogen bonds.

1.4.2 Translational repression and mRNA decay

After the effector complex has recognized its target, the miRNP mediates target regulation by translational repression or mRNA degradation. The former regulatory mechanism precedes or follows translation initiation. Besides several important factors that are

involved in the translation process (initiation, elongation, and termination), it is crucial that mRNAs posses a 5'-cap structure[1] and a 3'-poly(A) tail[2]. In the cytoplasm, the cap-binding complex eIF4F (composed of eIF4E, eIF4G, eIF4A) and the protein PABPC associate with the 5'-cap and the 3'-poly(A), respectively. Their physical interaction results in a circular mRNA which can be efficiently translated and is protected from degradation[51].

Predominantly, the miRNP targets the cap structure or interferes with the function of either eIF4F or PABPC. This regulation takes place at translation initiation[51]. Affected mRNAs are transported to P-bodies[3] for either degradation or storage. But also miRNPs and their targets were observed in functional units of protein synthesis (ribosomes). It has been suggested that peptide elongation is slowed down or aborted by a ribosome drop off. Further, proteolytic cleavage of the nascent polypeptide occurs co-translationally[52].

However, transcriptome profiling and studies of single miRNA:target pairs showed that the predominantly miRNP mode of action is mRNA degradation. In this case, the miRNA target level inversely correlates with the abundance of the miRNA. Since direct endonucleatic cleavage by miRNPs occurs only for fully complementary targets, this is a rare mechanism in animals. Instead, for partially complementary targets, the miRNP initiates the cellular 5'-to-3' mRNA decay pathway. Here, the 3'-poly(A) tail is removed by deadenylases (CAF1-CCR4-NOT complex), followed by decapping through DCP2 and subsequent 5'-to-3' exonucleolytic digestion by XRN1. This can occur either after or before translation initiation. In the latter case, mRNA polydeadenylation interferes the PABPC binding. Independent of the point in time, it has been reported that rapid mRNA destabilization provides the main contribution to protein output reduction in animal cell cultures[51].

As mentioned in Chapter 1.3.3, AGOs complex with GW proteins, particularly GW182[4], to silence miRNA partially complementary targets. The GW182 silencing domain at the mid and C-terminal regions interact with PABPC. This either blocks the PABPC:eIF4G interaction or reduces the affinity of PABPC for the mRNA 3'-poly(A). In both cases, target mRNA circularization is prevented and consequently translation is inhibited. Further,

1 5'-cap is a $m^7G(5')ppp(5')N$ structure at the mRNA 5'-terminus.
2 3'-poly(A) denotes multiple adenosine (A) monophosphates at the mRNA 3'-terminus.
3 P-bodies are distinct foci in the cytoplasm consisting of several enzymes involved in mRNA turnover. Absorbed mRNAs mainly undergo decay; some mRNAs will be released to re-initiate translation.
4 GW182 is also known as TNRC6A.

GW182 recruits the CAF1-CCR4-NOT complex to initiate the 5'-to-3' mRNA decay pathway. Since the open poly(A) tail conformation is more exposed to decay enzymes, target degradation is facilitated[51].

Notably, Vasudevan *et al.*[53] demonstrated that the interaction of AGO2 with the 3'-UTR of tumor necrosis factor-α, under specific cellular conditions, lead to upregulation rather than downregulation of translation. They suggested that AGO2 is part of a functional miRNP complexed with FXR1. The miRNP:mRNA pairing may induce translational activation by interfering with inhibitory RNA-binding proteins (RBPs) at the 3'-UTR – a scenario which has also been observed the other way round: RBPs, such as ELAV1 or DND1, interfere with the miRNP-mediated repression of translation and subsequently acts as translational activator[52]. Translational activation by miRNAs was also documented as a common function of miRNPs on cell cycle arrest. Intriguingly, translational regulation may oscillate between repression and activation during the cell cycle[54].

1.4.3 Regulatory roles

The initial paradigm of miRNA regulation was based on the lin-4:lin-14 interaction and, thus, complies with the role of miRNAs as binary 'off-switches'. In this case, the miRNPs decrease the protein output of their targets to inconsequential levels, i.e. they switch their targets off. Here, two temporal distinct scenarios can occur: either miRNAs repress translation of pre-existing mRNAs or target transcripts are trapped by already matured miRNAs[27].

However, miRNA transfection experiments have shown that miRNAs only modestly repress the translational output, rarely resulting in more than a $2-4$ fold reduction on protein levels[43,51]. In this case, miRNAs act as a rheostat rather than a binary off-switch conferring robustness to biological processes. This 'tuning' interaction dampens the protein output to a more optimal, but still functional, level[27].

Nevertheless, these small expression changes of any individual target are difficult to reconcile with the malformed phenotypes caused by perturbed miRNA regulation. Thus, it has been suggested that the observed effect of miRNAs is obscured by several factors: i)

timing of the experiment[1], ii) feedback loops involving expression of transcriptional regulators, and iii) feedback mechanisms with the maturation pathway, such as the regulation of Dicer by let-7 or the repression of GW182 by miR-30[43]. Further, miRNAs increase their impact by coordinated targeting of multiple transcripts of a particular pathway or protein complex[55].

Recently, Mukherji *et al.*[56] analyzed the level of miRNA repression in single cells. They observed an average modest level of repression which is in line with previous population-based studies. But, they also reported dramatically differential effects among individual cells. This cell-to-cell variation was strongly affected by the available miRNA concentration, the target mRNA level, and the strength and number of embedded miRNA binding sites at the target sequence. Given a specific miRNA target expression threshold, strong repression occurs at low mRNA levels (below the threshold) and weak repression at high mRNA levels (above the threshold). It has been suggested that if the miRNA pool is not saturated, then all targets of a specific affinity for the miRNA will be exposed to the same degree of repression. But, by increasing the mRNA concentrations, the miRNA pool becomes gradually saturated, i.e. all miRNAs are duplexed with their targets, and the number of mRNAs escaping miRNA-mediated regulatory mechanisms raises. Thus, miRNAs can act as both, off-switches for targets expressed below the threshold and as fine-tuners for targets with transcript levels ranging between the threshold and minimal repression at high mRNA concentrations. The regulatory impact is increased by higher complementarity (Chapter 1.4.1) and multiple target sites[56]. Since the majority of target sequences have more than four conserved binding sites of multiple miRNA families per 3'-UTR[57], strong repression can be a result of synergistic miRNA regulation. Notably, one miRNA is able to bind hundreds to thousands of expressed RNAs. Although a large fraction of interactions is operative, several of the participating RNAs denote off-targets or non-coding miRNA-sequestering agents, such as pseudogenes and long ncRNA[55]. Depending on the environment, also some mRNAs act as miRNA decoy: they bind miRNPs, but can be degraded without functional consequences. This phenomenon has been described as neutral interaction between a miRNA and its antitargets elsewhere[27].

1 The rate of AGO loading is about 10 h, the median half-life of target proteins is about 48 h[43]. Thus, early measurements will imply lower effects.

The more competitive endogenous RNAs (ceRNAs) [1] are available at the transcriptome, the lower the effective miRNA concentration, and the higher the impairment of miRNA activity. In total, the frequency of functional *bona fide* miRNA targets and the amount of natural coding and non-coding miRNA decoys[2] determine the location of the individual miRNA target threshold[55]. Therefore, different tissues or conditions that exhibit distinct expression profiles account for changing target thresholds and, in the end, different repression strengths.

1.4.4 miRNA-mediated genetic networks

Regulation of gene expression is crucial for cellular processes as it governs the availability and activity of cellular components. This regulatory control is conditionally modulated and expression is highly dynamic over a wide range from rapid, short responses to slow, lasting adaptions.

TFs bind to *cis*-regulatory elements on the DNA to regulate the flow of genetic information from DNA to RNA. Subsequently, miRNAs regulate post-transcriptionally the mRNA and protein levels. To understand the regulatory activity of a genome, the reconstruction of the whole ensemble of *cis* and *trans* elements is required. To elucidate the roles of miRNAs, an integrated network analysis is performed emerging from summation of the interactions of miRNAs and targets. The result is a so-called 'miRNA-mediated genetic network'[3] consisting of genes that are regulated by other gene products, i.e. RNAs or proteins.

Regulation is an interplay of various transacting factors on different layers, not only accomplished by a single force. As described in Chapter 1.3, miRNA genes are controlled by TFs; Enright *et al.*[61] observed that miRNAs preferentially target TFs. This enables beside co-operative and competitive, also mutual regulation. Consequently, the interplay of transcriptional and post-transcriptional interactions results in a regulatory landscape of high intrinsic complexity. Usually, the underlying networks are highly wired. However,

1 ceRNAs are *bona fide* coding and ncRNA targets competing for miRNA binding[58]. For a recent review refer to Tay *et al.*[59].

2 miRNA decoys are also known as miRNA 'sponges'[43].

3 Basic genetic networks contain only TF regulation. Current studies, such as Cohen *et al.*[60], have shown that miRNAs complete these networks.

there are various regulation circuits emerging repeatedly in miRNA-mediated genetic networks. Often TFs and miRNAs form feed-forward loop motifs[1] in which either i) a miRNA and its target is regulated by a common TF or ii) a TF and its target is regulated by a common miRNA[62]. For the former class of motifs, Hornstein *et al.*[63] suggested that miRNAs are dedicated to buffer stochastic perturbations. They distinguished their roles between 'coherent' and 'incoherent' feed forward loops.

The logic of this circuit is 'coherent' in that the TFs regulation of its targets is consistent. In example, the TF:miRNA interaction is positive (stimulation) whereas the remaining interactions are negative (repression) [2]. Here, the post-transcriptional repression is synergistic with the transcriptional inhibition of the same target. Thus, miRNAs antagonize 'leaky' mRNA of TGs which are already transcriptionally repressed. Reciprocally, TFs may stimulate target transcription and repress miRNA production forming another 'coherent' logic [3][63]. Here, existing miRNAs buffer stochastic bursts [4] of TF induced target transcription.

In 'incoherent' feed forward loops, the direct and the indirect regulation executed by the TF are opposing. In example, the TF:miRNA and the TF:target interaction are positive (stimulation) whereas the miRNA regulation is negative (repression) [5]. Here, the co-regulated miRNA performs fine-tuning. Target expression variation arising from extrinsic noise, such as TF concentration or activity, is reduced[55].

A more simple motif which was also found enriched in the architecture of these networks is the reciprocal regulation between TF and miRNA. In this feedback loop, the TF stimulates miRNA transcription and the miRNA induces mRNA degradation of the TF. Again, miRNAs hold a role as buffers against fluctuation in gene expression[55].

In the end, integrating miRNAs in genetic networks leads, at fist glance, to an increased dimension of connectivity. However, viewing miRNAs in this systems context reveals their role as buffers of transcriptional noise to provide robustness in these networks[43]. It should be noted that a markedly buffering effect requires both a rapid change in miRNA

1 Feed forward loop motif composed of A, B, and C: $A \rightarrow B$, $A \rightarrow C$, $B \rightarrow C$.

2 Type 3 'coherent' feed forward loop for TF A, miRNA B, and target C: $A \xrightarrow{+} B$, $A \xrightarrow{-} C$, $B \xrightarrow{-} C$[64].

3 Type 4 'coherent' feed forward loop for TF A, miRNA B, and target C: $A \xrightarrow{-} B$, $A \xrightarrow{+} C$, $B \xrightarrow{-} C$[64].

4 Transcription can occur in bursts (pulses) resulting from the stochastic nature of biochemical events[55].

5 Type 1 'incoherent' feed forward loop for TF A, miRNA B, and target C: $A \xrightarrow{+} B$, $A \xrightarrow{+} C$, $B \xrightarrow{-} C$[64].

concentration and a prompt target repression. A slow miRNA response will only slightly dampen the amplitude of the target mRNA fluctuation. Further, the local context of each circuit has to be considered. Since all motifs are embedded in a global network, the TF, the miRNA as well as their target are likely regulated by multiple regulators. Thus, more sophisticated models will be needed to describe the role of miRNAs in more complex conditions[55].

1.5 Experimental identification of miRNA targets

1.5.1 Transcriptome and proteome analyses

The first miRNA:target interaction, namely lin-4:lin:14, was determined using two techniques, genetic screening and reporter assays. In the former method, candidate interactions are selected by gene mutations that rescue a miRNA loss-of-function phenotype. Besides the advantage that the identified targets can be directly linked to a phenotype, this method holds several drawbacks. In example, the experiments are laborious, challenging to conduct in mammals, and they yield direct as well as indirect targets. Most miRNAs are not individually essential for the outcome of a specific phenotype. Other mechanisms, such as targeting by miRNA families, can act in a compensatory manner[43]. In the reporter assay, 3'-UTRs with computationally predicted target sites are cloned into luciferase reporter vectors. The quantification of the reporter gene expression following miRNA induction (e.g. by miRNA mimics) or inhibition (e.g. by anti-miRNAs) indicates that the gene of interest is regulated by the miRNA through their 3'-UTR[65]. To identify potentially regulated 3'-UTRs, computational predictions are performed. Hence, these experiments are biased and restricted to miRNA target sites defined by the implemented model. Further, genome-wide detection of miRNA targets is not feasible.

Since miRNAs mainly negatively regulate their targets, the loss/gain of miRNA function should lead to an increased/decreased target expression. Therefore, a series of miRNA overexpression and inhibition studies were performed. Early experiments transiently transfected tissue specific miRNAs into cells where they are normally not endogenously expressed. Subsequent microarray analyses or RNA sequencing (RNA-seq) were used to identify mRNAs which exhibited higher decay rates following miRNA transfection. Indeed, this approach enables the identification of a large set of targets, but it also holds a

high fraction of false positives caused by off-target effects [1]. To overcome this problem, the experiment was conducted the other way round: the target miRNA was inhibited in any cell of interest using complementary exogenous oligonucleotides. It was expected that the target mRNA is significantly upregulated following miRNA inhibition[65].

Similarly, stable isotope labeling by amino acids in cell culture (SILAC)[2] approaches were used to detect proteins that are affected by changes in miRNA expression. In contrast to transcriptome-based approaches, these methods are sensitive to mRNA destabilization and translational repression. Differences in protein synthesis are computed from mass spectrometry measurements of metabolically (pulse-)labeled peptides containing heavy isotopes of essential amino acids[65].

Transcriptome and proteome analyses are limited to targets that exhibit an expression change to a certain extent. Since most of miRNA effects are modest, it is difficult to distinguish primary from downstream effects[3]. In addition, several candidate targets do not follow the methodical assumption, i.e. these are upregulated post-transfection (Chapter 1.4.2). To overcome these drawbacks, the direct interaction between the miRNA and its target has to be identified. For this purpose, either the miRNA was labeled, e.g. with a biotin-tag, or the AGO protein was labeled, e.g. with an epitope-tag, in transfection experiments. This allowed the subsequent isolation and quantification (microarray or RNA-Seq) of associated mRNAs[65].

1.5.2 The AGO-bound CLIP-Seq protocol

While whole transcriptome and proteome analyses provide a quantitative view of the regulatory effect of miRNAs, they have the profound disadvantage that they do not directly reveal miRNA:mRNA interactions. Further, one experiment is restricted to dissect the targetome of only one of more than 2 500 mature human miRNAs. Considering that

1 Off-target effects arise when unintended or stochastic base pairings with the introduced RNA occur.

2 In the SILAC method, cells from two samples are grown in two different media, respectively: one medium containing amino acids labeled with light isotopes (miRNA overexpression/inhibition) and one medium with heavy-isotope-labeled amino acids (normal miRNA expression). The changes in protein abundance are computed by the ratio between the signal from the light and heavy isotopes obtained by mass-spectrometry[43].

3 Secondary effects arise when the activity of a true miRNA target is affected, e.g. suppression of a TF by a transfected miRNA leads to downstream effects in the TF target expression.

the transcriptome is cell-line dependent, an excessive number of experiments has to be conducted.

While the RISC targeting depends on stable physical association to the mRNA target, its isolation and extraction opens the possibility to identify miRNP:target interactions *in vivo*. Immunoprecipitation of proteins associated to the RISC, such as GW182 family members[66] or AGO[67], provided the means of direct identification of target mRNAs stably coupled with active complexes. Although this approach provides large datasets of high-confidence miRNA targets[68], the precise location of the miRNP binding site was still hidden.

Recent high-throughput methods based on AGO cross-linking and immunoprecipitation (AGO CLIP) overcome this drawback (Figure 1.4). Ultraviolet (UV) light is used to induce protein-RNA cross-links[1] between endogenous AGO and its associated guide miRNA:mRNA duplex. Then, partially RNase-digested AGO-RNA complexes are isolated by highly specific monoclonal antibodies and size-fractionated. Bound RNA molecules are recovered and converted to complementary DNA (cDNA) by reverse transcriptase. The resulting cDNA library is deep sequenced (CLIP-Seq) and the reads are mapped to the genome[70]. Based on that, clusters are computed, e.g. by estimating the enrichment of CLIP-Seq reads in relation to the expected number obtained from the relative mRNA abundance from RNA-Seq or microarray data. Typically, RNA sequence mutations are introduced by sample preparation at the cross-link regions. These so-called cross-linking-diagnostic mutations are used to increase the efficiency of miRNP target site identification. Further, CLIP-Seq reads which were mapped to miRNA hairpins denote AGO-associated miRNAs. Notably, the guiding miRNA and its target site are not captured together. The specific miRNA:mRNA hybrid remains to be inferred computationally[43].

Three variants of the AGO-bound CLIP-Seq protocol have been established. Firstly, high-throughput sequencing of RNAs isolated by CLIP (HITS-CLIP) uses UV C light[2] inducing cross-linking-diagnostic deletions. The approach was initially employed by Chi *et al.*[71] in mouse brains. They identified miRNP-binding regions, termed 'average AGO-mRNA footprints', where AGO bound within 62 nt of cluster peaks $\geq 95\%$ of the

1 Irradiation of cells with UV light causes the formation of covalent bonds (cross-links) between proteins and nucleic acids which are in close contact[69].

2 The used UV C light had a wavelength of 254 nm[71].

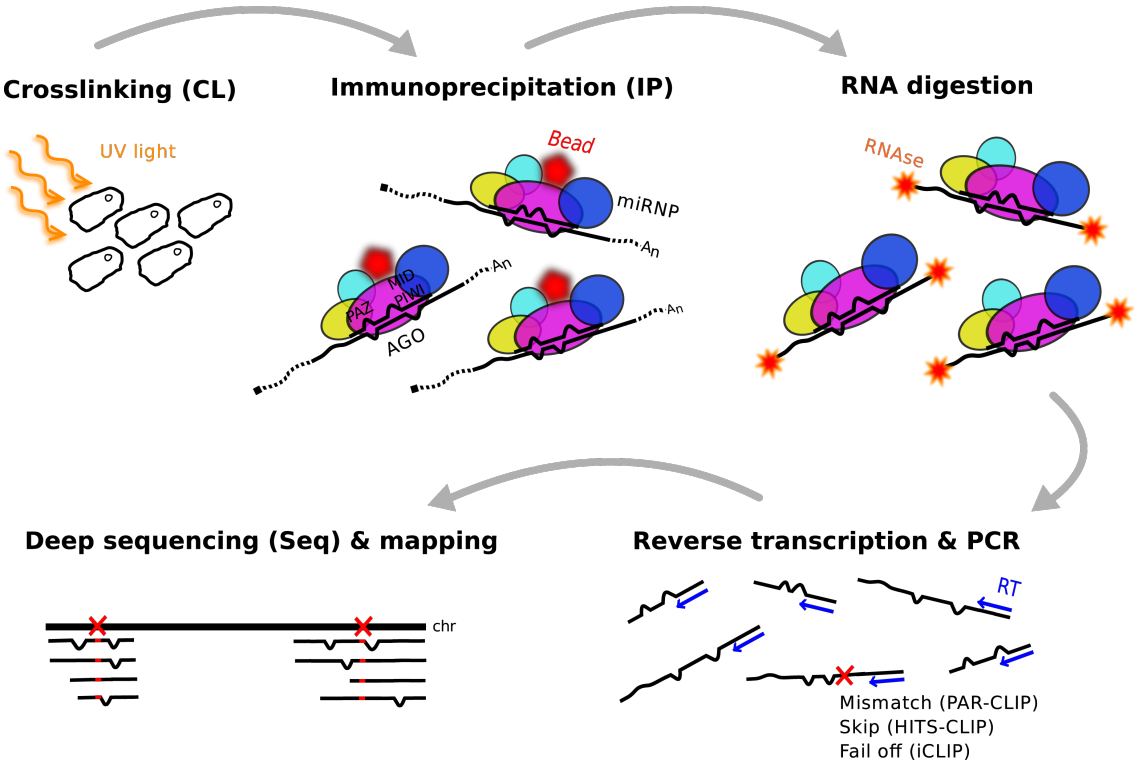


Figure 1.4 | **The AGO-bound CLIP-Seq protocol.** Illustration of the preparation of an AGO CLIP-Seq library. First, covalent bonds are induced between RNA and the AGO protein using UV radiation (254 nm or 365 nm). Here, using photoactivatable analogs of ribonucleosides has been shown to enhance RNA-AGO cross-linking[70]. The RNA-protein complexes are isolated by immunoprecipitation. Then, bound RNAs are partially RNase-digested. A cDNA library is prepared for subsequent high-throughput sequencing. Each protocol introduces typical RNA cross-linking-diagnostic mutations. These make the reverse-transcriptase enzyme (RT) error-prone at the cross-linked regions. HITS-CLIP induces RNA lesions, PAR-CLIP generates U to C mutations, and iCLIP causes truncation at the cross-linking site. The cross-linking-diagnostic mutations facilitate the genome-wide mapping of miRNP binding sites.

time. Secondly, Hafner *et al.*[70] presented the photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP) protocol: they incorporated 4-thiouridine into RNA and used UV A light[1] causing distinctive mutations to deoxycytidines in the CLIP-Seq reads. Compared to UV C light approaches, this method required less sequence reads to capture cross-link evidence due to its high rate of T to C changes. They identified 41 nt long clusters

1 The used UV A light had a wavelength of 365 nm[70].

which were centered over the predominant cross-linking site, termed cross-link centered regions (CCRs). Both protocols, HITS-CLIP and PAR-CLIP, have been successfully used to extract transcriptome-wide AGO2 binding sites in a human cell line[49]. In late 2013, a third variant, namely individual-nucleotide resolution CLIP (iCLIP), has been applied for *Caenorhabditis elegans*[72]. Similar to HITS-CLIP this method uses UV C light in the absence of photoreactive nucleotides. Its distinguishing feature is that it takes advantage of the propensity of the reverse transcriptase to stop polymerizing at cross-linked nucleotides. This enables the miRNP binding site capture at nucleotide-level resolution[43].

Notably, Helwak *et al.*[73,74] presented an extension of the AGO CLIP-Seq protocol recently. They included an additional step in which the miRNA is ligated to its target site. Thus, miRNA:target site chimaeras are sequenced rather than each part of the hybrid independently. Subsequent computational simulations were used to infer the structure of the hybrid. Indeed, the idea seems compelling for improved miRNA target identification. However, the efficiency of their cross-linking, immunoprecipitation and sequencing of hybrids (CLASH) protocol is low; the identified targets respond only weakly to miRNA perturbations. To be capable of mapping and modeling comprehensibly the whole miRNA targetome, the CLASH method requires further improvements[43].

1.6 Motivation and outline of this thesis

The first report stating the idea of regulatory RNA dates back 50 years. As recently as 1993, the first small ncRNA, the *Caenorhabditis elegans* RNA lin-4, was described. The later discovery of the RNA let-7 and its conservation from worms to humans initiated a small RNA revolution. It is now 10 years ago since the scientific imprinting of the term 'microRNA' (miRNA), but our biological comprehension of this molecule is still limited yet. Indeed, it became clear that post-transcriptional regulation by miRNAs is important for crucial cellular processes; consistently, its dysfunction can lead to fatal phenotypes. However, we are only just beginning to understand the nature and the extend of miRNA regulation. The number of known miRNA genes and mature transcripts raises continuously − almost all of which had evaded prior detection. Also the number of studies investigating

therapies targeting miRNAs[1] in human diseases is growing with some very promising first results[76,77]. Thus, the elucidation of this molecule in terms of biogenesis, target detection and regulation, and function is of high relevance − for our basic molecular biological understanding, but also to improve biomedical science.

Recent technologies enabled significant progress in this field, such as the successful isolation of the active miRNP. This enabled, amongst others, its structural delineation and the transcriptome-wide identification of miRNP binding sites. For the latter, two techniques were published in 2009 and 2010, denoting milestones compared to previous experimental protocols: AGO HITS-CLIP and AGO PAR-CLIP. The knowledge of miRNA target sites has been largely obtained from measurements of expression changes following miRNA transfection or inhibition. As discussed in Chapter 1.5.1, these transcriptome and proteome analyses bear a variety of limitations. The most critical issue is that they are not able to identify the miRNP binding sites. AGO-bound CLIP-Seq data identifies not only miRNA-target interactions with high specificity[71], but also reveals precise miRNP binding regions. This thesis aims to address the computational modeling of miRNA-mediated regulation in consideration of novel information obtained from AGO-bound CLIP-Seq data analysis.

In **Chapter 2**, I revisited the current model of miRNA target recognition. Since the miRNA targetome is elaborate and the experimental detection is a costly and time-consuming process, construction of miRNA-mediated regulatory networks heavily relies on computational miRNA target prediction. At this, it is generally accepted that the pairing between the target sequence and the seed sequence of the miRNA 5'-end presents the most important feature. However, prediction algorithms apply different seed paradigms to identify miRNA target sites. Limited by the experimental methods, previous studies defining the miRNA seed sequence have been restricted to noisy assessments, such as signal-to-noise ratio, and degree of mRNA or protein repression.

I present an approach to prepare murine and human CLIP-Seq data to construct interaction maps and to discriminate, for the first time, between functional and non-functional

1 Strategies addressing *decreased* miRNA levels in disease: miRNA gene re-expression by epigenetic drug treatment, transfection with exogenous pre-miRNAs by miRNA virus delivery systems, and enhanced miRNA processing by drugs (e.g. enoxacin).[75]
Strategies addressing *increased* miRNA levels in disease: complementarity-based miRNA inhibition by synthetic antisense oligonucleotides (anti-miRNAs, antagomirs or locked nucleic acids) or vectors containing multiple miRNA binding sites (miRNA sponges)[75].

target sites in a bulky and quantitative manner. By implementing a separate-and-conquer algorithm, I defined a canonical minimal and sufficient set of six seed types and examined their potential impact on target transcript stability. Here, I showed that the regulatory effect depends on the length and the start position of the seed pairing. Further, I evaluated the seed feature for miRNA target prediction. The specificity of long seeds was confirmed, but the majority of functional target sites was formed by less specific seeds of only 6 nt indicating a crucial role of this type. Since common target prediction is restricted to long seed sites, the majority of functional sites remains uncovered. Conservation analysis revealed that a substantial fraction of genuine target sites was non-conserved and, as such, lineage-specific.

Parts of this chapter were published in the journal *Bioinformatics*[1] in collaboration with Florian Büttner [2], Volker Stümpflen[2] and Hans-Werner Mewes[2] [78].

Chapter 3 extends the previous study. Although the base-pairing of the miRNA seed is a strong determinant of target site detection, the existence of a $6-8$ nt long miRNA seed complementary sequence does not necessarily imply a functional miRNA:mRNA interaction. As such, the false positive rate is considerably high rendering the prediction of reliable miRNA target sites still an unsolved computational challenge. Additional characteristics of the binding site context that influence target sensitivity to miRNA repression are required. Thus, I was interested whether discriminatory features can be found to predict mRNA regions preferentially bound by the miRNP using the AGO-bound CLIP-Seq data. For this purpose, I present an approach using a machine learning technique. A training and evaluation set of positive and negative instances was prepared. Several features of miRNP binding sites were extracted and scored by a sliding window approach. Subsequently, a support vector machine (SVM) classifier was trained. The novelty of this analysis is three-fold: i) an elaborate data basis composed of two CLIP-Seq libraries was used, ii) the features were selected and analyzed with the objective to describe miRNP binding sites, and iii) the resultant classifier is able to identify miRNP:mRNA interacting regions unbiased of any miRNA, i.e. the initial target region search does not require a miRNA seed match. By combining the SVM classifier with common prediction methods, the precision of determined targets was shown to be improved.

1 *Bioinformatics*, Oxford University press
2 Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Germany

Further, a biological use-case applying the results from Chaper 2 and Chapter 3 is presented. Pulmonary fibrosis is the most common and fatal form of idiopathic interstitial pneumonia. Königshoff *et al.* [79] reported that the WNT1-inducible signaling pathway protein 1 (WISP1) is a highly expressed pro-fibrotic mediator in idiopathic pulmonary fibrosis (IPF). However, its regulation remains to be elucidated. In collaboration with Barbara Berschneider[1] and Melanie Königshoff[1], the hypothesis was examined that WISP1 eludes post-transcriptional control by miRNAs in pulmonary fibrosis. For this purpose, I prepared and analyze miRNA expression studies to select a set of candidate regulators. By applying the novel classifier, screening for the set of canonical seed types and conducting subsequent structural analyses of the miRNA:mRNA hybrid, I predict miR-92a as most promising candidate regulating WISP1. Experimental verification of Barbara Berschneider and collegues showed that this miRNA and WISP1 are significantly associated in experimentally lung fibroblasts and lung tissue specimens of IPF patients. Notably, miR-92a reverses TGF-β1-induced WISP1 mRNA expression in lung fibroblasts and miR-92a inhibition increases WISP1 protein expression. Concluding, these findings constitute a novel regulatory role of miR-92a for WISP1 expression in pulmonary fibrosis.

Parts of this chapter were published in the journal *The International Journal of Biochemistry & Cell Biology*[2] in collaboration with Barbara Berschneider[1], Hoeke Baarsma[1], Cedric Thiel[1], Chiko Shimbori[3], Eric White[4], Martin Kolb[3], Peter Neth[5], and Melanie Königshoff[1] [80].

In **Chapter 4**, the question was examined whether genetic variance affects miRNP binding, e.g. by disrupting the miRNA seed complementary sequence or the local folding of the target segment. Our current knowledge on the function of non-coding variants, in particular on SNPs affecting the miRNA regulation pathway, is limited. In collaboration with Matthias Arnold[6], a set of trait-associated index single-nucleotide polymorphisms (SNPs) and proximal SNPs in strong LD was prepared. The analysis of their genomic position indicated that single nucleotide mutation may influence miRNA regulation: the SNPs

1 Comprehensive Pneumology Center, Helmholtz Zentrum München, Germany
2 *The International Journal of Biochemistry & Cell Biology*, Elsevier
3 Department of Medicine, McMaster University, Hamilton, Canada
4 Division of Pulmonary and Critical Care Medicine, University of Michigan, Ann Arbor, USA
5 Institute for Cardiovascular Prevention, Ludwig-Maximilians-Universität München, Germany
6 Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Germany

were found significantly enriched in the 3'-UTR of protein-coding transcripts, a prominent segment embedding miRNA target sites. Following extraction of miRNP binding sites from the AGO-bound CLIP-Seq data, I investigated several potential processes affecting miRNA regulation in *cis*. In the end, I describe three occurring mechanisms mediated by *cis*-miR-SNPs: i) alteration of the miRNA seed pairing, ii) alternative 3'-UTR splicing leading to a loss of miRNP binding sites, and iii) change of the 3'-UTR fold. 53 SNPs of a total of 288 trait-associated 3'-UTR SNPs were annotated as mediating at least one of these mechanisms. The validity of these mechanisms was supported by an expression quantitative trait loci (eQTL) survey. Here, *cis*-miR-SNP induced allelic expression imbalance (AEI) was observed with a noticeable change in target expression variance.

Parts of this chapter were published in the journal *PLoS One*[1] in collaboration with Matthias Arnold[2], Mara Hartsperger[2], Arne Pfeufer[3], and Volker Stümpflen[2] [81].

Chapter 5 presents COGERE, a novel method for the computational modeling of miRNA-mediated gene regulatory networks (GRNs) in human and mouse. In contrast to the previous chapters, in this part of the thesis the large-scale modeling of miRNA-mediated regulation was addressed. The experimentalist is confronted with large data sets of high dimensionality reflecting the interplay of thousands of cellular components. Therefore, it is an imperative computational challenge to develop predictive and actionable models to investigate functionality as well as spatial and temporal behavior of these components. As the availability of experimental evidence in databases and the biomedical literature sharply increased, the systemic integration of existing knowledge to support the analysis of genome-wide molecular expression signatures of complex diseases becomes a bare requirement. Here, the elucidation of gene regulatory networks is a valuable source of hypothesis-driven clinical research. In this chapter, the novel approach COGERE is presented addressing the computational modeling of global miRNA-mediated regulation. I integrated existing information of regulatory interactions from multiple sources to a comprehensive prior model. At this, I implemented a data integration framework using, amongst others, information from AGO-bound CLIP-Seq data. Evaluation showed that the developed scoring scheme outperforms common integrative approaches. Further,

1 *PLoS One*, Public Library of Science
2 Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Germany
3 Institute of Genetic Medicine, European Academy Bozen/Bolzano (EURAC), Bolzano, Italy

COGERE is capable to infer condition-specific regulation. This is performed by evaluating the mutual dependency between regulator (transcription factor or miRNA) and target gene expression using prior information. This dependency is scored by the non-parametric, non-linear correlation coefficient η^2 (eta squared) which is derived by a two-way analysis of variance (ANOVA). Thus, COGERE implements a robust inference method together with a concept of high-level data integration. A comparative benchmark revealed that COGERE significantly improves alternative methods in predicting GRNs on simulated datasets. Furthermore, by inferring the cancer-specific GRNs from a cancer expression study, I demonstrate the utility of COGERE to promote hypothesis-driven clinical research. Since COGERE is a generalizable approach that boosts signal-to-noise for the modeling of large-scale condition-specific regulatory landscapes in any cellular contexts, the application was made public available [1] for academic research.

Parts of this chapter were published in the journal *Nucleic Acids Research*[2] in collaboration with Jörn Leonhardt [3], and Hans-Werner Mewes [3] [82].

In the final Chapter **Chapter 6**, I summarize and conclude the results presented in this thesis. Further, perspectives on potential future studies are discussed.

1 COGERE, http://mips.helmholtz-muenchen.de/cogere
2 *Nucleic Acids Research*, Oxford University Press
3 Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Germany

CHAPTER 2

The canonical set of miRNA seed types

The relation between miRNPs and their targets in higher eukaryotes is part of the highly complex gene regulation network. To unravel the interactions controlling gene regulation post-transcription, the available information is insufficient to reliably predict all functional pairs modulating translation and mRNA decay[83–85].

The basic prerequisite for miRNP binding in metazoans is a short perfect match to the coupled miRNA complemented by imperfect matches in close vicinity. This miRNA response element (MRE) region is called the 'seed' sequence and is considered to be a $6-8$ nt long substring within the first 8 nt at the 5'-end of the miRNA[26]. It is regarded to be the most important feature for target recognition by miRNAs in mammalians[27,28].

Naturally, merely seeking for short sequence matches yields a plethora of putative target sites containing a large fraction of false positives. To dodge *a priori* the majority of false positives, computational miRNA target site prediction approaches concentrate on the subset of target sites equipped with long perfect seed matches. In addition, several miRNA targeting determinants beyond the seed have been proposed to extract authentic target sites from the set of seed matches[49,50,86]. Although the evolution of miRNA targets is not well understood, a common strategy to increase specificity is to require conservation of the seed match. However, there is evidence that non-conserved miRNA targeting is even more widespread[87,88] and that miRNA:target interactions may play a role in the evolution of organismal diversity[89].

To date the effect of different types of seed matches has been assessed by means of signal-to-noise ratio[26,90], degree of mRNA[28,50] or protein repression[87,91]. Based on that, a set of canonical seed types that differ in abundance and intensity of the regulatory

effect has been defined[27]. However, for theses studies precise information on specific RISC binding regions was missing and due to experimental constraints a quantitative assessment was impractical. Now, recent experimental approaches allow for the detailed identification of AGO-miRNA:mRNA ternary complexes using an *in vivo* cross-linking protocol and subsequent high-throughput sequencing (Chapter 1.5.2). Chi *et al.*[71] analyzed miRNA:mRNA interactions in *Mus musculus* neocortex tissue samples and published an interaction map containing a set of verified target sites in the transcriptome of the murine brain; Hafner *et al.*[70] conducted a transcriptome-wide identification of target sites in human embryonic kidney cells.

In this chapter, previous studies in this field are complemented by determining canonical seed-pairing target site types using the AGO CLIP-Seq interaction maps. A minimal and sufficient set of six seed types was identified and their potential impact on target transcript stability was examined. Further, the precise mapping of AGO binding regions allowed to distinguish between miRNA:target and higher resolved miRNA:target site interaction during an evaluation of the seed feature for miRNA target prediction. At this, the impact of individual seed types on recall and specificity was quantified. Additional target site conservation analyses revealed that short seed-pairing sites are less conserved than long sites.

Major parts of this chapter have been previously published in the following article:

- **Ellwanger DC**, Büttner FA, Mewes HW, and Stümpflen V. The sufficient minimal set of miRNA seed types. *Bioinformatics*, 27(10):1346-50, 2011.

The results of this chapter have been presented at the following scientific conferences:

⋆ **Ellwanger DC**, Büttner FA, Mewes HW, and Stümpflen V. The sufficient minimal set of miRNA seed types. *German Conference on Bioinformatics* (Freising, Germany), 2011.

⋆ Büttner FA, **Ellwanger DC**, Mewes HW, and Stümpflen V. Large scale analysis reveals novel insights into the characteristics of miRNA targeting. *Lecture Notes in Informatics Edts.*, Schomburg D & Grote A (Braunschweig, Germany), 2010.

2.1 Material and Methods

2.1.1 Preparation of CLIP-Seq data

AGO HITS-CLIP

Chi *et al.*[71] provided a transcriptome-wide miRNA:mRNA interaction map in P13 mouse brains. It contains the absolute chromosomal positions of sites full complementary to miRNA seeds (murine genome assembly of 2006). These sites are located almost at the center of an average AGO-mRNA footprint. This is a defined region of mRNA complexed with AGO determined by AGO-mRNA clusters, where AGO bound within 62 nt of cluster peaks $\geq 95\%$ of the time. For each chromosomal coordinate, I determined the longest protein-coding mature mRNA transcript and its corresponding relative position by means of the NCBI reference sequence database[92]. Sites that were located within an intron (4%) or upstream of the 3'-UTR (45%) were removed. AGO HITS-CLIP included 20 miRNAs, whereas 18 of which are broadly conserved (according to Friedman *et al.*[57]). All analyses were conducted for the set of conserved miRNAs. All mRNA and miRNA data were downloaded from the UCSC Table Browser[93] and miRBase[15] on October 2010.

AGO PAR-CLIP

Hafner *et al.*[70] identified clusters formed by at least five PAR-CLIP sequence reads and more than 20% T to C transitions in human embryonic kidney (HEK293) cells. These 41 nt long regions were centered over the predominant cross-linking site. I mapped the chromosomal locations of 17 318 AGO1-4 CCRs to the longest protein-coding mature mRNA transcript based on the NCBI reference sequence database annotation[92]. All CCRs located within an exon of a mRNA 3'-UTR (37%) were retained. The dataset contained 580 miRNAs having at least one sequence read derived from AGO PAR-CLIP. For miRNA families having the same seed sequence (position $1-8$ at the 5'-end), the set was reduced to the member holding the highest sequence read count. Only broadly conserved miRNAs[57] were retained. All mRNA and miRNA data was obtained from the UCSC Table Browser[93] and miRBase[15] on January 2011.

I obtained the processed (background-corrected, adjusted for non-specific binding, and

quantile normalized with the GCRMA algorithm[70]) microarray measurements of HEK293 cells transfected with 2'-O-methyl-modified antisense oligoribonucleotides of the most highly expressed 27 miRNAs in the PAR-CLIP study (let-7a, miR-10a, miR-15a, miR-15b, miR-16, miR-17, miR-18a, miR-19a, miR-19b, miR-20a, miR-20b, miR-21, miR-25, miR-27a, miR-30a, miR-30b, miR-30c, miR-92b, miR-93, miR-101, miR-103, miR-106b, miR-186, miR-301a, miR-378, miR-7, miR-124) and microarrays of mock-transfected HEK293 cells from Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under series GSE21577. Platform probe identifiers (IDs) (Affymetrix Human Genome U133 Plus 2.0 Array) were mapped to GeneBank[94] accessions using GEO platform annotation GPL570. The log-intensity of probe sets mapping to the same gene were averaged to obtain the expression level per single transcript. The log fold-changes of transcript intensities were calculated as the ratio of the mean transcript expression in miRNA antisense-treated samples and mock-transfected cells.

2.1.2 Definition of functional binding sites

Based on the set of conserved miRNA sequences and mRNA 3'-UTR sequences, all sites complementary to a minimum of six contiguous nucleotides beginning at either position 1, 2 or 3 relative to the 5'-end of the miRNA were determined. The seed matches were classified by means of their distance to nucleotides found in AGO HITS-CLIP footprints. To account for all seed start positions, each seed match located within a distance of 2 nt to an AGO HITS-CLIP nucleotide was tagged functional. Since the reported positions in the AGO HITS-CLIP data were located almost at the center of a 62 nt long average AGO-mRNA footprint, matches found within a distance of $3-31$ nt could also be functional. Since the chromosomal coordinates of the footprints were not available, an unambiguous classification was not feasible. To avoid false positives, these sites remained unclassified. All seed matches located beyond the AGO-mRNA footprint (distance > 31), i.e. outside of a miRNP binding site, were classified as non-functional. Further, two miRNAs whose target sites were not significantly enriched (χ^2 test *P-value* (P) < 0.05) in the footprints were removed from the dataset. Finally, the instances were composed of 7 342 functional, 64 689 non-functional and 1 755 unclassified seed-pairing sites. Verifying a required minimum target site length of 6 nt, all 5mer matches were determined. The frequency of seed matches within a footprint (distance ≤ 31) and beyond of it was calculated for each

seed match length.

Accordingly, for the AGO PAR-CLIP data, seed matches were classified by means of their distance to the predominant CCR site. The seed match located within the CCR and nearest to its center was classified functional for each miRNA. Again, to avoid false positives, miRNA target sites found beyond the CCR center remained unclassified. Seed matches lying beyond the CCR, i.e. outside of a miRNP binding site, were classified non-functional. Further, only miRNAs whose target sites were significantly enriched (χ^2 test $P < 0.05$) in the CCRs were retained. Transcripts having only non-functional sites were removed from the dataset. Finally, the instances were composed of 21 214 functional, 380 893 non-functional and 665 unclassified seed-pairing sites for 72 miRNAs and 3 166 3'-UTRs.

2.1.3 Determination of seed types

The background set Ω was defined based on the functional and non-functional sites. A seed match $S_{p,k} \in \Omega$ was distinguished by its start position p relative to the miRNA 5'-end ($1 = \alpha$, $2 = \beta$, $3 = \gamma$) and its length k. Due to the hierarchical structure of Ω, the application of a separate-and-conquer strategy was feasible (Algorithm 2.1). First, the target sites were divided by their seed match start position. Thus, one got three supersets composed of seed matches of a minimum length of 6 nt containing all seed types: $S_{\alpha,6}$, $S_{\beta,6}$, $S_{\gamma,6}$. These sets were separated into 6mers having a mismatch at their subsequent position ($\overline{S_{\alpha,6}}$, $\overline{S_{\beta,6}}$, $\overline{S_{\gamma,6}}$) and seed matches having a minimum length of 7 nt, $S_{p,7}$. The null hypothesis was tested stating that the distribution of functional and non-functional target sites is independent of a mismatch at the 3' most subsequent position of a seed match. Thus, if the proportions of functional to non-functional target sites between the $S_{p,6}$ and the $\overline{S_{p,6}}$ seed types were not significantly varying, the separation terminated otherwise the procedure was continued for the next seed type length. The P was calculated by means of a two-tailed Fisher's exact test[95]. The outcome of this were 20 match types $\overline{S_{p,k}}$ with a corresponding P. The distributions of all seed match types were disjoint because each seed match was graded by the longest possible type.

For a significance level of 0.05, the α-seed site separation terminated after three steps, the β-seed matches contained two significant subsets and γ-yielded no significant subsets. The found significant seed types were termed based on their start position and their length:

$\overline{S_{p,k}}$ = "$k$mer$p$". For standardization, the endmost subsets were renamed: $S_{\alpha,8}$ = 8merα, $S_{\beta,7}$ = 7merβ, $S_{\gamma,6}$ = 6merγ.

To estimate the significance of the seed type set, the distribution of the functional sites was compared with a randomized pool of functional seed matches. By drawing without replacement, a subset of 7 803 instances of the multinomial distribution from functional and non-functional seed matches was created. The P was calculated by means of a χ^2 test of independence.

Algorithm 2.1: Find canonical seed types

Data: Start position type p, consecutive seed match length k, set of class-divided seed matches $\Omega = \Omega^+ \cup \Omega^-$

Result: Significant seed types

1 **begin**
2 $\quad \Sigma \leftarrow \emptyset$ ▷ Initialize accumulator
$\quad$ ▷ Identification of subsets (separation step)
3 $\quad S_{p,k} \leftarrow \{\forall s \in \Omega : \text{starttype}(s) = p \wedge \text{length}(s) \geq k\}$ ▷ Match at $k+1$
4 $\quad \overline{S_{p,k}} \leftarrow \{\forall s \in \Omega : \text{starttype}(s) = p \wedge \text{length}(s) = k\}$ ▷ Mismatch at $k+1$
5 $\quad m \leftarrow [|S_{p,k} \cap \Omega^+|, |S_{p,k} \cap \Omega^-|, |\overline{S_{p,k}} \cap \Omega^+|, |\overline{S_{p,k}} \cap \Omega^-|]$ ▷ Contingency table
6 $\quad P \leftarrow \text{FisherTest}(m)$ ▷ Fisher's exact test[95]
7 $\quad$ **if** $P > 0.05$ **then** ▷ Reject null hypothesis for a significance level of 0.05
8 $\quad\quad$ **return** $\Sigma \cup S_{p,k}$
9 $\quad$ **else** ▷ Continue recursively with subsets (conquer step)
10 $\quad\quad \Sigma \leftarrow \Sigma \cup \overline{S_{p,k}}$
11 $\quad\quad k \leftarrow k+1$
12 $\quad\quad$ **go to** 3
13 $\quad$ **end if**
14 **end**

2.1.4 Analysis of miRNA target site prediction

The impact of the seed types to miRNA target site prediction was evaluated in terms of recall, specificity and precision. The recall estimates how many of the functional target sites Ω^+ are covered by a certain seed type S, the specificity computes the fraction of

correctly excluded non-functional target sites and the precision denotes the relative amount of functional sites of a seed type.

Let

$$p_t = |\{s : s \in S \wedge s \in \Omega^+\}|,$$

$$p_f = |\{s : s \in S \wedge s \notin \Omega^+\}|,$$

$$n_t = |\{s : s \notin S \wedge s \notin \Omega^+\}|,$$

$$n_f = |\{s : s \notin S \wedge s \in \Omega^+\}|.$$

One can define:

$$\text{recall} = \frac{p_t}{p_t + n_f}, \qquad \text{specificity} = \frac{n_t}{n_t + p_f}, \qquad \text{precision} = \frac{p_t}{p_t + p_f}, \tag{2.1}$$

Further, an aggregate measure was computed (Matthews correlation coefficient, MCC):

$$\text{MCC} = \frac{p_t n_t - p_f n_f}{\sqrt{(p_t + p_f)(p_t + n_f)(n_t + p_f)(n_t + n_f)}} \tag{2.2}$$

The quality metrics of each miRNA target prediction algorithm were determined in terms of pure seed finding. Their seed type selection was assigned as described in the related literature. Due to ambiguous seed type assignments based on the first position of the target sequence, the evaluation of TargetScan[50] was performed by executing predictions on the mRNA set.

2.1.5 Seed type characterization

To estimate the miRNA seed type usage, the relative frequencies f_s of a seed type S for a certain miRNA was calculated. These values were normalized by the mean μ and the standard deviation σ:

$$Z(f_S) = \frac{f_S - \mu_{f_S}}{\sigma_{f_S}} \tag{2.3}$$

The conservation of each seed site was determined using the software package PHAST[96] as described by Betel *et al.*[97]. The included algorithm PhastCons is based on a phylogenetic hidden Markov model which is fitted to the input sequence by maximum likelihood. Each nucleotide gets a score measuring the evolutionary conservation across 17 vertebrates. For each seed match the absolute chromosomal coordinates were determined and a conservation score was calculated. Only if the score of each nucleotide within a functional seed match exceeded the threshold of 0.57[97], the site was tagged conserved in mammals. The background conservation of a seed type was computed by calculating the fraction of conserved nucleotides of a non-redundant set of 3'-UTRs holding a specific seed type.

2.2 Results

2.2.1 The canonical seed types of miRNA target recognition

In this study, a set of canonical seed types was defined by analyzing the seed matches of experimentally verified functional target sites in the 3'-UTR. The AGO HITS-CLIP miRNA:mRNA interaction map (murine assembly of 2006)[71] lists 15 665 chromosomal positions of target sites belonging to 20 miRNAs frequently bound in AGO complexes. These sites were mapped to annotated protein-coding mRNA transcripts and retained if they were located within the 3'-UTR, respectively. For each miRNA, the 3'-UTRs of the transcript set were scanned for all sites complementary to a miRNA subsequence beginning at either position one (α-position), two (β-position) or three (γ-position) relative to the miRNA 5'-end. At this, a minimum length of 6 nt was required. Seed matches of length five, as reported by Brennecke *et al.*[48], were not significantly enriched in average AGO footprints (Table 2.1). The sites were classified by means of their distance to an AGO HITS-CLIP binding site. After filtering of 16 broadly conserved miRNAs of which target sites were significantly enriched in AGO footprints, the set contained 2 369 murine genes with 7 070 Ago HITS-CLIP sites.

Each contiguous seed match was defined by its start position type and its length. The dataset was composed of eight α-, seven β- and five γ- seed match types (Figure 2.1). Following the principle of Occam's razor, the simplest seed type setting for target prediction should usually be the correct one. To reduce unnecessary complexity of the seed type set, unique seed types differing significantly from their superset in terms of functional

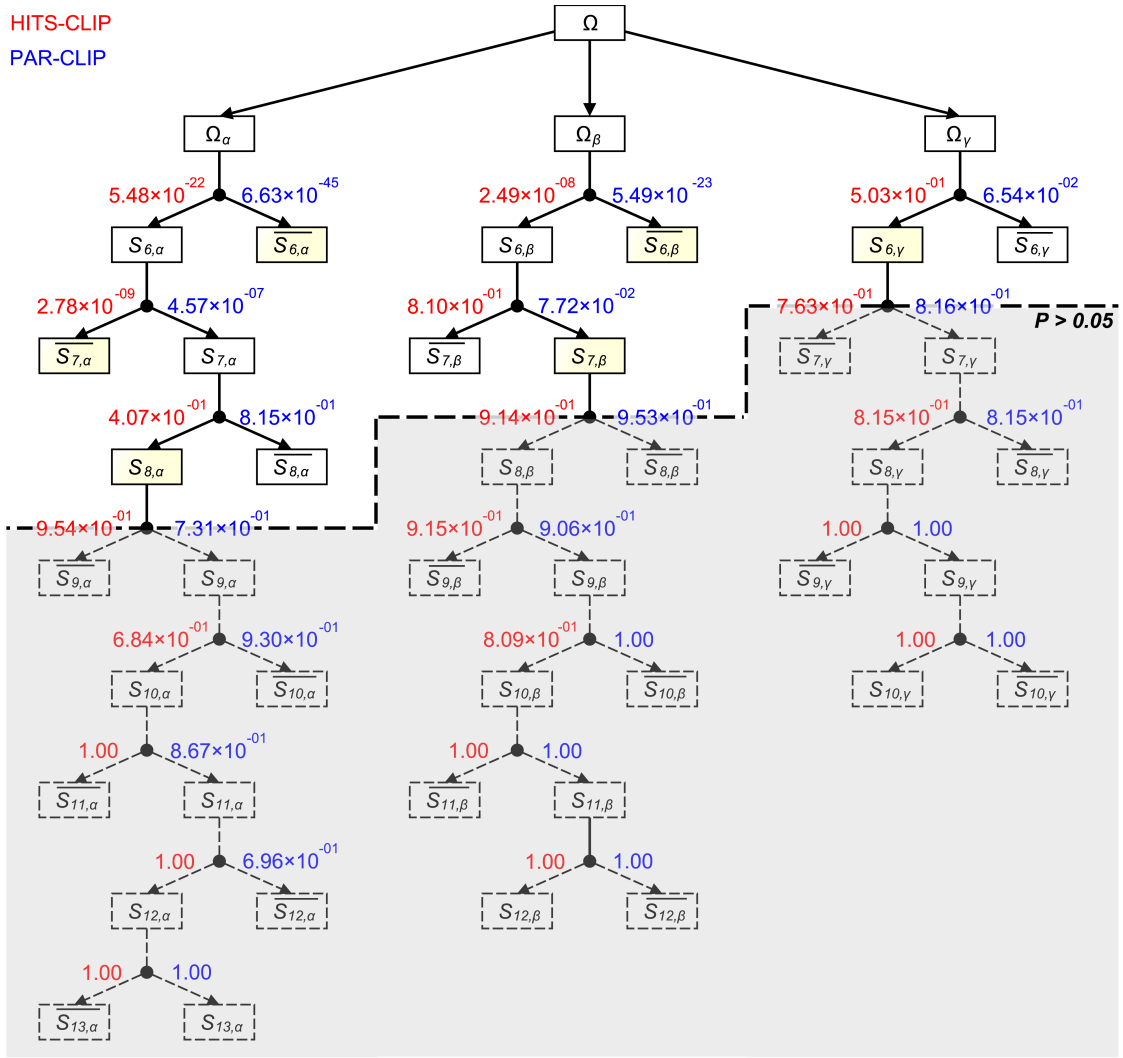


Figure 2.1 | **Determination of the sufficient minimal set of seed types.** The background set Ω is composed of functional and non-functional seed match sites. Each seed match $S_{p,k}$ is characterized by its start position p relative to the miRNA 5'-end and its minimum length k: $S_{p,k} \in \Omega_p$ with $p \in \{\alpha, \beta, \gamma\}$ and $k > 5$. In each step Ω_p can be further separated into a subset of seed matches with a length of at least k nt, $S_{p,k}$, and a subset of seed matches with a length of exactly k nt, $\overline{S_{p,k}}$. If the distribution of functional and non-functional target sites is independent of a mismatch at the 3' most subsequent position of a seed match ($P > 0.05$), the separation terminated otherwise the procedure was continued for the next seed type length. P are shown for each separation node (HITS-CLIP colored red, PAR-CLIP colored blue). The following seed types were identified: two seed types starting at position one ($\overline{S_{\alpha,6}}$, $\overline{S_{\alpha,7}}$, $S_{\alpha,8}$), two starting at position two ($\overline{S_{\beta,6}}$, $S_{\beta,7}$), and one starting at position three ($S_{\gamma,6}$).

Table 2.1 | **Enrichment of consecutive matching sites found in HITS-CLIP cluster peaks.**

Site length	Sites in peak	Sites out of peak	Log odds ratio	*P*
5	14 876	208 562	0.00	5.55×10^{-001}
6	6 239	54 346	0.21	5.92×10^{-289}
7	2 295	14 772	0.34	5.41×10^{-281}
8	948	3 963	0.53	2.39×10^{-279}
9	219	983	0.50	9.98×19^{-059}
10	45	242	0.42	7.03×10^{-010}
11	13	58	0.50	7.57×10^{-005}
12	2	16	0.25	4.44×10^{-001}
13	2	3	0.97	2.70×10^{-003}

and non-functional site distribution were identified. Six different, disjunct types of seeds were achieved: three 6mers either beginning at the first nucleotide (6merα), the second nucleotide (6merβ) or the third nucleotide (6merγ), two 7mers either starting at position one (7merα) or position two (7merβ) and one 8mer beginning at the first nucleotide (8merα). These canonical seed types terminated within the first 8 nt of the miRNA in 97% of cases. This underscores the importance of the octamer at the miRNA 5'-end. A fact which can be motivated by the AGO2 protein structure. The first 10 nt of the miRNA are located at the RNA binding groove (Figure 1.3) and prearranged in a geometry resembling an A-form helix[38]. The results suggest that the accessibility of a preformed helical segment longer than about 8 nt would not increase the effective nucleation surface. This may be reasoned by the fact that additional nucleotides would face opposing directions inducing topological challenges[27].

The significance of this seed type set was evaluated by a sampling approach. The log odds ratio of long seed types is above zero, pointing to a better discrimination between functional and non-functional sites (Table 2.2). Further, to exclude that the inferred seed type set is affected by an experimental bias, human AGO PAR-CLIP data[70] composed of 21 214 functional, 380 893 non-functional seed-pairing sites for 72 miRNAs and 3 166 3'-UTRs was used to validate the observation. By applying the presented separate-and-conquer algorithm, the identical seed type set was identified (Figure 2.1 and 2.2A).

Next, the destabilization effect of miRNA binding to a specific seed type was char-

Table 2.2 | **Determined canonical seed types.**

Seed type	Functional	%	Non-functional	%	LOR[a]	*P*
6merα	1 793	24	20 746	32	−0.12	1.20×10^{-028}
6merβ	1 382	19	13 500	21	−0.04	2.57×10^{-004}
6merγ	1 755	24	17 954	28	−0.06	2.26×10^{-009}
7merα	760	10	5 036	8	0.12	2.03×10^{-013}
7merβ	959	13	5 250	8	0.21	1.34×10^{-042}
8merα	693	9	2 203	3	0.44	7.60×10^{-132}

[a]Log odds ratio based on sampling.

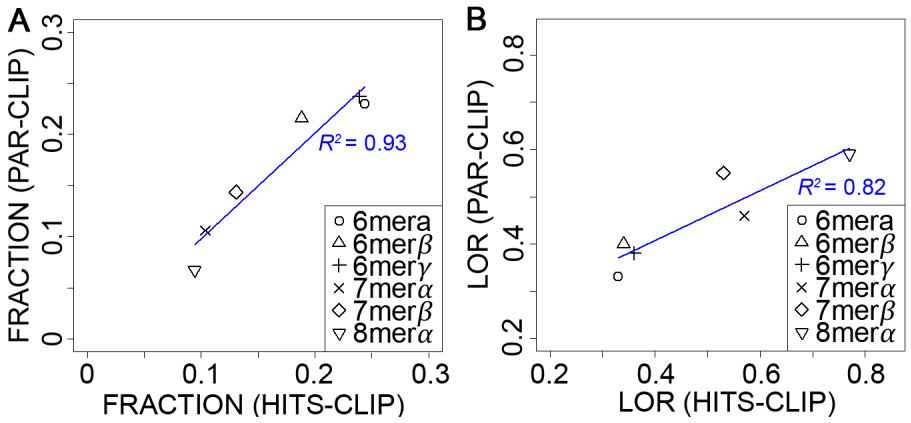


Figure 2.2 | **Correlation of HITS-CLIP and PAR-CLIP. A** | The seed type distribution of functional sites is equal in both datasets (F-test $P = 2.1 \times 10^{-3}$). **B** | The log odds ratio (LOR) was computed of finding functional sites in conserved regions. The log odds ratio (LOR) is equal in human and mouse (F-test $P = 1.3 \times 10^{-2}$).

acterized. For this purpose, transcriptome-wide expression data of embryonic kidney cells following transfection with antisense oligoribonucleotides of the most abundant 27 miRNAs in the PAR-CLIP study[70] was examined. Figure 2.3A shows that the stability of transcripts which contain a functional target site characterized by any of the six seed types was significantly increased post-transfection compared to transcripts without a functional seed match (Bonferroni corrected Wilcoxon rank sum test $P < 10^{-04}$). One can also observe an increasing order of regulatory effectivity: from 6merγ (lowest), 6merα, 6merβ, 7merα, 7merβ, to 8merα (highest). Consistent with previous studies[50] the repressive effect of the miRNA depends on the length of the seed-complementary region and rises

clearly from 6mer to 7mer to 8mer matches (Figure 2.3B).

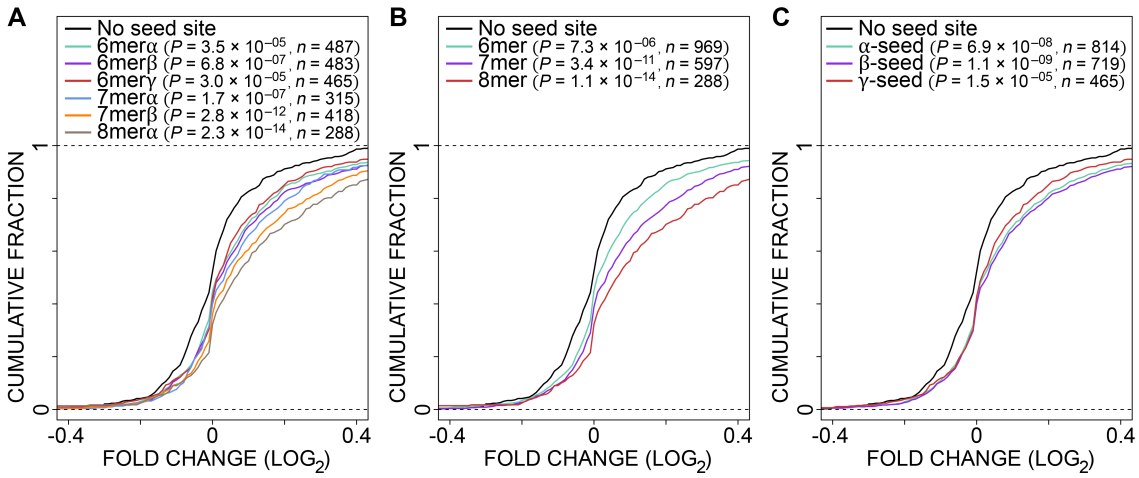


Figure 2.3 | **Effectiveness of canonical sites.** Transcripts were categorized according to the presence of a functional seed matching site of any of the 27 most abundant miRNAs in the PAR-CLIP study. The distribution of the expression fold-change of target transcripts following anti-miRNA transfection are shown for these categories: the magnitude of destabilization effects of transcripts containing a specific functional canonical site (**A**), the effectiveness based on either the seed match length (**B**) or the seed match starting position (**C**). *P* are given by the Wilcoxon rank sum test and indicate significant differences between the expression level changes of *n* transcripts with a functional target site versus 386 transcripts without a functional seed complementary site; shown values were adjusted for multiple-testing using the Bonferroni correction.

In a previous work, Bartel[27] defined seeds of miRNA target recognition. The AGO CLIP-Seq derived set of canonical seed type recovers the previous definition and extends it by additional seed types starting at the α-position (Figure 2.4). Interestingly, the major fraction of functional target sites is complementary to the very 5'-terminal nucleotide of the miRNA seed sequence (HITS-CLIP 44%, PAR-CLIP 40%). The set of miRNAs whose target sites were significantly enriched in AGO binding sites have a strong bias towards a uracil at their first position (χ^2 test $P < 1.1 \times 10^{-03}$). This observation has also been previously stated elsewhere[98–100] and was suggested to be justified by the AGO2 protein structure. Backbone atomes of a rigid loop in the middle domain of the AGO2 peptide chain exhibit a higher affinity for the base of uracil monophosphate (UMP) than for the base of adenosine monophosphate (two-fold lower), guanosine monophosphate (28-fold lower) and cytosine monophosphate (30-fold lower than UMP)[39]. Thus, mature miRNA

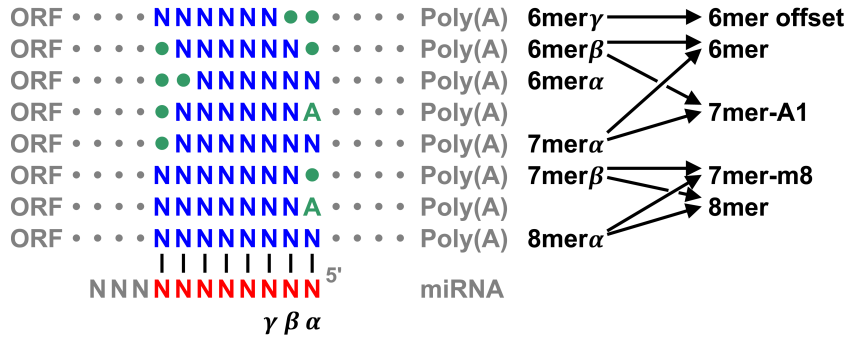


Figure 2.4 | **Definition of seed types.** The seed types were termed by the start position relative to the 5'-end of the miRNA and the length of the consecutive seed match. The defined set of canonical seed types can be surjectively projected to the seed type set of Bartel[27]. Equivalent definitions could be found for 6merβ, 7merβ and 6merγ. In the case of miRNAs having a seed sequence beginning with a uracil, 7merα complies with 7mer-A1 and 8merα is equal to 8mer. Otherwise 6merβ equates 7mer-A1 and 7merβ complies with 8mer. If the first position within the target sequence is not an adenine, 8merα equates 7mer-m8 and 7merα is equal to 6mer. Additionally, the set considered 6mer matches that are complementary to the first position of a miRNA seed (6merα). Core seeds can be found from position 1 to 6 (covered by 6merα, 7merα, and 8merα), 2 to 7 (covered by 6merβ, 7merα, 7merβ and 8merα) and 3 to 8 (covered by 6merγ, 7merβ and 8merα) of the miRNA sequence.

sequences starting with a uracil may be preferentially integrated into the RISC. Further, Lewis *et al.*[90] reported that the majority of conserved target sites exhibits a 3'-terminal adenine. They assumed that this so-called 'A anchor' is recognized simultaneously or sequentially to the interaction with the first nucleotide of the miRNA by a protein contained in the RISC. Resultant, as uracil binds to adenine via two hydrogen bonds, α-seed types can be indeed expected to be frequently observed. This raises the question how Watson-Crick pairing at the very 3'-terminus of the miRNA complementary site affects mRNA:miRNP complexing and to which extent target cleavage. The transfection data provides no evidence for a differential effectiveness between α- or β-paired seed regions (Figure 2.3C).

Previous studies defined the miRNA seed match starting at position two and requiring a length of at least 6 nt as miRNA core seed[27,50,57], i.e. a paired region covered by multiple seed types. In the canonical set, it is covered by the seed types 6merβ, 7merα, 7merβ and 8merα. In addition, two further core seeds can be identified: the former is ranging from position one to six covered by 6merα, 7merα and 8merα; the latter is ranging from position three to eight covered by 6merγ, 7merβ and 8merα.

2.2.2 Majority of functional sites are based on 6mer seeds

The effect of each seed type to recall and specificity was examined (Figure 2.5A). Focusing on the relative contribution of each seed type to functional sites, 6mer seeds make up the highest fraction of true target sites (recall = 0.67). On the other hand, 6mer types involve many false positives leading in sum to a low specificity (0.19) and precision (0.09). In terms of computational target site classification, the usage of a short seed type causes an inverse prediction (MCC < 0, Figure 2.5C), suggesting the avoidance of such a type. In this case, reversing the classification would yield a result superior to an average random prediction.

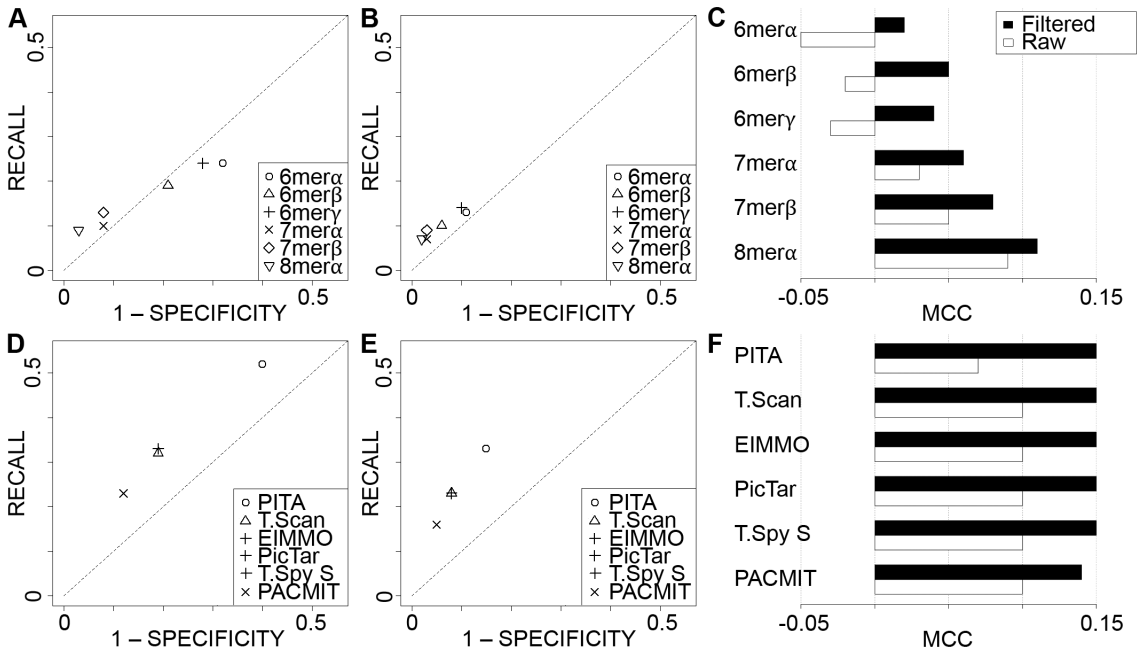


Figure 2.5 | **Accuracy evaluation. A**, **D** | The impact of each seed type on miRNA target site prediction was determined by means of recall and specificity. The effect of the (default) seed type selection is shown for several prediction algorithms. These values present the minimum specificity and the maximum recall of the six tools, respectively. **B**, **E** | Removing non-conserved target sites increases the specificity and the precision, but lowers the recall. **C**, **F** | MCC values for predicting sites with and without filtering for conserved sites. Note that panels **D**, **E** and **F** do not reflect the ranking of predictions based on the algorithms' scoring schemes; *T.scan* = TargetScan, *T.spy S* = TargetSpy Seed. The dashed line illustrates an average random prediction.

Barely one-third of all genuine target sites are covered by seeds of length 7 and 8. Among these seed types, 7merβ holds the highest recall (0.13) and 8merα shows the best specificity (0.97). The combined set of 7- and 8mer matches achieves a specificity of 0.8 (precision $= 0.19$).

Evaluation on the miRNA:mRNA interaction level resulted in an increased recall and specificity for each seed type (Figure 2.6). This evaluation level is more general, as only the presence of a site on a mRNA matters. Multiple matches of one miRNA on a target mRNA are combined into one miRNA:mRNA interaction. In contrast, the miRNA target site determination evaluation takes the location of a seed match relative to an AGO footprint into account. Consequently, evaluation on the more general level implies that multiple false positive seed matches may be combined to one true positive miRNA:3'-UTR interaction. Conversely, multiple true negative target sites may be combined to one false positive interaction.

The majority of functional sites are formed by short seed-pairing sites. These 6mers were found to be associated with low repressive effects. Three regulatory roles have been proposed for miRNAs in the literature: switches, fine-tuners, or natural targets (Chapter 1.4.3). In this study, 6mer seeds were found to be associated with low repressive effects. It can be suggested that marginal reduction, i.e. fine-tuning, of the mRNA level may be the predominant effect of global miRNA-mediated regulation. Also, short seed-complementary sites may likely play a major role in the miRNA decoy mechanism (Chapter 6).

The importance of short seed types gains further support by the observation that 37% of the 3'-UTRs contain exclusively seed matches of length six in their AGO footprints. Interestingly, the sequences of this subset of 3'-UTRs are significantly shorter than those of the superset (t-test $P = 4.53 \times 10^{-06}$). Stark *et al.*[101] studied the impact of miRNA regulation on 3'-UTR evolution and found that short 3'-UTRs indicate avoidance of miRNA regulation. This goes well with the observation that short 3'-UTRs are regulated by less effective 6mer matches.

Next, the question was addressed, whether miRNAs exhibit seed type propensities. The relative frequencies of the seed types were computed for each miRNA. A Z-score indicates miRNAs holding a frequency over or below the mean frequency given a specific seed type (Figure 2.7). It was observed that 6mer seed types and long seed types are grouped to clusters, respectively. Further, two main miRNA cluster appeared. The larger group contains miRNAs binding primarily to 6mer-based functional sites. Seven of the

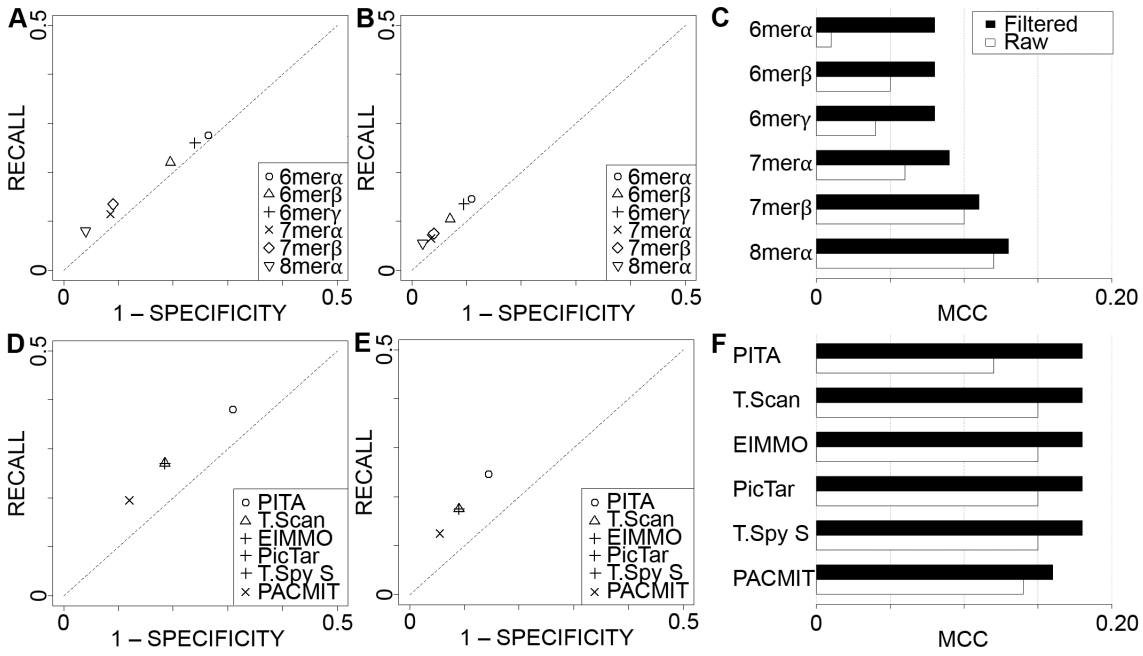


Figure 2.6 | **Accuracy evaluation of miRNA:mRNA interaction determination. A**, **D** | The contribution of a seed type to a miRNA:mRNA interaction was measured by the receiver operating characteristic (recall vs. 1− specificity). The corresponding minimum specificity and maximum recall values for a set of six miRNA target prediction algorithms were determined. **B**, **E** | The effect on accuracy by retaining only conserved sites was computed. **C**, **F** | MCC values for predicting miRNA:mRNA interactions with and without filtering for conserved sites. Note that panels **D**, **E** and **F** do not reflect the ranking of predictions based on the algorithms' scoring schemes; *T.scan* = TargetScan, *T.spy S* = TargetSpy Seed. The dashed line illustrates an average random prediction.

16 miRNAs carry out stronger repression by pairing to rather long seed matches. These results suggest that each miRNA likely has a transcriptome-specific bias towards long or short seed-binding sites.

2.2.3 Non-conserved targeting relies on short seeds

The strategy established by Betel *et al.*[97] was used to identify seed-pairing sites conserved across mammals (Figure 2.8). The majority of functional target sites is conserved (60%). All seed types have a higher fraction of conserved sites than one would expect by chance, given the conservation of their 3'-UTRs (6merα: log odds ratio $= 0.33$, $P = 1.23 \times 10^{-61}$; 6merβ: log odds ratio $= 0.34$, $P = 4.07 \times 10^{-50}$; 6merγ: log odds ratio $= 0.36$, $P =$

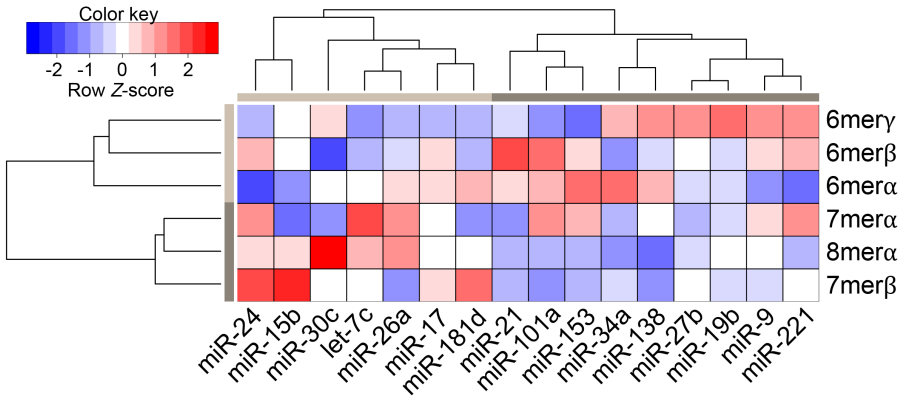


Figure 2.7 | **Seed type distribution for each miRNA.** The colors affected by the row Z-score indicate the propensity of miRNAs to bind specific seed-complementary sites in the murine neocortex transcriptome. A red/blue coloration implies a higher/lower usage of a seed type compared to other miRNAs.

2.53×10^{-71}; 7merα: log odds ratio $= 0.57$, $P = 2.75 \times 10^{-72}$; 7merβ: log odds ratio $= 0.53$, $P = 3.33 \times 10^{-82}$; 8merα: log odds ratio $= 0.77$, $P = 1.12 \times 10^{-108}$). This observation was validated with the human AGO PAR-CLIP data (Figure 2.2B).

The 6mer sites reveal an almost equal partitioning in conserved and non-conserved sites. A clear discrepancy between the numbers of conserved and non-conserved sites emerges for 7- and 8mer seeds. Particularly, 8merα seed matches exhibit a significant tendency to be conserved. The number of conserved sites in this case is more than three times as high as the number of non-conserved sites. In terms of 7mer seeds, about two-thirds of the seed matches are conserved, whereat 7merα exceeds 7merβ. Retaining only conserved seed matches lifts target prediction specificity of all seed types (Figure 2.5B). In particular, the 6mer seeds show a significant increase in specificity resulting in a classification better than an average random guess (MCC > 0, Figure 2.5C).

In summary, the mean probability to be conserved is about 55% for a 6mer seed. In contrast, 7mer and 8mer seeds have a probability of up to 77% to be conserved. Further, a total of 75% of the functional non-conserved sites are covered by 6mer seeds. Therefore, lineage-specific miRNA regulation relies to a large extent on target sites containing short seed-pairing sites.

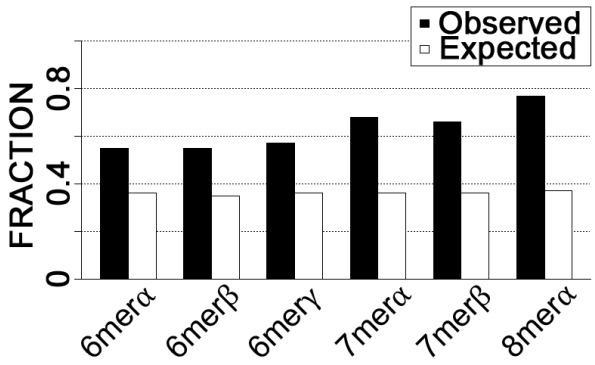


Figure 2.8 | **Conservation of seedtypes.** Observed and expected fraction of conserved functional seed matches for each seed type.

2.2.4 Target prediction focuses on 7- and 8mer seed matches

Frequently used approaches for target prediction in mammals were reviewed with regard to the implemented seed types (Table 2.3). The TargetScan algorithm[50] seeks mainly for seeds of length seven and eight via the seed types 7mer-A1, 7mer-m8 and 8mer. The 7mer-A1 sites may be of type 6merβ in the event that the miRNA sequence starts with a nucleotide different to uracil. However, the majority of mammalian miRNAs begins with a uracil[90]. Both PicTar[102] and EIMMO[103] require stringent seed pairing of $7-8$ nt starting at either the α or the β-position. A novel approach called TargetSpy (with seed setting)[104] restricts target predictions to transcripts encoding a perfect 7mer site.

Table 2.3 | **Default miRNA seed type selection of prediction algorithms.**

Algorithm	**6merα**	**6merβ**	**6merγ**	**7merα**	**7merβ**	**8merα**
PITA[a]		✓		✓	✓	✓
TargetScan[b]		✓[b]		✓[c]	✓	✓
PicTar				✓	✓	✓
EIMMO				✓	✓	✓
TargetSpy S				✓	✓	✓
PACMIT[a]					✓	✓

[a]Configurable seed length; default seed types ensure high precision.
[b]If miRNA seed sequence starts with an adenine, guanine, or cytosine.
[c]If miRNA seed sequence starts with a uracil.

Some algorithms allow for custom-defined seed searching: PITA[49] seeks by default for sites of length six, seven and eight that start at position two of the miRNA. The standard setting of PACMIT[105] is even more restrictive by considering merely sites matching to miRNA positions $2-8$. Both tools enable the adjustment of the site length by the user. RNAhybrid[106] as well as IntaRNA[107] are more flexible by providing a couple of additional parameters to customize the seed search, e.g. a user-defined setting of the start position. Notably, IntaRNA is a general approach to predict any RNA:RNA interactions. Both tools do not suggest default seed search parameters.

The impact of the (default) seed type selection of prediction algorithms on recall and specificity was evaluated. Prediction methods implement scoring schemes to value target site characteristics beside the seed. In contrast to common evaluation frameworks, the assessment presented in this work is not focusing on a subset of top scored instances, but uses all predictions. Therefore, the denoted specificity values represent the minima while the recall values show the maxima for the (default) seed choice, respectively. It should be noted that subsets composed of top scored predictions would achieve significantly higher specificity values.

It was observed that all prediction models exhibit a considerable constraint regarding their ability of finding potential target sites (Figure 2.5D and F). PITA holds the highest recall of 52% (specificity $= 60\%$) owing to the exhaustive search for 6merβ seed matches, whereas PACMIT has the lowest recall of 23% (specificity $= 88\%$). Remarkably, this tool is restricted to find less than a quarter of all functional seed-pairing sites. Additional filtering by removing conserved sites increases the specificity but consequently lowers the recall (Figure 2.5E). Here, PACMIT can only find 16% of all functional sites (specificity $= 73\%$). A higher recall but a lower specificity can be observed for the prediction of miRNA:mRNA interactions (Figure 2.6D, E, and F). Concluding, due to the significant gain of precision, tool developers prefer to use long seeds. This study quantified the loss of recall accompanied by this proceeding.

2.3 Conclusion

This study presented an analysis of the most important feature for miRNA target recognition, the so-called miRNA seed, using a large-scale dataset of functional target sites. Based

on the AGO HITS-CLIP and AGO PAR-CLIP miRNA:mRNA interaction maps, seed properties and their influences on miRNA target site prediction methods were analyzed. Due to the definite specification of AGO binding sites, the classification of MREs contained in the mRNA 3'-UTR as either functional or non-functional was feasible. A minimal set of seed types that is sufficient for accurate miRNA target site predictions was defined and its effect on transcript stability was examined. The data pool presented in this work allowed for enhanced analysis of miRNA target prediction algorithms compared to earlier studies that were restricted by experimental constraints (e.g. Alexiou *et al.* [108]; Selbach *et al.* [91]).

It was found that most conserved miRNAs interact predominantly with target sites endowed with short seed matches; 67% of functional sites are based on 6mer seeds. The common assumption that short seed matches are associated with low effects on target mRNA stability [57] was recovered. From this observation it was suggested that the predominant effect of global miRNA-mediated regulation is a marginal reduction of the mRNA level. This is consistent with the commonly accepted mode of action of miRNA regulation: these short regulatory ncRNAs were suggested to be responsible for fine regulation of target transcript abundance to adapt the cellular phenotype during crucial processes such as cell development and differentiation [109].

In terms of maintaining target predictions with a low false-positive rate, the common approach of current algorithms to focus mainly on seeds of length seven or eight was reconfirmed. At present, prediction algorithms have to accept severe deficiencies of recall to ensure high specificity that is naturally considered to be more important. In addition, such a restriction denotes a disregard of the majority of fine-tuned miRNA targets.

It was observed that the preferential search for long seeds lifts the proportion of conserved sites. However, a substantial fraction (40%) of all functional target sites is not conserved across mammals; 6mer complementary seed sites are enriched among these. It remains to be elucidated, how the conservation constraints for each seed type vary between close and more distant related species. Recently, Xu *et al.* [89] described that, in general, seed-based target sites are more conserved between closely related species, such as primates, but less conserved for distantly related species, such as birds and mammals. Although most target sites were suggested to be under, at least marginal, evolutionary constraints, several seed-sites are conserved only for a short evolutionary period. Based on the observation of this chapter, one can suggest that a fraction of 6mers evolved from 5mers rather from long seed-pairing sites that are assumed to be under positive selection. Since

one miRNP binding site embeds multiple seed-pairing sites, the evolutionary emergence of *bona fide* 'sponging' 6mer motifs is one probable scenario (Chapter 6). However, the current knowledge about miRNA target site evolution is limited − further investigations are required to clarify this question.

Concluding, omitting short seed-pairing sites and requiring target site conservation results in a lowered recall of current target prediction algorithms. Since the fraction of spurious matches is very high for this kind of seed-pairing sites, the problem of recall can be easily translated to a problem of precision. This strongly intensifies the need for features beyond seed pairing that realistically describe miRNA targeting, in particular non-conserved target sites. It may also raise the basic question for the potential of seed-based approaches in discriminating between functional and non-functional sites.

CHAPTER 3

miRNP features beyond miRNA seed pairing

After maturation in the cytoplasm, miRNAs are incorporated into the miRNP and function as primer for partially complementary base pairing mostly to the 3'-UTR of the target mRNA[52]. Early studies on target recognition revealed that Watson-Crick pairing between the target sequence and the 5'-end of the miRNA is a primary determinant of target specificity[110]. As discussed in Chapter 2, such a seed match by itself is a poor predictor due to high stochastic noise caused by the high number of random occurrences of any given 6mer, 7mer or 8mer motif in a 3'-UTR. Thus, for reliable target determination additional features beyond seed pairing are required.

However, the comprehension of the molecular basis of the miRNA:target pairing process is limited. Since the experimental detection is a costly and time-consuming process, the current knowledge about the exact location of miRNA target sites is limited and disproportional to the number of known miRNAs. A common database with target site-related information is miRecords[111]. The most recent release (April 27, 2013) contains 733 interactions of 162 miRNAs and 297 genes in human. In contrast, the miRBase (release 20, June 2013) sequence database lists 2 576 human mature miRNA transcripts. Considering that it has been estimated that miRNAs regulate hundreds of targets via multiple target sites[90], it is obvious that the reported number of verified information accounts only for a small fraction of the actual extent of miRNA targeting. This fact emphasizes the urgent need for computational miRNA target site prediction methods to guide wet lab experiments and, in the end, to facilitate the transcriptome-wide discovery of operative miRNA-mediated regulation.

Considerable advances have been made in *ab initio* target prediction[27,112]. Several

algorithms were developed implementing additional target site features, such as evolutionary conservation[102], structural accessibility[49,113,114], local nucleotide composition, and target site location[50]. All tools require an initial search for anchor sequences such as seed complementary sites. Subsequently they are evaluating the site by their respective feature scoring schemes. For the most part, all of the them generate a bulge of predictions − many of which are presumed to be false positives[65]. Thus, the prediction of reliable miRNA target sites is still an unsolved computational challenge.

Over the recent years, significant efforts have been made in the experimental high-throughput screening of biologically relevant miRNA:target interactions. By using cross-linking and AGO immunoprecipitation coupled with high-throughput sequencing (CLIP-Seq) in cells of interest, the binding regions of the miRNP can be reliably determined. This is a clear advantage to previous experimental approaches such as mRNA expression profiling and proteomics[85,87,91] as this technique allow the direct identification of a huge pool of short target sequences representing miRNP binding regions (Chapter 1.5.2). It has been suggested that the false-positive rate of prediction algorithms can be significantly reduced by restricting the search space of miRNA target sites *a priori*[70,71,115]. However, AGO CLIP-Seq libraries are limited to highly expressed transcripts. Further, the results are condition-specific, i.e. they depend on the environment, organism, tissue, and cell cycle state. Thus, the computational characterization and genome-wide prediction of miRNP binding sites is of particular interest.

This chapter analyzes features of miRNP binding sites beyond seed pairing. For thus purpose, an elaborate data pool of negative and positive instances from two AGO-bound CLIP-Seq libraries was prepared. Several features were collected from literature which were suitable for miRNP binding site prediction. This set was composed of attributes established for miRNA target prediction as well as relevant characteristics from other fields of RNA analysis. Following feature extraction, a SVM-based classifier was trained. The classifier ranks segments on target transcripts for their miRNP binding affinity. It is independent of any miRNA sequence and as such can be combined with a subsequent miRNA target site search, such as a naïve seed matching or any sophisticated target prediction tool. By using miRNA transfection data, it was shown that filtering of target sequences by predicted AGO-bound regions increases the precision of common target prediction algorithms. Further, the approach was applied to examine the hypothesis that the pro-fibrotic mediator WISP1 escapes post-transcriptional regulation by miRNAs in pulmonary fibrosis.

At this, miR-92a was identified as potential regulator. Subsequent experiments confirmed the miR-92a-mediated regulation of WISP1 in lung fibroblasts and lung tissue specimens of affected patients. Further, it was found that TGF-β1-induced WISP1 expression can be altered by modulation of miR-92a in human primary lung fibroblasts (pFBs).

Parts of this chapter have been previously published in the following article:

- Berschneider B, **Ellwanger DC**, Shimbori C, White ES, Kolb M, Neth P, and Königshoff M. miR-92a regulates TGF-β1 induced WISP1 expression in pulmonary fibrosis. *Int J Biochem Cell Biol.*, 53:432-41, 2014.

The results of this chapter have been presented at the following scientific conferences:

⋆ Berschneider B, **Ellwanger DC**, Mewes HW, Neth P, Königshoff M. Regulation of Wnt1-inducible signaling pathway protein 1 by miRNAs in pulmonary fibroblasts. *Am J Respir Crit Care Med*, 187 A6060 (Philadelphia, USA), 2013.

⋆ Berschneider B, **Ellwanger DC**, Thiel C, Stümpflen V, and Königshoff M. microRNA regulation of WISP1 in pulmonary fibrosis: an *in silico* approach. *Pneumologie*, 65 - A6 (Homburg/Saar, Germany), 2011.

3.1 Material and Methods

3.1.1 Processing of CLIP-Seq data

Kishore *et al.*[116] examined the effect of metabolic labeling with photoreactive nucleosides, cross-linking at distinct wavelengths and the use of different ribonucleases on the recovery of AGO2 binding-sites by PAR-CLIP and HITS-CLIP. They prepared libraries of AGO2 CCRs based on both experimental protocols in human embryonic kidney (HEK293) cells. Subsequent statistical analyses of the enrichment of binding sites and the frequency of various types of mutations within these sites allowed an accurate extraction. The complete RNase T1 treated samples were obtained from Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under series GSE28865. It has been described that the 3'-UTR represents the major miRNA binding region with the highest impact on target transcript stability[27,117]. Therefore, the CCRs were mapped to the longest mature 3'-UTR transcript of each gene based on the Ensembl[118] genes annotation (assembly GRCh37) resulting in 5 701 CCRs for PAR-CLIP and 2 407 CCRs for HITS-CLIP.

3.1.2 Feature extraction

To extract the features of an AGO2 CCR, a sliding window approach was applied. A window length of $n = 41$ nt was selected. Segments of this length were suggested to represent the central miRNP binding segment[70,116]. Starting at the first position of a 3'-UTR and moving by 1 nt, the following feature vector $x = (x_1, ..., x_{11})^T$ was computed for the central nucleotide.

Conservation score

Each nucleotide of a 3'-UTR was rated by a conservation score from the PhastCons[96] and PhyloP 46way[119] track provided by the UCSC Table Browser[93]. The central nucleotide of a window was assigned two scores by computing the total conservation of a window segment using either algorithm.

Local base content

The local content of adenine (A) and uracil (U)[50], the content of adenine and guanine (G)[120] as well as the uracil content[70] within a window was scored by a weighting function tailing off the distance from the central nucleotide[50]. The presence of the respective nucleotide increased the score for the site:

$$x_i = \sum_{d=1}^{n} \frac{s_d}{|\lceil 0.5n \rceil - d| + 1}, \text{ with } s_d = \begin{cases} 1 & \text{if nuc}(d) \in K \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

The function nuc(d) returns the nucleotide on position d of the sequence window. Three feature scores were computed with $K = \{\text{A, U}\}$, $K = \{\text{A, G}\}$, and $K = \{\text{U}\}$.

Base asymmetry bias

The basic asymmetry signals (skews) were calculated. These were proposed as evidence for the potential formation of strand-specific RNA structures between A and U as well as G and cytosine (C)[121]:

$$x_i = \frac{f_k - \overline{f_k}}{f_k + \overline{f_k}} \tag{3.2}$$

Here, f_k denotes the relative fraction of nucleotide k and $\overline{f_k}$ denotes the relative fraction of the nucleotide complementary to nucleotide k in the sequence window. Two scores were computed with $k = \text{A}$ (AT skew), and $k = \text{G}$ (GC skew).

Compositional entropy

The base compositional entropy of a sequence window was scored by[120]:

$$x_i = -\sum_{k \in \{A,T,G,C\}} f_k \log_2(f_k) \tag{3.3}$$

where f_k denotes the relative fraction of the nucleotide k in the window sequence.

Relative position

The relative distance of the central nucleotide to the center of the 3'-UTR was calculated[50], i.e. the higher the relative position score $\in [0,1]$, the closer was a window located to one of the 3'-UTR ends.

Structural accessibility

More accessible segments were proposed to be more effective target sites[49]. The Raccess algorithm[122] was applied to compute the structural accessibility of each nucleotide of a 3'-UTR. A minimum length of 20 nt was forced to be accessible. The maximal span of base pairs considered for pairing was restricted to 400 nt. The central nucleotide of a sequence window was assigned the total energy required to unfold the window segment.

miRNA pattern aggregration

It has been described that regions encoding miRNA target sites accumulate patterns composed of reverse complements of mature miRNAs[104,113]. To avoid a species specific bias, patterns were mined from the set of 687 broadly conserved miRNAs provided by Grimson *et al.*[50]. Identical and near-duplicate sequences were removed by computing optimal global pair-wise alignments using the dynamic programming method proposed by Needleman and Wunsch (NW)[123] (Algorithm 3.1). This resulted in a subset R of 236 sequences with less than 73% identity. All variable-length motifs in the reverse complement of R were discovered with the Teiresias[124] algorithm. The patterns were defined analogous to Miranda *et al.*[113]: minimum length of $L = 4$ nt and at least 30% of their positions had to be specified ($W = 12$). In addition, it was required that a pattern has to occur at least $K = 3$ times. Then, a second-order Markov chain was used to estimate the log probability of the 397 737 patterns to occur by chance on human Ensembl[118] 3'-UTR sequences. For this purpose, the occurrences of all tri-nucleotides f separated by any number of wild-cards, i.e. any nucleotide (denoted by '$\cdot$'), were computed. Then, the probability that a pattern, e.g. $\mathrm{UC}\cdot\mathrm{A}\cdot\mathrm{C}\cdots\mathrm{G}$, is generated from a random database was computed using the Bayes' theorem, e.g. $P(\mathrm{UC}\cdot\mathrm{A}\cdot\mathrm{C}\cdots\mathrm{G}) = P(\mathrm{UC}\cdot\mathrm{A}|\mathrm{UC})P(\mathrm{C}\cdot\mathrm{A}\cdot\mathrm{C}|\mathrm{C}\cdot\mathrm{A})P(\mathrm{A}\cdot\mathrm{C}\cdots\mathrm{G}|\mathrm{A}\cdot\mathrm{C})$. The probabilities were directly inferred from f, e.g. $P(\mathrm{UC}\cdot\mathrm{A}|\mathrm{UC}) = f_{\mathrm{UC}\cdot\mathrm{A}}/(f_{\mathrm{UC}\cdot\mathrm{A}} + f_{\mathrm{UC}\cdot\mathrm{U}} + f_{\mathrm{UC}\cdot\mathrm{G}} + f_{\mathrm{UC}\cdot\mathrm{C}})$. Each 3'-UTR was

scanned by all members of the pattern library. If a motif was found, then the scores of all matching nucleotides were increased by the negated log probability of the pattern. The central nucleotide of a sequence window was assigned the total pattern score found in a window segment.

Algorithm 3.1: Find divergent miRNA sequences

Data: Ordered set S of n miRNA sequences by decreasing length

Result: Heterogeneous miRNA sequences

```
1  begin
2      R ← {S[1]}                                   ▷ Initialize with first sequence of set S
3      for k ← 2 to n do
4          A ← NW(S[k], R, gap_open = 10, gap_extend = 0.5)      ▷ Global alignment[123]
5          I ← Identity(A)          ▷ Number of identical positions for each alignment
6          if ∀i ∈ I : i < 0.73 then
7              R ∪ S[k]
8          end if
9      end for
10     return R
11 end
```

3.1.3 Preparation of training and test data

Feature scaling

To avoid difficulties during model learning, such as domination of features with greater numeric ranges over attributes with smaller numeric ranges, the feature vectors had to be prepared. Some machine learning algorithms will not work properly as their objective functions computing the distance between instances will be governed by the feature(s) with the highest range. In particular, kernel based functions computing the inner products of feature vectors will suffer from numerical problems. Scaling induces an approximately proportionately contribution of each attribute to the final distance. Thus, each feature

distribution x_i was standardize by scaling to the range $[-1,1]$:

$$x_i' = 1 + \frac{2(x_i - \max(x_i))}{\max(x_i) - \min(x_i)} \tag{3.4}$$

Instance selection

The 10% most reliable AGO CCRs were selected from the CLIP-Seq libraries according to the ranking provided by Kishore *et al.*[116]. The central nucleotide of all miRNP binding-sites as well as all nucleotides located 10 nt up- and 10 nt downstream to the center were classified positive (miRNP$^+$). All other nucleotides located on the same 3'-UTR and not within any other CCR measured by AGO PAR-CLIP were classified negative (miRNP$^-$). By this procedure 11 706 positive and 1 307 123 negative instances were created for AGO PAR-CLIP and 5 061 positive and 570 971 negative instances for AGO HITS-CLIP. For subsequent processing, i.e. feature analysis and machine learning, the dataset was balanced by extracting a sample of negative instances.

3.1.4 Feature analysis

To analyze the miRNP binding site features, each feature score distribution x_i was discretized to bins of equal width. Then, the fraction of positive and negative instances per bin was computed. This resulted in two matrices $S_c^{10\times 11}$ of true ($c = \text{miRNP}^+$) and random sites ($c = \text{miRNP}^-$). Using these matrices, the difference between the fraction of miRNP$^+$ and miRNP$^-$ sites for each score bin $b_{i,j}$ was calculated:

$$D(b_{i,j}) = P(b_{i,j}|c = \text{miRNP}^+) - P(b_{i,j}|c = \text{miRNP}^-) \tag{3.5}$$

Here, $D(b_{i,j}) = 0$ indicates no difference in feature score frequency between positive and negative instances; otherwise the feature interval is over-represented in positive instances ($D(b_{i,j}) > 0$) and *vice versa* ($D(b_{i,j}) < 0$).

Since the score intervals were all equal spaced, the common information gain metric was applicable to rank each feature by its overall information content[125] for each AGO CLIP-Seq library d:

$$IG_d(c, x_i) = H(c) - H(c|x_i) \tag{3.6}$$

with

$$\begin{aligned} H(c) &= -\sum_{c} P(c)\log_2 P(c) \\ H(c|x_i) &= -\sum_{c}\sum_{b_{i,j}} P(c)P(c|b_{i,j})\log_2(P(c|b_{i,j})) \end{aligned} \tag{3.7}$$

$H(c)$ denotes the total entropy; $H(c) = 1$ as the data set is balanced ($P(c) = 0.5$). $H(c|x_i)$ is the total entropy considering the information based on feature x_i. The entropy characterizes the impurity of an arbitrary collection of instances. The information gain, i.e. the difference of $H(c)$ and $H(c|x_i)$, describes the expected reduction in entropy caused by separating the instances according to a specific feature. Finally, the average information gain was computed over each data set, i.e. AGO PAR-CLIP and AGO HITS-CLIP. This value will be referred as information gain IG in this study.

3.1.5 Model learning

For model learning, a balanced training set was sampled with $N = 2\,000$ instances. Further, an instance i was defined as $(x,c)_i^T$ with x^i is the feature vector and c^i the class label ($+1$ for miRNP$^+$, -1 for miRNP$^-$). The classification problem was formulated as follows. To separate the two classes linearly by a supervised learning function $\xi(x^i)$, a hyperplane had to be defined[126]:

$$\xi(x^i) = \xi_1(x^i) - \xi_2(x^i) = (w_1^T x^i + w_{1,0}) - (w_2^T x^i + w_{2,0}) = w^T x^i + w_0 \tag{3.8}$$

with the weight vector w and $\xi(x^i) > 0$, then $c^i = +1$, otherwise $c^i = -1$. Thus, w_0 defined the threshold because if $c^i = +1$, then $w^T x^i > -w_0$.

To lower the generalization error, each instance was required to be located at the correct side and also exhibit a certain distance to the hyperplane, i.e. if $\xi(x^i) \geq +1$, then $c^i = +1$ and if $\xi(x^i) \leq -1$, then $c^i = -1$. Here, the optimal hyperplane had to be found to maximize this margin with the least error (instance is located on wrong side or within margin), i.e. $\xi(x^i) \geq 1 - \lambda$. The slack variable λ stores the deviation from the margin, i.e. $\lambda = 0$ is correctly classified, $0 < \lambda < 1$ is correctly classified but located within the margin, and $\lambda \geq 1$ is wrongly classified[126].

An algorithm which is capable of solving this problem is the so-called SVM[127].

It applies a product of basis functions $K(x^i,x) = \phi(x)^T\phi(x)$ (kernel) to perform non-linear transformations to a higher-dimensional space where the problem can be linearly solved[126]. New instances are then mapped into that space and are assigned a class label based on which side of the margin they are located. To train the SVM the Gaussian radial basis function $K(x^i,x) = \exp(-\gamma||x^i - x||^2)$ was selected[126]. This kernel has been shown to be eligible to classify miRNA target sites[128]. To determine the parameters of the SVM, namely the regularization cost C (trade-off between misclassification and complexity of decision surface) and the kernel width γ, a two dimensional grid search was performed with a 5-fold cross-validation to optimize the accuracy of the prediction:

$$\text{accuracy} = \frac{|\text{miRNP}_p^+ \cap \text{miRNP}_o^+| + |\text{miRNP}_p^- \cap \text{miRNP}_o^-|}{|\text{miRNP}_o|} \tag{3.9}$$

Here, miRNP_p denotes the predicted and miRNP_o the experimentally observed sites respectively. Finally, a SVM model was trained using the determined parameters. The entire training process was performed by applying the LIBSVM library[129].

3.1.6 Model evaluation

The accuracy of the SVM was assessed using 5-fold cross validation. Further, it was tested against an independent sample of size 2 000 of the HITS-CLIP data set. To evaluate whether the filtering by predicted miRNP binding sites improves the precision of existing approaches, miRNA transfection experiments were used. The measurements from let-7c, miR-15a, miR-16, miR-17-5p, miR-20, miR-103, miR-106b, miR-141, miR-192, miR-200a, and miR-215 were obtained from the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession GSE6838[130]. The mRNA expression levels in HCT116 Dicer^{ex5} miRNA transfected cells were computed relative to mock-transfected cells 24 h post-transfection. Probe IDs were mapped to the accession number of the longest Ensembl[118] transcript. The probe with the lowest log fold-change for each transcript was selected. Potential miRNA binding sites were predicted by applying TargetScan[50] and TargetSpy[104]. Here, both, the 'sensitive' and the 'specific' version of TargetSpy was evaluated. For each transfection experiment the null hypothesis was tested that the set of filtered predicted targets of any tool exhibit equal or lower fold-changes than a random selection of equal size by means of the one-sided Mann-Whitney U-test.

100 random samples each were extracted allowing the computation of the combined *P* test proposed by Fisher[131]. This test measures whether all of the separate null hypotheses are true:

$$\chi^2_{200} \sim -2 \sum_{i=1}^{100} \ln(P_i) \tag{3.10}$$

Here, the test statistic follows a χ^2 distribution with 200 degrees of freedom. If P_i of each sample i tend to be small, the test statistic χ^2 will be large, suggesting that the null hypotheses can be rejected for every individual test.

3.1.7 Case study

The workflow of the data analysis and modeling is outlined in Figure 3.1. In the following each step is described in detail.

Identification of target sites

All potential miRNP$^+$ sites were predicted for the WISP1 3'-UTR (Ensembl[118] transcript ENST00000250160). Segments composed of 20 nt up- and 20 nt downstream of a miRNP$^+$ site were classified as region with increased affinity to the miRNP. It has been shown that complementarity to the miRNA seed region is most predictive to changes in mRNA levels in response to changes in miRNA concentration[90]. Thus, all potential seed-binding sites were determined using the set of miRNA seed types found significantly enriched in AGO CCRs (Chapter 2). All sites located in miRNP$^+$ regions were assumed to be functional. The stability defined by the hybridization energy ΔG_{hybrid} of the miRNA:mRNA heteroduplex was predicted by the tool IntaRNA requiring the given seed pairing[107].

Preparation of IPF expression data

Two human miRNA expression profiles were obtained from GEO under series GSE13316[132] (10 IPF, 10 control samples) and GSE21394[133] (9 IPF/UIP, 6 control samples). The raw data was normexp background corrected[134], quantile normalized[135], and $\log_2$ transformed. Processed miRNA expression data of a murine bleomycin-induced lung fibrosis model was obtained from Liu *et al.*[136] (ArrayExpress; http://www.ebi.ac.uk/arrayexpress/,

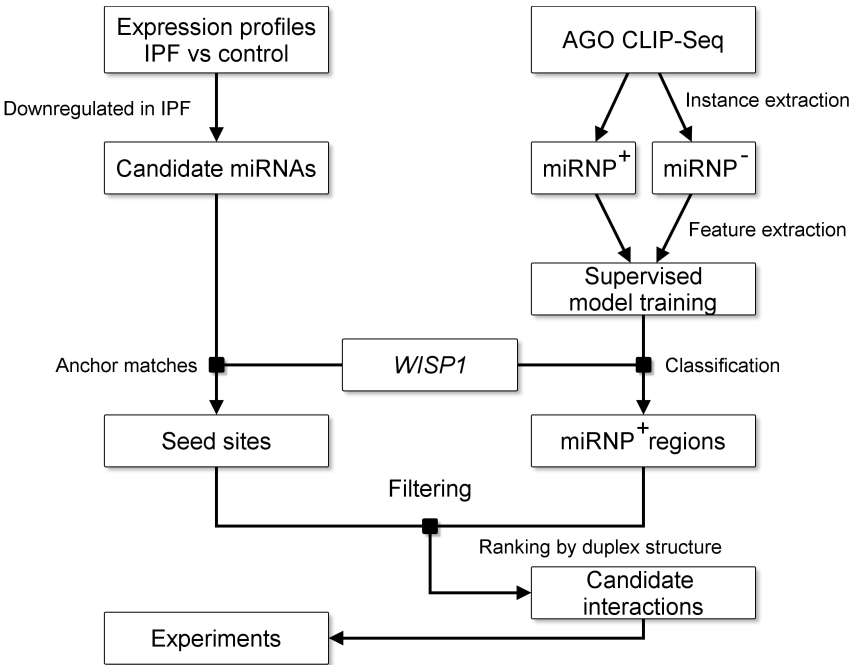


Figure 3.1 | **Flow chart for the identification of candidate miRNA:WISP1 interactions.** Since WISP1 is strongly upregulated in IPF, a set of candidate miRNAs which were downregulated in pulmonary fibrosis was determined. For this purpose, several independent expression studies were examined. Then, the supervised machine learning approach developed in this study was applied (right branch). This classifier was used to rank potential miRNP binding regions (miRNP$^+$) on the WISP1 3'-UTR. Seed matching sites of all candidate miRNAs were determined on the WISP1 3'-UTR. Potential miRNA:mRNA hybrids found within miRNP$^+$ sites were ranked by their duplex energy. The most promising interaction was selected for further experimental investigations.

accession E-MEXP-2749; 9 samples, pooled reference). Values for within-array replicate spots were replaced with their average, at which only probes with an intensity of greater than 95% of the negative controls were considered. For the human IPF studies, the expression fold-change of each miRNA between IPF and control samples was calculated. For the mouse model data, the miRNA expression after 7 and 14 days of bleomycin instillation was compared to day 0. Statistical significance was determined by the Wilcoxon rank-sum test.

Experimental procedures

Experimental material and methods (human tissue, primary human fibroblasts, fibrotic rat lung, cell treatments and transfections, RT-qPCRs, Western blots, enzyme-linked immunosorbent assays, and Luciferase reporter assays) were conducted by external collaborators. For more details, please refer to our corresponding publication (Berschneider *et al.*, 2014[80]).

In this chapter, results of RT-qPCR experiments were specified using the $\Delta\Delta_{Ct}$ method[137]. In a nutshell, the relative value of the RNA concentration of the target gene in the PCR reaction was denoted by the threshold cycle (Ct) metric. The Ct was defined as the number of cycles required for the fluorescent signal to cross a specified threshold, i.e. exceeds the background level. Replicates were averaged by the arithmetic mean. The relative transcript abundance of a target gene relative to a reference (housekeeping) gene was computed by $\Delta_{\mathrm{Ct}} = \mathrm{Ct}^{\mathrm{reference}} - \mathrm{Ct}^{\mathrm{target}}$. The relative changes of RNA levels between conditions was calculated by $\Delta\Delta_{\mathrm{Ct}} = \Delta_{\mathrm{Ct}}^{\mathrm{treated}} - \Delta_{\mathrm{Ct}}^{\mathrm{control}}$.

3.2 Results

3.2.1 miRNP binding site features

Recent studies using cross-linking and immunoprecipitation (CLIP) provide a properly mapping of transcriptome-wide binding sites of RNA-binding proteins. In particular, CLIP of AGO proteins provides specific regions of AGO binding and miRNA target sites. This raises the question whether this information can be used to facilitate miRNA target prediction by filtering candidate 3'-UTRs for regions likely bound by the miRNPs. Further, since the AGO CLIP-Seq protocol is difficult to perform and, *per se*, restricted to a specific transcriptome, it is of particular interest whether the *ab initio* identification of miRNP binding sites is feasible.

To build a computational model by learning from CLIP-Seq data, selective features had to be extracted from a set of representative instances. For this purpose, a set of *bona fide* miRNP binding sites (miRNP$^+$) and off-sites (miRNP$^-$) was prepared using the libraries of two recent AGO CLIP-Seq experiments[116]. In detail, a miRNP$^+$ site corresponds to the nucleotide located at the center of a CCR of 41 nt length whereas miRNP$^-$ sites were

not found in any experimental measurement. In total, two balanced data sets with 23 412 AGO PAR-CLIP and 10 122 HITS-CLIP instances were generated, respectively.

Several characteristics have been proposed to classify RNA segments in general, and miRNA target sites in particular. In this study, 11 of these features were collected and adapted to enable an unbiased characterization of miRNP binding sites. These allude a variety of characteristics: sequence-based (the asymmetry bias of AU and GC, the local content of U, AU, GU, and the compositional entropy), thermodynamic-based (the required energy to access the site), homology-based (the local conservation by PhastCons[96] and PhyloP[119]), motif-based (miRNA pattern aggregation), and position-based (the relative location of the binding site on the target transcript). The features were analyzed by measuring the differences of the score distributions between positive and negative instances. Further, the information content was rated by the information gain (IG) metric (Figure 3.2).

In general, all features exhibited a non-zero IG and, as such, contain relevant site information. In particular, it was found that miRNP target regions were preferentially conserved and accessible, i.e. less energy is required to unfold the mRNA. This observation goes well with the current common consensus on feature selection for miRNA target prediction. The windowed PhastCons scoring scheme performed slightly better than the per-base PhyloP framework in weighting conserved functional sites. PhastCons scores the probability that a nucleotide belongs to a conserved element whereas the PhyloP score denotes the $-\log(P)$ under a null hypothesis of neutral evolution.

It should be noted that, although different in data basis and calculation, the presented study reproduces the previous observations of Kertesz *et al.*[49]. The difference in accessibility between positive and negative instances perfectly followed their reported distribution. As expected, the high IG of the accessibility attribute was accompanied by a characteristic local sequence composition: miRNP target sites exhibited a high AU content which indicates local structures composed of pairings with only two hydrogen bonds. The AU content and the total energy required to unfold this region were negatively correlated (Figure 3.3).

Interestingly, the distribution of G and C was remarkably right skewed, i.e. genuine miRNP target sites hold a higher fraction of C than G (median GC skew in AGO PAR-CLIP $= -0.22$ and in AGO HITS-CLIP $= -0.28$). By contrast, the nucleotide composition of negative instances followed Chargaff's second parity rule[138] (median GC skew ~ 0). This feature gave one of the best discriminability and, to my best knowledge, was not

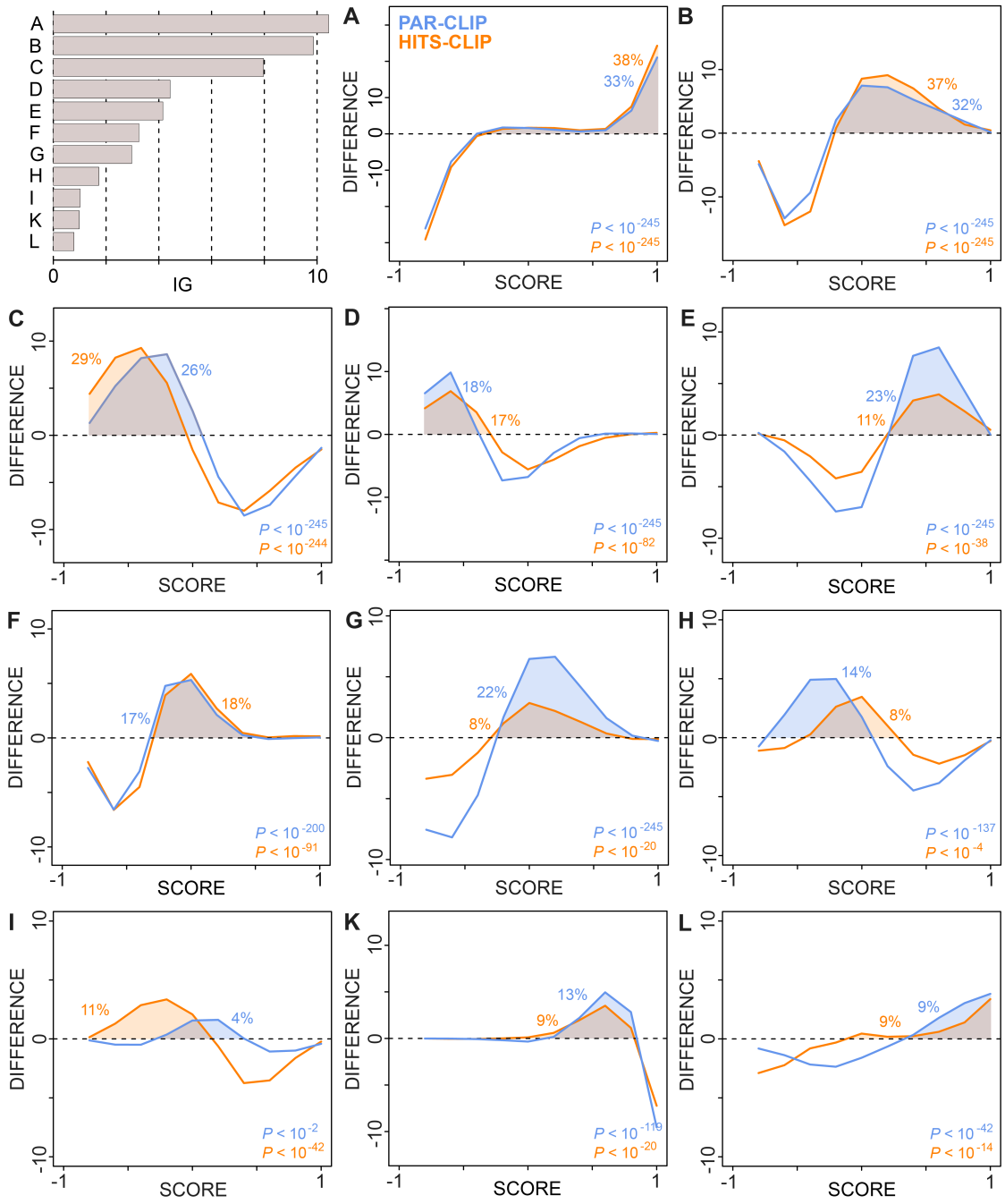


Figure 3.2 | **Analysis of AGO site characteristics.** For each score value (x-axis) the difference between the distributions of miRNP binding sites and randomly selected regions of the 3'-UTR is plotted (in %). The fraction of miRNP binding sites that differ from random sites is indicated (shaded area). *(continued on next page)*

Figure 3.2 *(previous page)* | The total information gain metric (IG) was used to rank each feature by its information content (upper left panel). The following features were analyzed for PAR-CLIP (blue curves) and HITS-CLIP (orange curves): PhastCons conservation score (**A**), PhyloP conservation score (**B**), GC asymmetry bias (**C**), total energy required to unfold the segment (**D**), local AU content (**E**), miRNA pattern aggregation (**F**), local U content (**G**), AU asymmetry bias (**H**), local GU content (**I**), compositional entropy (**K**), and relative position (**L**). *P* are given by the Wilcoxon rank sum test and indicate the significance level of the difference between the distributions. Note that the y-axes were not scaled equally to foster a better comprehension.

directly used for miRNA target prediction yet. The GC skew described the data better (IG $= 8.0$) than established features, such as the target site accessibility (IG $= 4.4$), the local AU content (IG $= 4.2$), and the relative site location (IG $= 0.8$). An explanation for this skew may be a propensity to form specific local RNA structures[121,139]. These folds are likely more accessible for the RISC as the energy required to unfold these segments is positively correlated with the C% $>$ G% bias (AGO PAR-CLIP $\rho = 0.34$, AGO HITS-CLIP $\rho = 0.32$). While the purine G pairs with C by three hydrogen bonds, it is also capable to form common non-Watson-Crick G-U and U-G wobble pairs in RNA structure[140]. Thus, lowering the fraction of G may decrease the number of stable hydrogen bonds. Consistently, the local GU content is positively related with the C% $>$ G% ratio (AGO PAR-CLIP $\rho = 0.52$, AGO HITS-CLIP $\rho = 0.59$).

Further, an enrichment of miRNA binding motifs was observed. This result suggests that these miRNP target segments may embed multiple target sites for several miRNAs. Notably, the difference in score distributions was not growing on a linear basis. This indicates that the number of operative miRNA target sites is limited to a certain amount within miRNP-bound regions. All pattern matches were preferentially located in regions of higher AU and lower GU content, and as such, segments of superior accessibility (Figure 3.3). Notably, these regions exhibited a higher fraction of A than U. This is interesting as adenines have been assigned a prominent role as miRNA target site anchors[90] and the very 5'-terminal nucleotide of the major fraction of guide miRNA sequences is U enabling Watson-Crick base pairing to the target sequence (α-seed types, Chapter 2).

The local U and the local GU content varied to a considerable amount between the AGO PAR-CLIP and the AGO HITS-CLIP library. This effect is likely caused by the experimental protocols. Indeed, Hafner *et al.*[70] previously found an elevated U-content

in their AGO PAR-CLIP data (HEK293 cells). They argued that this was as expected according to previous analyses of functional miRNA binding sites. The direct comparison of AGO PAR-CLIP and AGO HITS-CLIP in the presented study confirmed that there is a very likely propensity of AGO to bind U enriched sites, albeit this holds only to a certain extend. Although AGO2 has been suggested to exhibit a higher affinity for the base of U monophosphate than for other bases[39], a binding preference of AGO proteins for U-enriched target sequence tracts has not been reported in the literature. The observed discrepancy between both protocols suggests that the PAR-CLIP technique holds a noticeable experimental bias. It can be suggested that the sequence propensity is intrinsic to the usage of 4-thiouridine in this protocol[141].

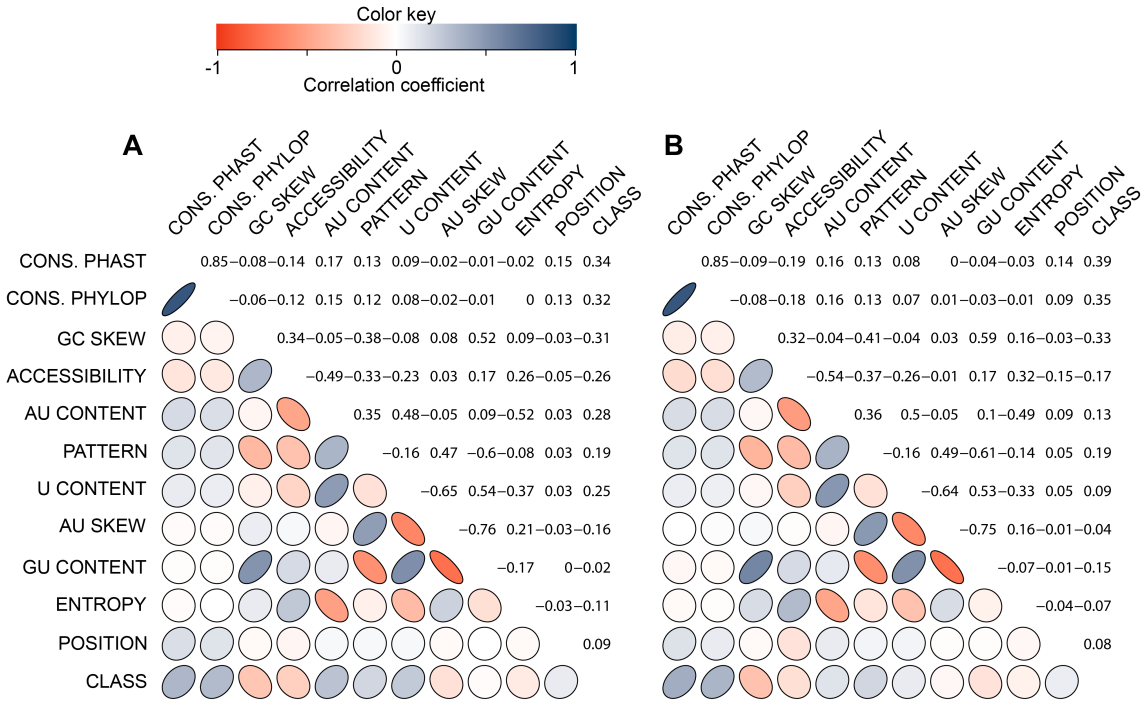


Figure 3.3 | **Correlation between miRNP site features.** Correlation is measured by Pearson's ρ and is shown for the AGO PAR-CLIP (**A**) and AGO HITS-CLIP (**B**) library. The color and shape of the ellipses denotes the strength and type of association (negative correlation: red, bent backwards; positive correlation: blue, bent forwards). Please note that the feature accessibility is defined as the total energy required to unfold the segment.

3.2.2 'Hot spot' filtering raises precision of miRNA target prediction

Miranda *et al.*[113] proposed the idea of the association of 'hot spots', so-called 'target islands', regions with aggregations of putative miRNA binding sites on the 3'-UTR. In their work they showed the statistical importance and advantage of filtering target mRNA sequences. This study extends their idea by integrating novel biological information about AGO binding sites from recent AGO CLIP-Seq libraries. A SVM model was trained using the set of 11 miRNP binding site features. The classifier achieved 75.3% accuracy by 5-fold cross-validation. Further, the SVM was tested against a set composed of 2 000 instances from the AGO HITS-CLIP library and obtained a decent accuracy of 68.85%. Again, this drop in performance may point to an experimental bias of AGO-PAR CLIP as already discussed. However, the classifier evaluation suggests that the prediction of miRNP *cis*-regulatory regions without guide miRNA sequence information is feasible. It also emphasizes the complex binding specificity of AGO2.

Next, it was investigated whether the classifier is able to improve common miRNA target prediction. An evaluation based on miRNA transfection data exhibited promising results. Since available target prediction methods are based on varying principles, two common approaches were selected exemplary, representing two basic paradigms: i) TargetScan[50] which requires a Watson-Crick pairing between the miRNA seed and the target sequence, and ii) TargetSpy[104] which predicts target sites regardless of the presence of a seed match. The latter is a machine learning approach which has been trained to optimize prediction accuracy in two ways: a very restrictive version referred as $\text{TargetSpy}_{\text{spec}}$ and sensitive version referred as $\text{TargetSpy}_{\text{sens}}$. The log fold-change distribution of predicted target mRNAs following transfection of 11 miRNAs was evaluated (Figure 3.4A). It became apparent that filtering by miRNP^{+} sites reduces the fraction of targets with no or less repressive effects for each prediction algorithm. The significance of this observation was tested against 100 random samples of target sites of equal size for each transfection experiment and reached an α-level of $P < 0.05$ for at least four measurements each (number of miRNAs: $\text{TargetScan}^{+} = 11$, $\text{TargetSpy}^{+}_{sens} = 10$, $\text{TargetSpy}^{+}_{spec} = 4$). The filtering effect can be interpreted as an increase in precision, i.e. the fraction of predicted targets that are assumed to be functional. To specify the value of the filtering effect, the top 20% most downregulated targets were defined as true positives (according to Betel *et al.*[142]). This enabled the estimation of the precision of each prediction algorithm. Figure 3.4B shows

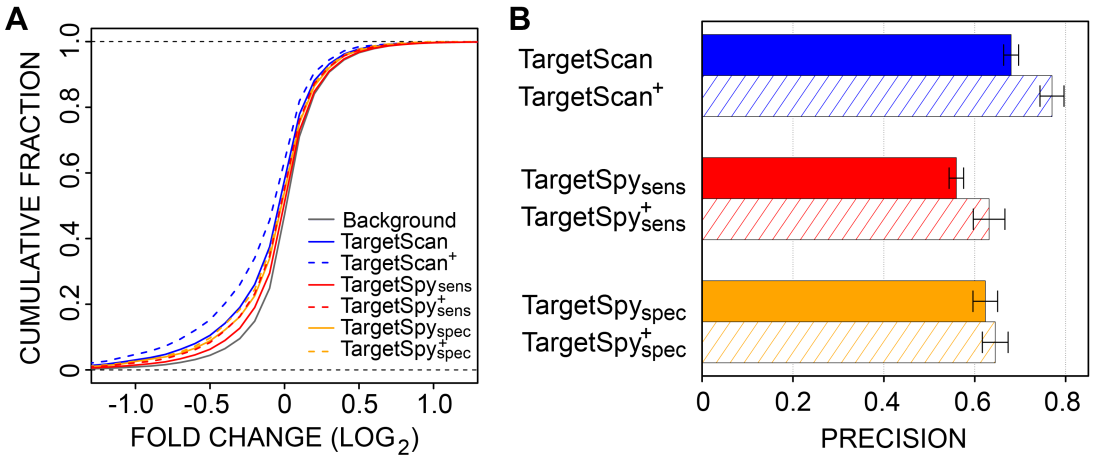


Figure 3.4 | **Evaluation of miRNA target predictions filtered by miRNP$^+$ segments. A** | Distribution of $\log_2$ fold-changes of predicted target mRNAs following miRNA transfection (straight lines). A downward shift in the cumulative distribution of mRNAs holding a target site located in a predicted miRNP$^+$ region can be observed (dashed lines; TargetScan$^+$, TargetSpy$^+_{sens}$, TargetSpy$^+_{spec}$). The cumulative fraction is computed by the arithmetic mean of all 11 miRNA transfection experiments. **B** | Average precision of miRNA target prediction over all 11 transfection experiments. The top 20% downregulated transcripts of each miRNA were defined as positives. The filtering by miRNP$^+$ segments advances the precision. The error bars denote the 95% confidence interval for the mean.

that filtering by miRNP$^+$ regions improved the precision of common prediction algorithms (gain in precision: TargetScan$^+ = 9.0\%$, TargetSpy$^+_{sens} = 7.2\%$, TargetSpy$^+_{spec} = 2.1\%$).

3.3 Case study: miRNA-mediated regulation in IPF

Recently, several studies proposed a role for miRNAs in IPF[25,132,136], the most common and aggressive form of idiopathic interstitial pneumonia[143]. Its etiology is uncertain and therapeutic interventions are still limited[144,145]. IPF is characterized by aberrant remodeling and profound changes in the phenotypes of alveolar epithelial cells and lung fibroblasts[25]. Normally, the pulmonary epithelium and underlying mesenchymal cells in the epithelial-mesenchymal unit communicate with each other through cytokines and growth factors coordinating growth and response to injury[146]. One of the key pathological hallmarks is the disturbed growth factor signaling within the epithelial-mesenchymal unit

whereat the transforming growth factor-β1 (TGF-β1, TGFB1) has been identified as key pro-fibrotic mediator[147]. Also Wnt signaling was suggested to contribute to epithelial as well as mesenchymal cell dysfunction and/or reprogramming in experimental and human pulmonary fibrosis[148,149]. At this, Königshoff *et al.*[79] reported a markedly increased Wnt target gene in IPF, WISP1. Its repression resulted in attenuated pulmonary fibrosis development *in vivo*. They suggested that the secreted multicellular protein WISP1 is an important mediator of perturbed epithelial-mesenchymal crosstalk. It is largely unknown how its gene expression is regulated. In this respect, the modulation of aberrantly regulated miRNAs was proposed as potential treatment to affect the development of pulmonary fibrosis[150,151]. Interestingly, the WISP1 transcript exhibits a long 3'-UTR – a region preferentially targeted by miRNAs. Thus, it was hypothesized that miRNA-mediated WISP1 regulation may get lost in pulmonary fibrosis. This event may induce WISP1 upregulation which in turn contributes to the pathogenic phenotype.

3.3.1 Candidate miRNAs regulating WISP1

To reduce the set of candidate miRNAs, mature sequences that were measured downregulated in pulmonary fibrosis were of particular interest. Therefore, two miRNA array studies of human and mouse fibrotic lung tissues[132,152] were compared. Here, 30 miRNAs significantly ($P < 0.05$) decreased in human IPF tissue specimens were identified. Of these, 13 miRNAs also exhibited decreased transcript levels after 7 and 14 days of bleomycin instillation in a murine lung fibrosis model. For each of these miRNAs (let-7d, let-7g, miR-26a, miR-26b, miR-30a-5p, miR-30b, miR-30d, miR-92a, miR-101, miR-203, miR-326, miR-375, and miR-598), the WISP 3-'UTR was scanned for seed complementary sites (according to the seed type set described in Chapter 2). The structure and stability, ΔG_{hybrid}, of the resulting 30 miRNA:mRNA duplices were computed, respectively (Table 3.1). Then, 14 miRNP$^+$ segments were identified on the WISP1 3'-UTR (Ensembl[118] transcript, 2 634 nt length) by the novel SVM-based classifier. Here, 13 miRNA target sites were located in a miRNP$^+$ region. In the end, miR-92a (also known as miR-92a-3p) was found the most likely regulator of WISP1 (Figure 3.5). This miRNA exhibited a differential expression in IPF and was predicted to bind to a target located within a miRNP-preferred region forming a stable RNA duplex geometry ($\Delta G_{\text{hybrid}} = -13.30$ kcal/mol, seed type = 7merα).

Table 3.1 | **Potential target sites of miRNA candidates.** Listed are seed matches and their position (Pos.) on the WISP1 3'-UTR. Symbol names and miRBase MIMAT accession numbers are based on GSE13316 (miRBase 9.1, GEO platform GPL6955, human). Fold-changes (FC) for E-MEXP-2749 (mouse) were computed after 7 and 14 days of bleomycin instillation (d_7; d_{14}).

Symbol	**MIMAT**	**$FC_{GSE13316}$**	**$FC_{E\text{-}MEXP\text{-}2749}$**	**Pos.**	**Seed**	**Class**	**ΔG_{hybrid}**
miR-92a	0000092	0.71	0.98; 0.96	1 600	7merα	miRNP$^+$	-13.30 kcal/mol
miR-101	0000099	0.62	0.72; 0.59	455	8merα	miRNP$^+$	-12.73 kcal/mol
let-7g	0000414	0.80	0.88; 0.93	516	6merα	miRNP$^+$	-12.50 kcal/mol
miR-203	0000264	0.53	0.86; 0.82	125	6merγ	miRNP$^+$	-12.30 kcal/mol
miR-598	0003266	0.89	0.98; 0.93	473	6merβ	miRNP$^+$	-11.01 kcal/mol
miR-203	0000264	0.53	0.86; 0.82	78	6merα	miRNP$^+$	-10.50 kcal/mol
miR-92a	0000092	0.71	0.98; 0.96	2 583	6merβ	miRNP$^+$	-8.70 kcal/mol
miR-598	0003266	0.89	0.98; 0.93	148	6merγ	miRNP$^+$	-8.29 kcal/mol
miR-30a-5p	0000087	0.54	0.84; 0.75	1 811	7merβ	miRNP$^+$	-8.26 kcal/mol
miR-375	0000728	0.70	0.91; 0.88	415	6merβ	miRNP$^+$	-7.40 kcal/mol
miR-30b	0000420	0.54	0.80; 0.74	1 811	7merβ	miRNP$^+$	-6.40 kcal/mol
miR-30d	0000245	0.51	0.82; 0.71	1 811	7merβ	miRNP$^+$	-5.86 kcal/mol
miR-101	0000099	0.62	0.72; 0.59	1 314	6merβ	miRNP$^+$	-5.00 kcal/mol
let-7d	0000065	0.76	0.93; 0.97	1 055	6merα	miRNP$^-$	-18.50 kcal/mol
miR-30a-5p	0000087	0.54	0.84; 0.75	1 183	6merβ	miRNP$^-$	-16.20 kcal/mol
miR-30d	0000245	0.51	0.82; 0.71	1 183	6merβ	miRNP$^-$	-15.50 kcal/mol
miR-326	0000756	0.84	0.93; 0.84	1 480	6merβ	miRNP$^-$	-13.90 kcal/mol
miR-30b	0000420	0.54	0.80; 0.74	1 183	6merβ	miRNP$^-$	-13.50 kcal/mol
miR-30a-5p	0000087	0.54	0.84; 0.75	1 194	6merγ	miRNP$^-$	-13.39 kcal/mol
miR-30d	0000245	0.51	0.82; 0.71	1 194	6merγ	miRNP$^-$	-13.29 kcal/mol
miR-26a	0000082	0.83	0.74; 0.65	1 976	7merβ	miRNP$^-$	-11.51 kcal/mol
miR-326	0000756	0.84	0.93; 0.84	726	6merβ	miRNP$^-$	-10.83 kcal/mol
miR-26b	0000083	0.74	0.77; 0.67	1 976	7merβ	miRNP$^-$	-10.11 kcal/mol
miR-203	0000264	0.53	0.86; 0.82	1 125	8merα	miRNP$^-$	-9.92 kcal/mol
miR-26a	0000082	0.83	0.74; 0.65	827	6merα	miRNP$^-$	-9.83 kcal/mol
miR-203	0000264	0.53	0.86; 0.82	2 109	6merβ	miRNP$^-$	-9.46 kcal/mol
miR-92a	0000092	0.71	0.98; 0.96	1 947	6merγ	miRNP$^-$	-7.83 kcal/mol
miR-30b	0000420	0.54	0.80; 0.74	1 194	6merγ	miRNP$^-$	-7.70 kcal/mol
miR-101	0000099	0.62	0.72; 0.59	1 918	6merβ	miRNP$^-$	-5.03 kcal/mol
miR-26b	0000083	0.74	0.77; 0.67	827	6merα	miRNP$^-$	-4.53 kcal/mol

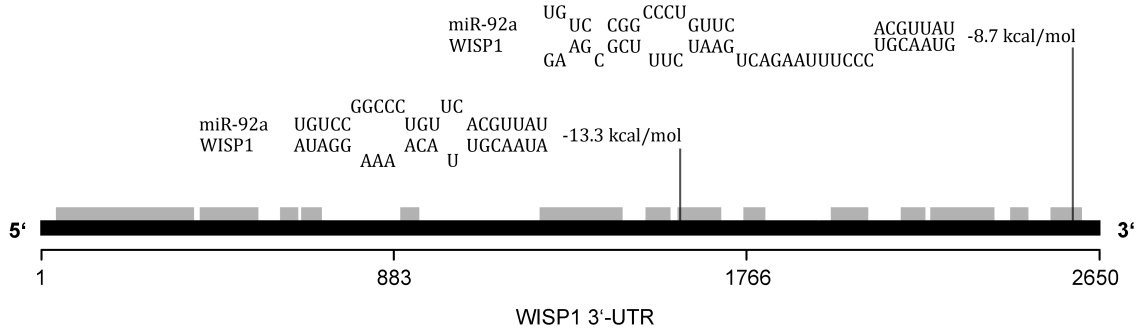


Figure 3.5 | **Predicted target sites of miR-92a on the WISP1 3'-UTR.** Schematic representation of miR-92a binding sites on the human WISP1 3'-UTR. Solid boxes represent miRNP$^+$ segments; two miR-92a seed-complementary sites were found located within these regions (7merα seed match at position 1 600 and 6merβ seed match at position 2 583). For each candidate binding site, the predicted miRNA:target duplex structure and the hybridization energy is shown, respectively.

Next, it was aimed to confirm the candidate selection approach by additional experimental measurements. First, the miR-92a downregulation was confirmed in a miRNA array of human IPF tissue samples[133] (GEO GSE21394; fold-change $= 0.45$, $P = 1.2 \times 10^{-02}$). Second, the expression of miR-92a was measured in pulmonary fibrosis tissue specimens by RT-qPCR (Figure B.1A). Again, significantly lower levels in IPF compared to control samples (donors from unaffected lung tissue biopsies) were observed ($\Delta_{\mathrm{Ct}}^{\mathrm{donor}} = 2.92$, $\Delta_{\mathrm{Ct}}^{\mathrm{IPF}} = 1.63$, $P = 2.9 \times 10^{-02}$). Third, the WISP1 protein level was measured in six whole IPF lunge tissue homogenate and six donor samples using Western Blotting. Here, WISP1 was found upregulated (Figure B.1B)[80]. Fourth, the miRNA target prediction was confirmed by firefly luciferase reporter gene assays (Figure B.1C). It was of interest whether miR-92a affects the protein concentration of human WISP1 via its 3'-UTR. Thus, HEK cells were transfected with a miR-92a inhibitor and a reporter plasmid containing the WISP1 3'-UTR cloned downstream to firefly. The inhibition of miR-92a led to a significantly increased luciferase activity indicating that miR-92a regulates WISP1 by binding to its 3'-UTR (ratio of firefly activity with and without WISP1 3'-UTR $= 1.2$, $P = 2.3 \times 10^{-2}$)[80].

3.3.2 miR-92a regulates TGFB1-induced WISP1 expression

WISP1 is a matricellular protein and as such is poorly expressed under homeostatic conditions and likely modulated by growth factor signaling[79]. Berschneider *et al.*[80] investigated whether TGF-β1, one of the most potent pro-fibrotic cytokines involved in myofibroblast differentiation[147], impacts WISP1 expression. They observed increased WISP1 mRNA and protein levels in human pFB after 24 h of TGF-β1 treatment. At this, WISP1 upregulation was accompanied by myofibroblast differentiation. Further, Western blotting and an enzyme-linked immunosorbent assay showed that WISP1 protein concentration was induced in supernatants of TGF-β1-treated human pFB indicating enhanced secretion of WISP1.

To examine whether TGF-β1-induced WISP1 expression is altered by miRNA regulation in a pro-fibrotic environment *in vitro*, miR-92a and a negative control was transfected to human pFB cells (Figure B.2A). It was observed that miR-92a buffered the TGF-β1-induced WISP1 mRNA level ($\Delta\Delta_{\mathrm{Ct}}^{\mathrm{Control}} = 1.24$, $\Delta\Delta_{\mathrm{Ct}}^{\text{miR-92a}} = 1.02$, $P = 9.6 \times 10^{-2}$). The expression of the TGF-β1 target genes COL1A1 and FN1 was induced by TGF-β1 treatment but not altered by miR-92a. This indicates that miR-92a targets WISP1 specifically[80].

These findings were corroborated by analyzing WISP1 protein expression in pFB upon miR-92a inhibition (Figure B.2B). Significantly increased WISP1 protein concentrations were observed (fold-change $= 2.1$, $P < 0.05$)[80]. These results suggest that miR-92a modulates WISP1 protein expression and that this regulatory interaction may contribute to WISP1 stimulation in pulmonary fibrosis.

Next, it was elucidated if the post-transcriptional regulation of WISP1 by miR-92a can also be observed *in vivo*. For this purpose, lung tissue homogenate samples were analyzed by RT-qPCR at day 7 and 21 after viral overexpression of TGF-β1 in the rat lung, a fibrosis model proposed by Sime *et al.*[153] (Figure B.3A−C). Increased levels of WISP1 mRNA ($\Delta\Delta_{\mathrm{Ct}}^{\mathrm{day7}} = 2.59$, $\Delta\Delta_{\mathrm{Ct}}^{\mathrm{day21}} = 1.82$) and decreased levels of miR-92a ($\Delta\Delta_{\mathrm{Ct}}^{\mathrm{day7}} = -0.87$, $\Delta\Delta_{\mathrm{Ct}}^{\mathrm{day21}} = -0.85$, $P = 2.0 \times 10^{-02}$) were found on both days compared to controls from day 21. At this, the expression levels of WISP1 and miR-92a exhibited a significant negative correlation as given by the coefficient of determination from linear regression $R^2 = 0.43$ ($P = 6.1 \times 10^{-03}$)[80].

It was assessed whether this association can be translated from the rat fibrosis model to

human IPF. Therefore, human pFB from fibrotic and non-fibrotic patients were analyzed (Figure B.3D and E). Increased mRNA levels of WIPS1 in primary fibroblasts from IPF patients were observed ($\Delta\Delta_{\mathrm{Ct}} = 4.67$, $P = 4.6 \times 10^{-02}$). The correlation between WISP1 mRNA and miR-92a levels was confirmed ($R^2 = 0.69$, $P = 1.1 \times 10^{-02}$)[80]. These results suggest that miR-92a regulates WISP1 expression in experimental and human fibrosis *in vivo*.

3.4 Conclusion

Chapter 2 of this thesis focused on the most important feature determining miRNA efficiency, the specific miRNA seed-pairing to the target sequence. In this study, the analysis was extended by a superordinate level, i.e. not the target attributes of a particular miRNA-pairing site, rather general features specifying binding sites of the miRNP complex were elucidated. Several studies have proposed the idea that miRNA binding sites can be found aggregated in 'target islands' on the mRNA[104,113]. Indeed, Miranda *et al.*[113] have shown in their work that this idea is meaningful from a statistical point of view. A crucial shortcoming of miRNA target prediction is the bulk of spurious matches, non-functional miRNA complementary sites. This problem has been discussed by Karlin and Altschul[154] in the context of sequence similarity detection during the development of their BLAST algorithm[155]. In particular, the number of complementary sites m of length l_m in the target sequence t of length l_t having a score $s \geq S$ expected by chance can be approximated by a Poisson distribution with $m = K \times l_m \times l_t \times \mathrm{e}^{-\lambda S}$. Here, K and λ denote constants depending on the scoring system. Using this equation the probability of a target prediction algorithm to report one ore more spurious matches can be calculated by $P = 1 - \mathrm{e}^{-m}$[113]. Obviously, P depends on l_t (i.e. the longer the target sequence, the higher the likelihood to detect one or more spurious matches), S (e.g. miRNA:mRNA hybridization energy), and l_m (i.e. the number of pairing nucleotides). The scoring scheme is given by the miRNA target prediction approaches and l_m defined by its search paradigm (i.e. with or without seed requirement). By shortening l_t to 'target islands', the probability that a method will report spurious matches will be reduced. Miranda *et al.*[113] described that if an algorithm examines only $1/p$-th of a target sequence, the number of false positives is reduced by a factor approximately equal to p. This advantage holds across a very wide range of values

for *m* and was suggested to be valid for a magnitude of real-world cases[113].

In the presented study, two recent AGO CLIP-Seq libraries were used to extract attributes describing miRNP binding regions on target mRNAs. At this, an elaborate set of instances composed of 16 767 positive and 16 767 negative sites was used. Amongst others, the conservation, the accessibility of the target region and a skewed G and C distribution was found to contain high information contents. These features are independent of any miRNA sequence and thus are relevant for the AGO target region. By training a SVM classifier it was shown that miRNP binding sites can be distinguished from random segments. It can be suggested that the basic idea proposed by Miranda *et al.*[113] which was motivated by a statistical point of view, can also be motivated by a biological prospect as there are characteristic regions preferentially targeted by RISC (referred as miRNP$^+$). Filtering of existing miRNA target prediction algorithms by miRNP$^+$ reduced the number of interactions with low regulatory effects and improved their precision, respectively. However, the comprehension of the molecular level of miRNP:mRNA pairing is still limited and, as such, the lack of features handicaps the classification of these sites resulting in false predictions. There is still potential for improvement as was shown by the identification of novel discriminating attributes in this study. Additionally an experimental bias was found for AGO PAR-CLIP likely induced by the requirement of 4-thiouridine transfected cells. Thus, target prediction will certainly also benefit from future experimental advancements. At this, unbiased protocols with a higher resolution, i.e. revealing genuine miRNA:mRNA duplices, are of particular interest.

The SVM was used to predict miRNA:WISP1 interactions in IPF. High WISP1 levels have been associated with *de novo* collagen synthesis in bleomycin-induced lung fibrosis[156]. Berschneider *et al.*[80] detected increased WISP1 protein secretion upon TGF-β1 *in vitro* and *in vivo* and suggested an autocrine and paracrine function of WISP1 within the epithelial-mesenchymal trophic unit in IPF.

A set of candidate miRNAs was composed that were found downregulated in three microarray studies of pulmonary fibrosis. Among these, miR-92a exhibited two target sites with a low miRNA:mRNA duplex energy located in miRNP$^+$ segments. This miRNA is a member of the polycistronic miR-17$\sim$92 cluster which was suggested to play an important role in lung development and homeostasis. Animals with an introduced miR-17$\sim$92 gene knock-out died postnatal likely caused by hypoplastic lungs and ventricular septal defects[157]. The cluster has been found expressed in epithelial lung progenitor

cells promoting their proliferative and undifferentiated phenotype[158]. Interestingly, other members of the miR-17∼92 cluster were not found downregulated in the data of Pandit *et al.*[132] (GEO GSE13316). Recently, it has been reported that post-transcriptional regulation of polycistronic clusters led to different mature miRNA levels under hypoxia conditions[159]. It can be suggested that this may occur in pulmonary fibrosis as well. Dakhlallah *et al.*[160] suggested that miR-92a is downregulated due to epigenetic modifications in IPF. They observed an enhanced DNA methylation at the miR-17∼92 cluster locus in IPF fibroblasts. Further investigations are required to elucidate whether miR-92a downregulation in IPF is caused by cytokine signaling and/or epigenetic modifications during fibrotic reprogramming.

Subsequent experiments validated that miR-92a downregulation is accompanied by increased WISP1 protein levels in IPF samples. Further, miR-92a reversed TGF-β1 induced WIPS1 mRNA expression *in vitro* and its inhibition led to WISP1 protein upregulation. It was found that miR-92a and WISP1 expression were significantly negative correlated in lung tissue homogenate samples in a fibrosis rat model *in vivo* and in primary human fibroblasts from IPF patients *ex vivo*. These findings indicate, for the first time, an altered post-transcriptional regulation of WISP1 in pulmonary fibrosis.

It should be noted that it remains unclear if the activation of Wnt signalling is a causal or a reactive process confounded by the IPF onset. Further, it is evident that miRNAs modulate signal propagation by regulating compounds in all horizontal layers of signaling networks[161]. Thus, it can be expected that also other members of the TGF-β1 and/or Wnt/β-catenine pathways are modified. While it was shown that miR-92a is able to directly target WISP1, its repression may also lead to altered TGF-β1/Wnt signaling and, as such, may induce offsite effects downstream to WISP1 expression.

CHAPTER 4

Genetic variation affecting the miRNA regulome

Aberrant miRNA expression contributes to significant cell biological consequences, perturbed organismal homeostasis, and ultimately leads to pathogenesis of fatal diseases[162,163]. Recent databases list hundreds of miRNAs linked to more than 100 human disorders[164,165]. Several studies reported miRNAs as valid biomarkers for complex traits[165–168]. Thus, the mechanisms affecting miRNA-mediated regulation are of particular interest. In Chapter 1.4.3 it was described that miRNAs exhibit an average modest level of target repression. However, an ectopic miRNA expression impacts the post-transcriptional regulatory effect on target transcripts. Also the local target site composition has been reported as important determinant of regulatory efficiency. Here, several key attributes, such as sequence complementarity (Chapter 2) and target site geometry (Chapter 3), were discussed in the previous chapters. Some first reports suggest that a disruption of *bona fide* target sites affects complex traits and diseases, such as cancer[169].

The genomic diversity in humans arises from 1% of variation[170], mostly induced by SNPs. These denote the occurrence of several, most common two, different nucleotides at the same genomic locus (allele) within a particular population. Here, the allele of lower frequency defines the minor allele frequency (MAF) of a SNP. Since the genetic component tend to run in families without following the classical Mendel's laws of inheritance, an association approach is applied. These kind of studies comprise the scanning of markers, such as SNPs, across complete sets of genomes of many individuals to identify genetic variations which are significantly more common in affected than in unaffected individuals[171,172]. Until now several hundred genome-wide association (GWA) studies have been performed associating a huge amount of predisposing variants to over a

hundred disorders and traits[173]. Since associations are conducted by correlation, GWA approaches are limited in the determination of causal loci, nor reveal their functional basis underlying disease risk. Thus, an ever growing amount of signals found in GWA studies is awaiting mechanistic characterization.

The genomic distribution of SNPs is not homogeneous. Over 90% of the common trait-associated variations identified by GWA studies are located within non-coding regions of the human genome[174] of which functional annotation remains limited. However, it has been supposed that many of these polymorphisms are likely to underlie perturbed dynamics of gene expression between individuals[175].

One of the first reports of miRNA-related SNPs was in 2005 by Abelson *et al.*[176], describing a mutation in the target site of miR-189 at SLITRK1 which was associated with Tourette's syndrome. Since then, several attempts were conducted to identify potential interrelations of aberrant miRNA regulation and genetic variation[177–181]. Amongst others, a signal from a GWA study was suggested, for the first time, to be explained by polymorphic miRNA targeting in the risk for Crohn's disease[182]. However, these studies were lacking comprehensive data on trait-associated polymorphisms and are often limited to *in silico* MREs which exhibit a high false positive rate (Chapter 2). Genetic variants in the 3'-UTR, and in particular in the miRNP binding region, have long been neglected for the most part of all GWA studies. Hence, published polymorphisms affecting the miRNA regulation pathway are rare[180,181].

Recent experimental technologies using cross-linking and AGO immunoprecipitation coupled with high-throughput sequencing (CLIP-Seq) enables an accurate transcriptome-wide determination of miRNP binding sites (Chapter 1.5.2). This study aims to identify mechanisms affecting miRNA-mediated regulation integrating AGO CLIP-Seq data and SNPs from public GWA studies. First, it was shown that trait-associated SNPs were enriched in the 3'-UTR − a region encoding the major fraction of operative miRNP binding sites. At this, affected genes were found enriched in lipid metabolism processes. Using computational analyses, the impact of 3'-UTR SNPs on several target site features was investigated: i) the loss of 3'-poly(A) signals which has been described to cause genetic diseases by cap-dependent miRNA-mediated degradation of the mRNA[183,184], ii) local changes of the target RNA structure which impacts an important feature for the binding affinity of the miRNP[49,185,186], iii) the alteration of MREs which affects miRNA:mRNA duplexing (Chapter 2) and iv) modifications of pre-mRNA splice sites which has been

reported to originate transcript variants with an altered translational efficiency[187]. In the end, 53 *cis*-acting miR-SNPs were annotated as mediating at least on of these mechanisms. By computing SNP-gene expression associations across different tissue types/populations, it was observed that *cis*-miR-SNPs induce AEI.

Parts of this chapter have been previously published in the following article:

- Arnold M†, **Ellwanger DC**†, Hartsperger ML, Pfeufer A, and Stümpflen V. *Cis*-acting polymorphisms affect complex traits through modifications of microRNA regulation pathways. *PLoS One*, 7(5):e36694, 2012.
  † equal contributors

4.1 Material and Methods

4.1.1 Preparation of the SNP data set

SNPs were obtained from the HapMap Project describing common patterns of human genetic variation (release 22, CEU[1] panel)[188]. Since most SNPs were expected to be silent[169], the set was filtered by SNPs associated to a trait increasing the specificity of the subsequent analysis. For this purpose, the NHGRI catalog of published GWA studies (accessed 2011)[189] was mined for significant ($P < 10^{-5}$) SNP-trait correlations[190]. Further, linkage disequilibrium (LD) patterns were identified using the tool SNAP[191] by computing the correlation between two loci:

$$r = \frac{\pi_{1,2}\pi_{2,1} - \pi_{1,1}\pi_{2,2}}{\sqrt{\pi_{1,.}\pi_{2,.}\pi_{.,1}\pi_{.,2}}} \tag{4.1}$$

where $\pi_{i,j}$ is the frequency of a haplotype for two loci with two alleles each; $\pi_{i,.}$ is the frequency of allele i of the first locus, and $\pi_{.,j}$ is the frequency of the allele j of the second locus[192]. The correlation coefficient r between a GWAS index locus and a proximal (within 500 kb) locus ranges in $[0,1]$, with $r^2 = 1$ denotes perfect LD. Proximal SNPs with $r^2 \geq 0.8$ were extracted. Finally, a set of 5 101 index and 13 783 proximal SNPs was obtained.

4.1.2 Mapping SNPs to the miRNA regulome

All SNPs were mapped on genomic locations of protein-coding genes (NCBI Reference Sequence[193] annotation, genome build NCBI36). Then, regional classes were defined as follows: 1) intergenic, 2) intragenic with its subclasses 2.1), intronic and 2.2) exonic with its subclasses 2.2.1) 5'-UTR, 2.2.2) CDS and 2.2.3) 3'-UTR. For the location enrichment analysis, each SNP was assigned to one of the five endmost classification levels.

To test if SNPs affect miRNA targeting by altering mature miRNA sequences, SNPs were mapped to miRNA genes. Since the annotation of pri-miRNAs is largely unknown, the chromosomal coordinates of sequences encoding miRNA hairpins, i.e. pre-miRNAs,

1 Utah residents with northern and western European ancestry from the CEPH collection

was used (miRBase[15] release 18, genome build GRCH37). For consistency, the GRCH37 coordinates were transformed to the NCBI36 genome assembly using the UCSC liftOver tool[194].

Integration of AGO CLIP-Seq data

To elucidate potential mechanisms affecting miRNA targeting, the human CLIP-Seq libraries of two miRNP proteins, AGO and TNRC6, were used from the starBase database[195]. The available chromosomal coordinates of the CLIP-Seq clusters were converted to the NCBI36 genome build using the UCSC liftOver tool[194] and mapped to transcripts of protein-coding genes according to the NCBI Reference Sequence annotation (RefSeq)[193]. The final set contained 139 254 locations of miRNP binding regions on 24 442 transcripts. The study was processed using the 48% of sites located within a 3'-UTR.

Examination of polyadenylation signals

Chromosomal coordinates of poly(A) signals were obtained from the PolyA DB[196]. Beaudoing *et al.*[197] proposed that poly(A) sites are located $10-30$ nt downstream of the poly(A) signals. Therefore, all trait-associated SNPs located within this segment were determined. Then, a 11 nt long sequence window centered at each SNP was extracted and examined for the most abundant poly(A) signal[197]. Variants were annotated effecting a poly(A) signal if they induce the creation of a new signal sequence or disrupt an existing pattern. SNPs with alleles maintaining the signal character of the sequence, i.e. variations creating another valid signal, were considered as synonymous mutations without any effect on mRNA stability.

Determination of splice sites

The NNSplice algorithm from the Berkeley Drosophila Genome project[198] was applied to predict alterations in transcript splicing. A genomic DNA sequence window of 60 nt centered at the SNP position was used as input. Predicted splice sites with a likelihood greater than 0.5 were retained neglecting cases with marginal changes[199]. The following events were considered: the total loss/gain of a splice site and the increase/decrease of the splice site likelihood. Lost acceptor sites or sites exhibiting an increase/decrease in their likelihood were filtered if they were located between 100 nt upstream and 10 nt downstream

of a reference intron/exon border (according to the NCBI Reference gene annotation[193], human genome build NCBI36). Lost donor sites or sites with an increased/decreased likelihood were retained if they were located between 10 nt up- and downstream of a reference exon/intron border. A gain of a completely new splice site was always kept[199]. Computed alternative spliced transcripts were explored whether they gain intronic or lose exonic miRNP binding sites.

Analysis of RNA structural properties

To account for structural changes, the RNAfold algorithm (version 1.8.5) from the Vienna RNA Package[200] was applied. Here, the complete ensemble of possible RNA conformations was considered. The partition function and the base pairing probability matrix of each 3'-UTR sequence encoding the respective alleles were computed[201]. The matrix row sums determined the pairing score for each nucleotide. A score vector S of length $n = 41$ nt centered at a miRNP:mRNA interaction site was extracted. The linear correlation between the reference S_i and the mutated structure S_j was measured by the Pearson product-moment correlation coefficient[202]:

$$\rho_{S_i,S_j} = \frac{\sum_{k=1}^{n}(S_{i,k} - \mu_{S_i})(S_{j,k} - \mu_{S_j})}{\sigma_{S_i}\sigma_{S_j}(n-1)} \tag{4.2}$$

where μ denotes the arithmetic mean and σ the sample standard deviation.

Since multiple miRNP binding sites were measured for a single transcript, the smallest correlation coefficient was taken for each SNP per transcript, i.e. the strongest effect on RNA folding was selected. To filter the RNA structural ensemble for significant variants, the minimal correlation coefficient was computed for 1 000 random samples obtained from the SNP background set. Based on this distribution, a correlation coefficient of 0.55 was found having a probability of less than 5% for a type I error (Figure B.4). Thus, SNPs inducing a minimal structural correlation coefficient of less than 0.55 between its alleles were filtered.

Identification of altered MREs

For each allele, all sites complementary to a canonical miRNA seed sequence (Chapter 2) were scanned in the 3'-UTRs of protein-coding genes. MREs were filtered if they

were located within a distance of 21 nt to the center of a miRNP interaction site[70]. To additionally reduce the false positive rate, it was required that a miRNA had at least one sequence read in its accordant CLIP-Seq experiment. Further, miRNAs were removed of which target sites were not significantly enriched within miRNP binding segments (log odds ratio > 0, χ^2 test $P < 0.05$). Finally, 258 miRNAs were found of which seed-pairing sites were disrupted and 324 miRNAs of which seed-binding affinity was increased.

Site conservation

To determine the maximum likelihood of a locus to be conserved across species, the algorithm PhastCons from the PHAST package[96] was applied using a whole genome alignment of 17 vertebrates. According to Betel *et al.*[97], a score greater than 0.57 was used to classify a site as conserved in mammals.

4.1.3 Statistical testing with simulated data

To test the significance of the observations, a background set was created. The 2.7 million SNPs from the CEU panel of the joint HapMap Phases I, II and III (release 27) were filtered by entities with an available genotype information[170,188,203]. Official SNP IDs were determined using the tool SNAP[191]. Chromosomal coordinates (genome build NCBI36) were assigned according to the UCSC Table Browser annotation[93]. Samples were generated comparable to the SNP set used for the analysis framework: 1 000 samples with 5 101 index SNPs were drawn with replacement and extended with SNPs in strong LD ($r^2 \geq 0.8$). The analysis pipeline, i.e SNP localization enrichment and identification of mechanisms affecting miRNP function, was conducted for each sample. Then, the cumulative empirical distribution was computed from the resulting values allowing the inference of the probability that the observed value x is stochastic, i.e. the probability to obtain a value $\geq x$ by chance.

4.1.4 Genotype-gene expression survey

To associate SNPs with transcript level changes, the data from Nica *et al.*[204] was used. They provided normalized gene expression profiles of three tissue types (166 adipose samples,156 lymphoblastoid cell line samples and 160 skin samples; Illumina HT-12v3 chip)

from healthy female twins (1/3 monozygotic, 2/3 dizygotic) derived from the MuTHER pilot phase project (Multiple Tissue Human Expression Resource; http://www.muther.ac.uk/). Missing measurements were complemented by data from Stranger *et al.*[205]. They measured the expression of lymphoblastoid cell lines (Illumina HumanWG-6 v2 chip) from 726 HapMap[203] individuals (CEU[1], CHB[2], GIH[3], JPT[4], LWK[5], MEX[6], MKK[7] and YRI[8]). The linkage between allele occupancy and gene expression intensity were scored by the Spearman's rank correlation coefficient, i.e. Pearson correlation between the ranked variables[206]. Statistical significance is assessed by computing a non-parametric *P* by data permutation. Expression values are shuffled between individuals' genotypes and the nominal *P* are recomputed (using a *t*-statistic[207]). A probability distribution is constructed by repeating this procedure 10 000 times under the null hypothesis of no SNP-probe linkage. All calculations were performed by applying Genevar[207], a platform of database and web services designed for the analysis of SNP-gene associations.

The regulatory effect of each SNP was quantified by the coefficient of variation (CV) metric[208–210]:

$$CV_i = 100\frac{\sigma_i}{\mu_i} \tag{4.3}$$

where μ_i denotes the arithmetic mean and σ_i the standard deviation of the expression levels of transcript i across individuals. Please note that the variance, σ_i^2, gives a measure of how far a set of data values is spread out. Here, the rooted variance is used to quantify the amount of variation or dispersion in the set of expression values. Thus, the CV metric has eligible properties to quantify variation of gene expression[210].

1 Northern and Western European ancestry in Utah, US
2 Han Chinese in Beijing, China
3 Gujarati Indians in Houston, US
4 Japanese in Tokyo, Japan
5 Luhya in Webuye, Kenya
6 Mexican ancestry in Los Angeles, US
7 Maasai in Kinyawa, Kenya
8 Yoruba in Ibadan, Nigeria

4.2 Results

4.2.1 Enrichment of SNPs in 3'-UTRs

Previous studies reported that the major fraction of trait-associated SNPs was located at non-coding regions[174]. Since these variants do not impact the gene product directly, it has been suggested that gene expression dynamics are likely affected[175]. The predominant non-coding region bound by miRNPs is the 3'-UTR of target mRNAs. Thus, SNPs impacting gene expression controlled by miRNAs were expected to be found within these segments.

A set of 18 884 SNPs was created filtering the HapMap CEU panel[188] by index SNPs from the catalog of published GWA studies[190] and proximal SNPs in strong LD ($r^2 \geq 0.8$). Each locus was classified by its chromosomal position relative to a protein-coding gene: intergenic, intronic, 5'-UTR, CDS, and 3'-UTR. Analysis of the regional classifications revealed a location bias towards intragenic, and in particular, terminal untranslated regions. 436 SNPs were located in the 3'-UTR of 326 human genes (odds ratio $= 2.3$, χ^2 test $P < 10^{-52}$). The enrichment was also observed for index SNPs only (odds ratio = 2.1, χ^2 test $P < 10^{-10}$). The significance of this observation was validated by calculating the probability of achieving an equal or stronger 3'-UTR enrichment by chance in a sample of equal size ($P = 1.1 \times 10^{-7}$). The robustness of this observation was examined by testing for potential dependencies between the odds ratio and different thresholds for r^2. Adjusting for r^2 during the LD computation showed that the distribution of odds ratios locally stabilizes around the threshold of $r^2 \geq 0.8$ (Figure 4.1A).

In this context, the MAF was analyzed. In general, the SNPs used in this study exhibited a higher MAF, i.e. these were more common than the complete HapMap SNP background − independent of the chromosomal location. However, the MAF distribution of the 3'-UTR SNPs had a slight trend towards moderate frequencies of $0.1 - 0.4$. This trend becomes more pronounced when comparing the 3'-UTR SNPs to polymorphisms located in the other two exonic regions, i.e. 5'-UTR and CDS. The 3'-UTR SNPs were under-represented in the intervals $[0.0, 0.1]$ (odds ratio $= 0.88$), $[0.2, 0.3]$ (odds ratio $= 0.70$) and $[0.4, 0.5]$ (odds ratio $= 0.78$). The other two intervals were significantly ($P < 0.05$) enriched: $[0.1, 0.2]$ (odds ratio $= 1.40$) and $[0.3, 0.4]$ (odds ratio $= 1.59$). At this, it was of particular interest whether the enrichment of trait-associated SNPs in the 3'-UTR compared against the

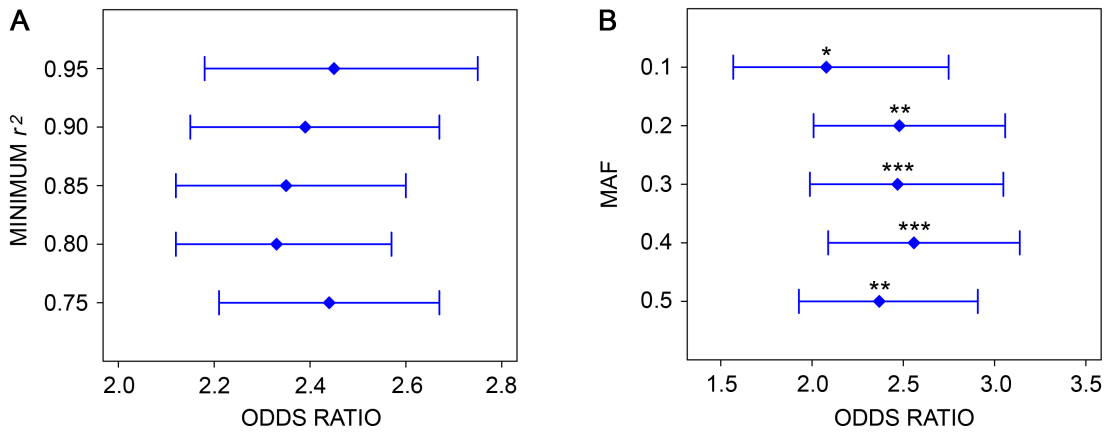


Figure 4.1 | **Statistical analysis of 3'-UTR enrichment values. A** | Enrichment of SNPs in 3'-UTRs of coding genes with respect to different r^2 thresholds during LD computation. Shown are the odds ratios and the confidence intervals for five cut-offs $\leq r^2$. The odds ratio stabilizes at a minimum of 0.8 for accumulative 3'-UTR sets. **B** | Enrichment of SNPs in 3'-UTRs of coding genes with respect to the MAF. Shown are the odds ratios and confidence intervals for five different MAF bins; * $P < 10^{-6}$, ** $P < 10^{-13}$, *** $P < 10^{-15}$.

complete HapMap background is only valid for a specific MAF. Figure 4.1B shows that the odds ratio always reached significance.

4.2.2 3'-UTR SNPs are involved in lipid metabolism

Next, it was assessed whether the 326 genes embedding a 3'-UTR SNP share common characteristics in terms of disease involvement and functional annotation. For this purpose, relevant traits were mapped to MeSH terms (Medical Subject Headings; http://www.ncbi.nlm.nih.gov/mesh/). The most abundant categories found in the 49 disease classes were immune system diseases, mental disorders, digestive system diseases, nervous system diseases, and neoplasms. Notably, the distribution of the 3'-UTR SNPs over these disease classes showed no significant enrichment compared to the number of studies performed for the single disorders in the NHGRI GWA study catalog[190]. Comparing the number of 3'-UTR SNPs per disease to the number of all non-3'-UTR SNPs, lipid concentrations were found to be significantly enriched ($P = 1.3 \times 10^{-3}$).

Testing for enrichment of disease terms using alternative databases, three categories reached statistical significance (Bonferroni corrected $P < 0.05$): dyslipidemia (OMIM;

http://omim.org/), neurological diseases, and infections (Genetic Association Database[211]).

Additional functional gene set enrichment analysis by means of Gene ontology terms[212] revealed four significantly enriched (Bonferroni corrected $P < 0.05$) annotations in this set: lipid metabolism, axon growth, activation of the immune response/inflammation, and regulation of/response to cell signaling.

4.2.3 3'-UTR SNPs affect miRNP binding site features

To elucidate the mechanisms influencing miRNP targeting, the analysis was continued with transcripts featuring both, 3'-UTR SNPs and experimentally determined miRNP binding sites by AGO-bound CLIP-Seq measurements. This data set contained 288 SNPs located at 219 genes of which 409 transcripts were affected. Please note that the analysis is *in silico*, i.e. transcripts were not sequenced *in vivo*, rather SNPs were introduced to reference sequences.

The efficacy of miRNPs to control target mRNAs relies, in a broader sense, on two important features: the target sequence composition, such as the encoding of MREs (Chapter 2) and the local target structure (Chapter 3). To determine, how 3'-UTR SNPs affect miRNA-mediated regulation in *cis*, four potential mechanisms compromising targeting features were examined. Of these three were found in the data of this study (Figure 4.2).

The basic prerequisite for miRNP binding in metazoans is a short perfect match to the coupled miRNA complemented by imperfect matches in close vicinity. This MRE region is called the 'seed' sequence and is considered to be a $6-8$ nt long substring within the first 8 nt at the 5'-end of the miRNA[26]. 22 SNPs (7.6%) were predicted impairing MREs (Table C.1), and 28 SNPs (9.7%) creating new or enhancing (i.e. extending an already existing seed match) MRE sequences (Table C.2). The number of SNPs substituting the MRE of one miRNA by a MRE of another miRNA, amounts to 13 variants. Accordingly, a total of 37 unique SNPs (12.8%) directly impact MREs (impairment $P = 1.3 \times 10^{-2}$, enhancement $P = 8.8 \times 10^{-4}$). Additionally, it was found that only 11% of SNPs enhancing or creating a MRE were conserved across mammals. This was a lower fraction than for SNPs mediating one of the other mechanisms (splicing $= 29\%$, structure $= 29\%$).

The geometry of miRNA target sequences is an important determinant of miRNP binding affinity[49,185,186]. 14 SNPs (4.9%) were predicted to impact the binding of the RISC through significant changes ($P < 0.05$) of the local 3'-UTR structure (Table C.4).

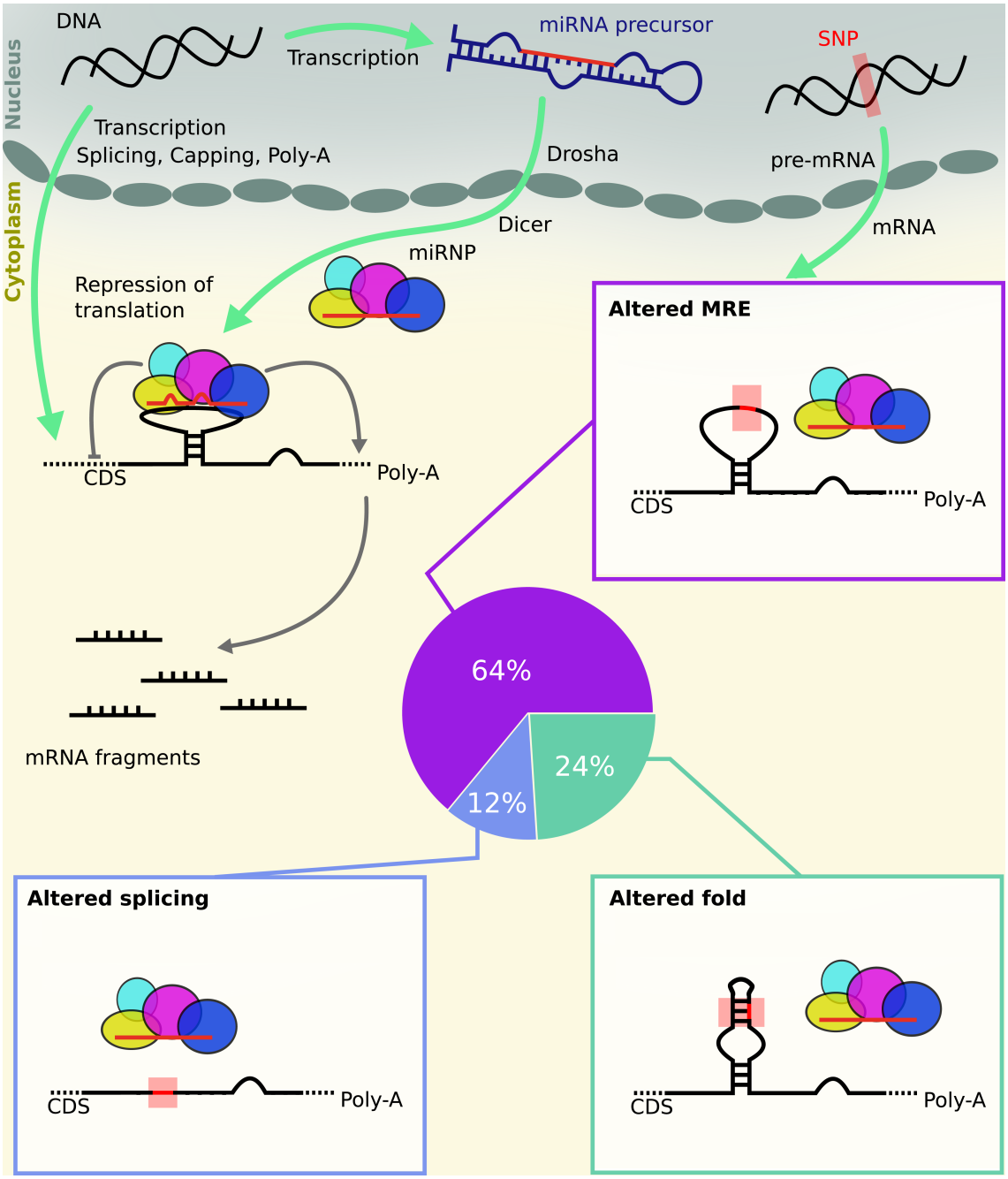


Figure 4.2 | **Mechanisms of 3'-UTR variants impacting miRNA function in *cis*.** Usually protein-coding genes are transcribed to pre-mRNAs which are subsequently maturated by splicing, capping, and polyadenylation; transcripts of miRNA genes are processed by the RNases Drosha and Dicer and incorporated in a protein-complex (miRNP). *(continued on next page)*

Figure 4.2 *(previous page)* | The miRNP binds to characteristic sites at the target mRNA and represses translation or induces degradation. Three *cis*-acting miR-SNP mechanisms perturbing this process were identified: i) 64% of variants caused an altered MRE sequence increasing/decreasing miRNA binding affinity, ii) 24% of SNPs detached miRNP binding sites by alternative splicing, and iii) 12% of variants induced a changed local secondary structure of the miRNA target region.

The alteration of pre-mRNAs by modifying splicing signals has been reported to influence translational efficiency[187]. Seven SNPs (7.4%) were predicted to interfere with RNA splice sites (Table C.3). Of these, six were predicted to create new acceptor sites and one to create a new donor site (acceptor sites $P = 1.8 \times 10^{-2}$, donor sites $P = 1.4 \times 10^{-2}$). In all seven cases, the predicted gain of splice sites results in exon shortening, leading to a noticeable loss (46% on average) of miRNP binding sites in the accordant transcripts. SNPs interfering with splice sites located at an exon/intron or intron/exon junction (as annotated in NCBI RefSeq[193]) were not observed.

The loss of 3'-poly(A) signals has been described to cause genetic diseases by cap-dependent miRNA-mediated degradation of the mRNA[183,184]. Four SNPs affecting hexamers with a sequence characteristic for poly(A) signals were identified. However, none of these hexamers was located in close proximity to a known poly(A) site.

In total, 53 3'-UTR SNPs mediating at least one mechanism impacting miRNA function were found (Table 4.1). Of note, 14 of these were index SNPs as reported in the NHGRI GWAS catalog (rs7097, rs7119, rs9253, rs12916, rs1379659, rs2071518, rs2244967, rs2282301, rs2564921, rs4770433, rs4819388, rs4973768, rs7528419, and rs11542478). In the following, SNPs affecting miRNA-mediated regulation in *cis* are referred as '*cis*-miR-SNPs'.

Table 4.1 | ***Cis*-miR-SNPs.** Listed are the SNP IDs (* denotes index SNPs), the affected gene, the type of MRE affection (disruption: $-$, creation: $+$, both: $\pm$), the occurence RNA folding impairments, the type of splice site affection (acceptor gain: Acc+, donor gain: Don+), the SNP-gene expression linkage score from two cohorts ($\rho > 0$: mRNA increase, $\rho < 0$: mRNA decrease) and the associated trait. Missing values in the MuTHER study were substituted by HapMap3 measurements (footnoted is the reference population and all populations with the same score sign).

SNP	Gene	MRE	Folding	Splicing	ρ_1	ρ_2	Disease/Trait
rs1121	PDXDC1	$+$			-0.13	-0.11	Height
rs4564	DLD	$-$			0.05	0.06	Ulcerative colitis
rs6706	TRIP6	$-$			-0.07	0.06	Resting heart rate
rs7089	TMUB2	$\pm$			-0.52	-0.63	Bone density
rs7097*	POLR1D	$+$			-0.34	-0.39	Large B-cell lymphoma
rs7118	ZFP90	$\pm$			0.42	0.43	Ulcerative colitis
rs7119*	HMG20A	$-$			-0.05	0.09	Type 2 diabetes
rs7371	GNAI3			Acc+	-0.47	-0.33	Major depressive disorder
rs7444	UBE2L3		✓		0.69	[a]	Crohn's disease
rs7444	UBE2L3		✓		0.69	[a]	Lupus
rs8523	ELOVL2	$\pm$			-0.27	-0.15	Phospholipid levels
rs9253*	MEAF6	$-$			N/A	N/A	Hematological phenotypes
rs9927	PYGB	$+$			-0.06	0.11	Liver enzyme levels
rs10923	SMC4	$\pm$			0.19	0.26	Primary biliary cirrhosis
rs11700	E2F4	$+$			0.16	0.11	Coronary heart disease
rs12439	CLIC4	$\pm$			-0.21	-0.08	Height
rs12916*	HMGCR	$+$			-0.21	-0.03	Cholesterol levels
rs12916*	HMGCR	$+$			-0.21	-0.03	Metabolic traits
rs12956	RYBP		✓		0.03	-0.10	Height
rs13099	TMED10		✓		0.22	[b]	Height
rs42038	CDK6		✓	Acc+	0.11	[c]	Height
rs42039	CDK6	$+$			-0.01	-0.10	Rheumatoid arthritis
rs232775	MYSM1	$+$			-0.05	-0.13	Diabetic retinopathy
rs699779	NOTCH2	$-$		Acc+	0.18	0.14	Type 2 diabetes
rs823136	RAB7L1	$+$			-0.17	[d]	Parkinson's disease
rs835575	NOTCH2	$\pm$	✓		-0.23	-0.12	Type 2 diabetes
rs835576	NOTCH2	$\pm$			-0.23	-0.12	Type 2 diabetes
rs1045100	ATG16L1	$\pm$			-0.24	-0.18	Crohn's disease

[a] JPT; CEU, CHB, GIH, LWK, MEX, MKK, YRI.
[b] LWK; CEU, CHB, GIH, JPT, MKK, YRI.
[c] LWK; CEU, GIH, MKK.
[d] CHB; GIH, LWK, MEX.
[e] LWK; CHB, MEX, MKK, YRI.

Table 4.1 *(continued)*

SNP	Gene	MRE	Folding	Splicing	ρ_1	ρ_2	Trait
rs1045407	ZNF678	+	✓		0.22	[e]	Height
rs1046917	FN3KRP		✓		0.42	0.28	HbA1c levels
rs1047440	CEP120	±			-0.21	-0.11	Body mass index
rs1058588	VAMP8	−			-0.44	-0.30	Prostate cancer
rs1379659*	SLIT2	−			-0.02	0.17	Echocardiographic traits
rs2032933	RMI2	+			-0.32	-0.14	Celiac disease
rs2071518*	NOV	+			-0.03	-0.17	Blood pressure
rs2077579	DDX6		✓		-0.36	[f]	Primary biliary cirrhosis
rs2229302	HOXB2	±			-0.31	-0.33	Primary tooth developm.
rs2244967*	VSTM4			Acc+	0.41	[g]	Serum uric acid
rs2282301*	RIT1		✓		0.13	[h]	Conduct disorder
rs2293578	SLC39A13	+			0.04	-0.07	Body mass index
rs2564921*	RTF1		✓		-0.16	[i]	Height
rs3816661	CD276	−			-0.08	0.17	Liver enzyme levels
rs3821301	TANC1		✓		0.20	-0.06	Sudden cardiac arrest
rs4770433*	SACS	−			0.19	-0.06	Protein QTL
rs4819388*	ICOSLG	+	✓		-0.36	-0.03	Celiac disease
rs4973768*	SLC4A7			Don+	0.13	0.08	Breast cancer
rs6722332	WDR12			Acc+	0.19	0.02	Coronary heart disease
rs6722332	WDR12			Acc+	0.19	0.02	Myocardial infarction
rs7350928	KIAA1267	±			-0.28	-0.17	Parkinson's disease
rs7528419*	CELSR2			Acc+	0.19	0.01	Cardioviscular disease
rs7528419*	CELSR2			Acc+	0.19	0.01	Cholesterol levels
rs7528419*	CELSR2			Acc+	0.19	0.01	Metabolic traits
rs7528419*	CELSR2			Acc+	0.19	0.01	Myocardial infarction
rs7528419*	CELSR2			Acc+	0.19	0.01	Response to statins
rs8176751	ABO	+			0.14	0.12	Hematolog. phenotypes
rs10892082	PAFAH1B2		✓		0.04	0.23	Protein QTL
rs10892082	PAFAH1B2		✓		0.04	0.23	Triglyceride levels
rs11067231	MMAB	+			-0.44	-0.49	Cholesterol levels
rs11542478*	FAM110C		✓		-0.15	0.01	Inform. processing speed
rs11713355	SLC6A6	±			-0.05	-0.18	Cognitive performance
rs17574361	KIAA1267	±			-0.28	-0.17	Parkinson's disease

[e] LWK; CHB, MEX, MKK, YRI.
[f] JPT; CHB, GIH, MEX.
[g] YRI; CEU, GIH, MEX.
[h] YRI; JPT, LWK, MEX
[i] CEU; GIH, MKK.

4.2.4 *Cis*-miR-SNPs exhibit allelic expression imbalance

Since *cis*-miR-SNPs were predicted to change miRNP binding specificity, the question was examined whether these variants imply turnover of miRNA targeting accounting for AEI of miRNA target genes. For this purpose, gene expression profiles from genotyped individuals were collected (twin samples from the MuTHER pilot project quantified by Nica *et al.* 2011 [204]; missing measurements were filled by samples of HapMap individuals quantified by Stranger *et al.* 2012 [205]). The linkage between *cis*-miR-SNPs (allelic occupancy) and target gene expression was conducted using Spearman's rank correlation and scored by its coefficient ρ.

The canonical model of miRNA-mediated regulation postulates that miRNPs repress expression of target mRNAs [27]. Under this simple model, 72% of all SNP-mRNA associations were found being in accordance with the expected differential expression for each allelic occupancy in at least two cohorts: i) increased mRNA levels ($\rho > 0$) induced by miRNA regulation loss due to impairment of MREs or RNA folding, or alternative spliced exons, and ii) decreased mRNA levels ($\rho < 0$) caused by miRNA regulation gain due to enhanced MRE affinity or RNA structures attracting miRNP binding (Table 4.1). The statistical significance of the associations was assessed using a permutation approach. Here, 50% of all *cis*-miR-SNPs reached a significance level of $P < 0.08$ in at least one cohort. Of these, two affected splice sites, nine altered the target structure and 18 *cis*-miR-SNPs manipulated MREs. These results suggest that the *cis*-miR-SNP mechanisms presented in this study are coupled with AEI.

However, the correlation coefficient does not quantify the extent of expression variation induced by *cis*-miR-SNPs. For this purpose, the coefficient of expression variation (CV) was employed. This metric has been previously well justified for the analysis of variance in expression profiles [208–210]. In line with the SNP-expression linkage analysis above, CV values were conducted for each transcript across the subset of individuals from the MuTHER pilot expression study provided by Nica *et al.* [204].

The expression variance of transcripts with predicted *cis*-miR-SNPs ranged from 0.95 to 5.18 with a median of 2.22 (Figure 4.3A). Since the turnover of miRNA-mediate regulation has been reported to be decent (Chapter 1.4.3), the variance in expression levels was expected to be moderate. However, the observed effect was significant higher ($\sim 61\%$) than expected by chance from the whole background distribution (Kolmogorov-Smirnov

test $P = 9.3 \times 10^{-07}$). The strongest variation was induced by altered splice signals (median = 2.75), followed by affected MREs (median = 2.16) and local folding (median = 2.15).

Next, the effect of each *cis*-miR-SNP on target variance was examined individually. Here, the cumulative distribution function of the CV values was computed. This allowed to estimate the location of each *cis*-miR-SNP CV value in the background distribution. In other words, the fraction of genes exhibiting a lower expression variation than the *cis*-miR-SNP affected gene was estimated (Figure 4.3B). 86% of variants exhibited an allelic expression variance beyond the average. Of these 27% had a higher CV than 90% of the background. Therefore, the CV analysis suggest that the mechanisms conducted by the *cis*-miRNA-SNPs are coupled with increased gene expression variability.

Since the set of *cis*-miRNA-SNPs was composed of either index SNPs from the NHGRI GWAS catalog or proximal SNPs in strong LD, the respective trait associations were available. This raised the question whether the found *cis*-miRNA-SNPs may be reasonable in the context of their predicted phenotype. For this purpose, for each mechanism an example was examined by literature research.

As a first example, the *cis*-miRNA-SNP rs11067231 was found associated with perturbed cholesterol levels. This SNP is located at the chromosome 12q24 region which was associated with high-density lipoprotein-cholesterol[213]. Its G>T nucleotide change at the reference genome sequence originates a 6merγ binding site for miR-624 at the 3'-UTR of the transferase MMAB. Fogarty *et al.*[213] observed a significant change in relative allelic expression of MMAB whereat the rs11067231 allele with lower high-density lipoprotein-cholesterol correlated with higher MMAB transcript abundance. They suggested that MMAB may be the most likely gene influencing high-density lipoprotein-cholesterol levels. An AEI was confirmed by the MuTHER study: the rs11067231 allele creating the MRE for miR-624 was significantly linked with reduced MMAB transcript levels ($\rho_1 = -0.438$, $P_1 = 9.4 \times 10^{-5}$, $\rho_2 = -0.488$, $P_2 = 3.2 \times 10^{-6}$).

Secondly, the SNP rs10923 was found in high LD ($r^2 = 0.86$) with the primary biliary cirrhosis GWA-associated risk loci rs4679904[214]. The minor G allele of SNP rs10923 disrupts the MRE of miR-299-5p located at the 3'-UTR of SMC4. This miRNA has been reported to be upregulated in patients suffering from primary biliary cirrhosis[215]. A considerable AEI was observed by the MuTHER study exhibiting increased SMC4 transcript levels for the G allele ($\rho_1 = -0.191$, $P_1 = 1.0 \times 10^{-1}$, $\rho_2 = -0.264$, $P_2 =$

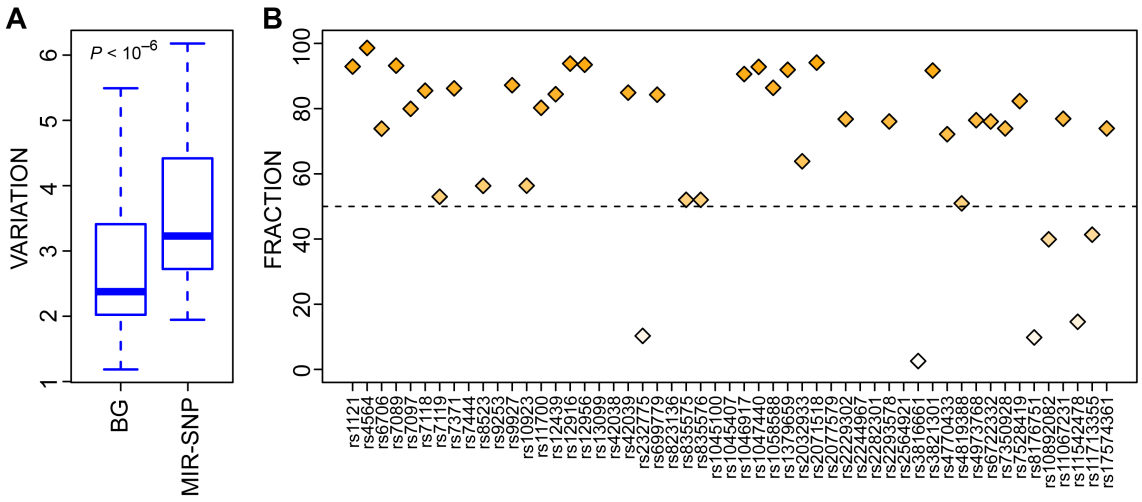


Figure 4.3 | **Expression variance induced by *cis*-miR-SNPs. A** | Boxplot of the distributions of the CV values of all transcripts (BG) and the transcripts predicted to be affected by a *cis*-miR-SNPs. *P* is given by the Kolmogorov-Smirnov test and indicates the significance level of the difference between the distributions. **B** | For each *cis*-miR-SNPs, the fraction of all measured transcripts holding a lower expression variance than the affected transcript are shown. The node coloration denotes a higher location of the observed CV in the empirical cumulative CV distribution of all transcripts, respectively.

1.8×10^{-2}). SMC4 is part of the condensin I complex[216] which has been suggested to be involved in the single-strand break repair mechanism by complexing with PARP1 and XRCC1[217] – two genes which were observed overexpressed in active cirrhosis[218].

Thirdly, the SNP rs1046917 was associated with altered glycated hemoglobin levels which are reflecting the effective plasma glucose concentration. Structure analysis revealed an altered folding of the FN3KRP 3'-UTR induced by the A>G nucleotide change at the reference genome sequence. The mutated RNA conformation decreases the affinity of the miRNP to bind to the target region. Consistently, an increased expression of FN3KRP induced by rs1046917 was observed by the MuTHER study ($\rho_1 = 0.421$, $P_1 = 1.0 \times 10^{-4}$, $\rho_2 = 0.275$, $P_2 = 9.9 \times 10^{-3}$). FN3KRP is assumed to protect proteins from nonenzymatic glycation to restore their function[219]. However, its specific physiological function remains largely unknown. Its protein sequence is highly similar to its neighboring enzyme FN3K which has been described to affect the glycation level at specific sites of haemoglobin[220]. Conceivably, an increased FN3KRP activity can be suggested to result in reduced amounts

of glucose circulating in the blood plasma and to account for the etiology of hypoglycaemia.

Lastly, the final product of purine degradation is uric acid, a compound of which elevated serum concentrations have been associated with several diseases such as gout[221], hypertension, and disorders of the cardiovascular system[222] in previous epidemiological studies. At this, it is very likely that immunoglobulins facilitate the crystallization of uric acid precipitating monosodium urate cristals out of serum[223]. The A allele of the serum urate associated SNP rs2244967 was predicted to generate a new acceptor splice site which shortens the mature 3'-UTR of VSTM4 causing a loss of 50% of all miRNP binding sites. Consistently, an AEI of VSTM4 comprising a higher expression level for the A allele was found ($\rho_1 = 0.409$, $P_1 = 1.2 \times 10^{-5}$, $\rho_2 = 0.176$, $P_2 = 6.7 \times 10^{-2}$). The function of VSTM4 is still unknown. However, its protein embeds an immunoglobulin-like domain (V-set domain) which can be also found in immunoglobulin light and heavy chains. Since it has been demonstrated that all three major classes of immunoglobulin have an affinity to the surface of monosodium urate crystals[223], it can be suggested that VSTM4 may also contribute to the crystallization of uric acid. Thus, increased activity of VSTM4 contributes to raised monosodium urate crystal levels which in turn triggers the phagocytosis of these crystals by neutrophils. This mechanism has been reported to affect the volume of urinary excretion of uric acid, and consequently, the level of serum uric acid. Interestingly, hyperuricemia has been associated to the etiology and inflammatory attacks of gout[224].

4.3 Conclusion

In the previous chapters of this thesis (Chapter 2 and Chapter 3), relevant characteristics of miRNP binding sites were analyzed. Since our current knowledge on potential effects on miRNA target selection by SNPs is limited, the impact of genetic variation on these features was of particular interest in this study. A set of human SNPs was generated composed of trait-associated index and proximal variants in strong LD. By analyzing their genomic position, a significant enrichment of loci in the 3'-UTR of protein-coding genes was observed. Since the 3'-UTR of the mRNA is the major host for *cis*-regulatory elements of miRNA regulation, it was investigated whether 3'-UTR variants impact post-transcriptional regulation. Experimentally determined miRNP binding sites by the CLIP-Seq protocol

were mapped to affected transcripts. Subsequently, the potential mechanistic model of a single nucleotide mutation on miRNP targeting was determined. It is of note that the used sequences were reference transcripts as provided by the RefSeq database[193]. As such, the presented mechanisms were *in silico* predictions requiring further experimental validations, such as *in vivo* RNA sequencing.

The computational analysis revealed that, in the majority of cases, SNPs created/disrupted MRE sequences or enhanced/damped their complementarity to miRNA sequences. Although using alternative MRE definitions, this finding confirmed previous studies[225,226]. Also a third novel scenario with a less straightforward rationale was observed: the substitution of the MRE of one miRNA by the MRE of another miRNA. Such a mutation may constitute concurrent but simultaneously diverging effects in different cell types, depending on the respective expression patterns of the two miRNAs. The fraction of 'regulator switches' denotes one third of the whole set of variants affecting MREs which is a surprisingly high number. This result proposes that the transcripts embedding a 'regulator switch' may represent rather interesting subjects for further studies.

The second most abundant effect was the alteration of the mRNA secondary structure leading to a modified local conformation of the miRNP binding region. Previous studies reported the importance of the RNA folding on the binding affinity of the miRNP complex[49,185]. The results of the presented study suggest that single nucleotide mutations may lead to considerable impacts on the miRNA regulatory network. By the example of FN3KRP, its effect was hypothesized in the context of the etiology of hypoglycaemia. However, the extent to which this mechanism translates into the development of traits or diseases remains unknown. In fact, further investigations are required to shed more light on this particular mechanism. It should be noted that a simultaneous work of Haas *et al.*[186] confirmed an interrelation between disease-associated 3'-UTR variants causing mRNA structural change and perturbed miRNA targeting.

The third mechanism effects splice signals. This event has been well described in the context of altered composition of amino acids and domains in proteins and was linked to disease susceptibility[199,227]. The splicing machinery is confronted with multiple attributes that guide the recognition of exon-intron boundaries. At this, the sequences in splice junctions are of particular importance as mutations alter the recognition efficiency of splicing factors[228]. In the presented study, not the CDS, rather the 3'-UTR was investigated for SNP induced alternative splicing. In average, it was predicted that 46%

of miRNP binding sites got lost due to exon shortening, entailing the escape of the target gene from translational repression by miRNAs. Although coherent in the biological sense, it should be noted again that this observation relies on a computational prediction. The expression of miRNAs and target transcripts is, in fact, condition-specific. Novel high-throughput technologies, such as RNA sequencing, will reveal to what extent alternative transcripts are processed that evade miRNA regulation. Further, alternative splicing may also lead to unstable transcripts. In example, rs7371 was predicted to induce a shortened GNAI3 3'-UTR due to creating a novel splicing acceptor site. However, the variant allele exhibited a lowered transcript concentration (Table 4.1). As the affected region was predicted to be conserved across mammals, its deficiency may impair mRNA stability or translocation.

Another mechanism which was not investigated in this study, but expected to be also relevant, is the interference of miRNA biogenesis. Changing the miRNA seed sequence, or impairing the function of a specific miRNA will change its specific regulome and, as such, have drastic global effects by rewiring gene regulatory networks. The data of the presented study contained no conclusive evidence for variations in miRNA genes, and as such, direct perturbation of miRNA processing and function in *trans*. Only one SNP (rs2168518) was found located within the hairpin transcript of miR-4513. The usage of a more sensitive data basis, e.g. the recent 1 000 Genomes Project[229] and an upcoming miRBase release, will reveal likely further SNPs in known miRNA genes. Also the genotyped SNPs may not had significant MAFs to be included in the used panel. However, based on the observations in this study, it can be suggested that miRNA-mediated regulation is primarily affected by common *cis*-miR-SNPs.

The validity of the presented *cis*-miR-SNPs mechanisms was corroborated by an eQTL analysis. The SNP-gene expression linkage was conducted for affected transcripts by computing the correlation between the two variables. Additionally, the respective *cis*-miR-SNP induced variance in the expression levels was measured. For all three classes of *cis*-miR-SNPs, significant AEIs of affected genes were observed. Although expected to be moderate, the effect of *cis*-miR-SNPs on transcript level variance was considerable high. For the following reasons, this observation was striking. While it has been described that miRNAs mainly acts as fine-tuners of target gene expression (Chapter 1.4.3) several other factors may interfere *cis*-miR-SNPs induced effects. First, as described in Chapter 2 and Chapter 3, regulatory efficiency relies on the complex interaction of various target site

features. The change of a single feature may not be expected to impact miRNP binding specificity. Second, human mRNAs contain multiple target sites, for the same as well as distinct miRNAs, which may attenuate the impact due to the loss of a specific MRE. Third, miRNA and mRNA expression levels change under varying environments. Thus, unlike non-synonymous coding variants consistently alter the amino acid sequence of a protein, *cis*-miR-SNPs may show condition-specific effects.

On the example of four *cis*-miR-SNPs, the potential impact on the phenotype was discussed. Here, it should be noted that the primary focus of this work was to investigate the potential genetic mechanisms affecting miRNA-mediate regulation. However, several hundred GWA studies resulted in a huge number of reported SNP-trait associations but only a small fraction revealed a functional explanation; the majority of SNPs are far from being proven causal variants of disease incidence[230,231]. It has been suggested that non-coding variants constitute the main fraction of SNPs identified by GWA studies[174]. Interpreting the identified *cis*-miR-SNPs in the context of the associated phenotype showed the potential of the presented results to generate novel hypothesis. It remains elucidated if these variants are, in fact, causal. Please, also consider that GWA studies are based on the assumption that complex traits are caused by multiple loci with low effects on the phenotype which exhibit alleles that are quite common in the population [1]. Although no direct experimental validation was provided in this study, the most suspicious *cis*-miR-SNPs merit further detailed investigation. This study also suggests that it may be inevitable to overcome the current examination bias of GWA studies towards the coding sequence.

1 The heavily discussed 'common disease, common variant hypothesis'[171,172].

CHAPTER 5

Global modeling of miRNA-mediated regulation

Cellular processes are programed through regulatory control and are conditionally modulated. Gene expression is a highly regulated mechanism that has a profound impact on crucial processes such as cell division, differentiation and apoptosis. Its malfunction can lead to the pathogenesis of fatal diseases[232,233]. The regulation of gene expression covers a number of sequential processes controlling the RNA concentration of TGs selectively regulating the quantity of gene products in the cell. Transcriptional regulation is controlled through proteins called TFs. Combinatorial interactions of RNA-binding proteins and non-coding RNAs with regulatory elements located on target RNA molecules determine the functional outcome of target RNA processing, such as splicing, polyadenylation, export, stability and translation[234]. At this, the miRNA family has attracted a lot of attention. Integrated within a multiprotein complex (miRNP) they bind to target sites preferably located in the 3'-UTR[235] or the coding sequence[236] of mRNAs to govern stability and translational efficiency. Post-transcriptional regulation by miRNAs is an essential regulation layer for higher eukaryotes. One miRNA is able to regulate a large number of protein-coding genes and *vice versa* one mRNA can be regulated by several miRNAs. By intertwining with transcriptional GRNs, miRNA regulation induces extensive interacting control structures. Both types of regulator genes (RGs), namely TFs and miRNAs, span a global GRN that controls thousands of mammalian TGs and forms multilayer regulatory circuits[62].

Novel technologies promote the ongoing transformation of biology from a data-poor to an increasingly data-rich science. The attendant increase in the number, size and diversity of data sources features knowledge for both, TF:TG and miRNA:TG interactions.

The integration of this information offers unprecedented and as yet, largely unrealized opportunities for discoveries from the analysis of large-scale GRNs. However, each data source has its unique bias and inherent potential drawbacks. Sequence-based predictions are rather exhaustive but yield a significant fraction of false positives due to the limited comprehension of the molecular basis of the regulator:target pairing process. Databases with experimentally verified data and high-profile studies provide an impressive amount of information but are far from complete. The biomedical literature is rich in known regulatory interactions but these are difficult to extract. All biological data sources naturally exhibit semantic differences that are caused by varying levels of granularity or abstraction at which objects and their relationships are described. These aspects illustrate the potential and importance of sophisticated data-driven integration approaches.

The fact that integrated networks contain regulatory interactions that were described under varying conditions makes these GRNs comprehensive, but also unspecific and static; also the regulatory sign (stimulation/repression) of potential relations is largely unknown. Since transcriptional and miRNA-mediated post-transcriptional regulation is context-dependent, it is evident that static GRNs are not sufficient to represent regulatory interactions taking place under changing conditions. Modeling condition-specific GRNs using prior information from integrated networks aims to overcome these problems and will facilitate a better understanding on how gene expression is modulated.

With rapidly increasing amounts of gene expression profiles, an exhaustive insight into their underlying large-scale condition-specific GRNs becomes feasible and attractive. Therefore, the method COGERE (modeling of COndition-specific GEne REgulation; from the Latin 'to collect') was developed, an approach to infer condition-specific gene regulation from gene expression data integrating existing knowledge of regulatory interactions. This approach enables the interpretation of multi-dimensional expression profiles reflecting the dynamic interplay of thousands of cellular components in the context of known regulatory relations. A data structure of transcriptional and miRNA-mediated gene regulation (prior model) was build by integrating automatically and manually mined interactions from all available biomedical text with information from relevant databases, recent studies and computational predictions from sequence data. In addition to an increased sensitivity, COGERE is able to suggest references for inferred interactions that were described in the literature. This will facilitate the generation of novel, testable hypotheses. To compute the condition-specific strength of association from gene expression data, COGERE

uses a two-way nonlinear non-parametric ANOVA considering prior information. This association metric overcomes the disadvantages of common approaches utilizing linear correlation[237–239] and mutual information[239,240]. Linear correlation requires miRNA and mRNA expression profiles to be obtained from the same set of individuals (matched data), and inherently detects only linear relations. Mutual information needs careful discretization of the expression data to avoid loss of signal and, in addition, is non-negative, and as such does not provide information about the condition-specific sign of interaction.

In this chapter, the construction of the COGERE framework is presented and it is shown that this approach significantly improves existing methods for the large-scale modeling of miRNA-mediated condition-specific GRNs. Further, the utility of COGERE is demonstrated by inferring a cancer-specific regulatory network from the NCI-60[241] microarray project.

Major parts of this chapter have been previously published in the following article:

- **Ellwanger DC**, Leonhardt JF, and Mewes HW. Large-scale modeling of condition-specific gene regulatory networks by information integration and inference. *Nucleic Acids Res.*, Oct 7, 2014.

The results of this chapter have been presented at the following scientific conference:

⋆ **Ellwanger DC**, Leonhardt JF, and Mewes HW. Large-scale modeling of condition-specific gene regulatory networks by information integration and inference. *Workshop Computational Biology @ Bayer* (Boston, USA), 2014.

⋆ **Ellwanger DC**, Leonhardt JF, and Mewes HW. COGERE: modeling of condition-specific gene regulation and regulator gene centrality by information integration and inference. *International Conference on Systems biology* (Copenhagen, Denmark), 2013.

5.1 Related work

The combination of sequence-based target predictions with high-throughput experimental data improves the prediction accuracy by reducing the false positive rate[242]. Therefore, a lot of effort has been put to integrate computational tools and expression data. The mirAct web server[243] enables the determination of condition-specific miRNA activity based on a three-step procedure. First, miRNA targets are determined by a sequence-based target prediction algorithm. For each TG its expression values are transformed to the Z-score across all samples or the average of its ranks within each sample and across all samples. Second, for each miRNA a sample score is computed using either t-statistics (in the case of Z-score transformation) or the difference of the average ranks between targets and non-targets. Finally, for each miRNA the null hypothesis is tested that all conditions have identical sample scores (Kruskal-Wallis test[244], Jonckheere-Terpstra trend test[244]). If the null hypothesis can be rejected, a miRNA is predicted to have a condition-specific effect. mirAct computes miRNA activity based on the expression of its potential targets across several conditions. Apperantly, by using this approach, it is not feasible to explicitly identify each condition-specific miRNA:TG interaction. This can be obtained by the widely used MATLAB tool GenMir++[245]. It implements a complex Bayesian framework to calculate the posterior probability that a candidate interaction is likely to have participated in degrading the TG transcript given the observed patterns of miRNA and mRNA expression. A linear function formulates the expression of a TG as being negatively shifted with respect to a background level of expression due to the regulatory effects of its candidate targeting miRNAs. The GenMir++ model incorporates parameters accounting for differences in the regulatory potential of miRNAs, varying hybridization conditions, and normalization between the expression data sets. To learn the parameters, GenMir++ applies the variational Bayesian algorithm, an Expectation Maximization method which may be extremely slow and computational inefficient for a large number of genes (convergence rate highly depends on the priors and the likelihood) and approximates a factorized variational posterior which may diverge from the real posterior. Since the Bayesian model requires sequence-specific scores of predicted miRNA target sites, it is impractical to integrate prior information from experimental databases. In contrast, the very prominent TaLasso web server[246] enables the usage of candidate

interactions composed of the union or intersection of several sources neglecting the sequence-based interaction score. To infer regulation, a linear relationship between mRNA and miRNA is assumed and approximated by LASSO (least absolute shrinkage and selection operator) regression, an alternative regularized version of least squares. TaLasso has been shown to outperform GenMiR++ in some cases[246]. For a complete review on this topic please refer to Muniategui *et al.*[242].

The major drawback of these tools is their limitation to miRNA interactions. Several studies observed a significant co-regulation between the transcriptional and post-transcriptional layer[62,247,248]. Its importance has been emphasized by Chen *et al.* [249] who suggested a close relation between the perturbation of co-regulation and carcinogenesis. Thus, to get a more comprehensible insight into the condition-specific regulatory landscape, it is crucial to model the combined regulation of both, miRNAs and TFs. This facilitates the elucidation of direct, indirect and co-regulatory mechanisms. However, the large-scale modeling of miRNA- and TF-mediated regulatory systems is still in its infancy. The combination of transcriptional and post-transcriptional regulation is challenging as they involve not only RG:TG pairs, but also the interactions between the regulators themselves. A common procedure is to filter a set of potential interactions by means of differential gene expression, i.e. retain only significantly up- and downregulated RGs and TGs[250]. Besides the fact that this approach omits potential significant correlations between non-differential expressed regulators and their targets, it is also only applicable for case-control study setups. To find a remedy, two prominent approaches were developed: mirConnX[237], MAGIA[238] and its update MAGIA2[239]. Both approaches model GRNs by superimposing an integrated network and inferred interactions from expression data. The integrated network is composed of predicted and experimental verified RG:TG pairs; the inferred network is composed of linear correlation (Pearson, Spearman or Kendall[251]) between each RG:TG pair across all conditions. Both tools apply a simple integration function neglecting individual prediction scores. The utilization of a linear correlation coefficient requires that miRNA and mRNA expression profiles have to be obtained from the same set of individuals (matched data). Further, it has been suggested that linear associations between regulator and target is a weak indicator of true condition-specific regulatory relationships in real biological data[252]. Please note that MAGIA2 also provides the option to use the non-linear mutual information as metric of association. This measure requires a careful discretization of the expression data to avoid a loss of signal and predicts only

unsigned interactions.

5.2 Material and Methods

COGERE maps regulatory complexity by reconstructing GRNs involving TFs or miRNAs as regulators (Figure 5.1A). The workflow of COGERE is outlined in Figure 5.1B. In the following each step of the framework (information integration, inference), the evaluation and the data analysis of the use-case are described in detail.

5.2.1 Construction of the prior model by information integration

The prior network was composed of *in vivo*, *in vitro* and computationally determined regulator:target interactions. Several heterogeneous data sources were combined to a single directed, weighted graph data model $G = (V, A, W)$. All genes with their symbols, gene synonyms and IDs as listed in NCBI Entrez Gene[253] and miRBase version 19[254] were added as vertices $v \in V$ to the regulatory graph. Regulatory associations were stored as directed interactions between two gene nodes. Each interaction was weighted by a prior score that ranks its regulatory potential with $\text{prior} \in W : A \rightarrow \mathbb{R}$; the weights were stored in a weight matrix $W \in \mathbb{R}^{|V| \times |V|}$ where $\text{prior}_{i,j}$ denotes the weight of the interaction between v_i and v_j. Note that the final weight matrix was scaled to $[0, 1]$ and G may contain self-loops. In the following, it is specified how the weight matrix is computed from the integrated evidences.

Integration of transcriptional regulatory interactions

To predict transcriptional regulatory associations, human and murine promoter sequences of protein-coding genes were obtained from the ElDorado database version 08-2011 (ElDorado; http://www.genomatix.de). For miRNA genes, promoters were collected from Fujita *et al.*[255] and CoVote[256] and transcriptional starts from CoreBoost_HM[257], Corcoran *et al.*[258], Marson *et al.*[259], Ozsolak *et al.*[21], miRStart[260], and Eponine-TSS[261]. Given a median promoter length of 448 nt in the study of Fujita *et al.*[255] and 350 nt predicted by CoVote, adequate promoter sequences from 500 nt upstream to 100 nt downstream relative to a transcriptional start site were extracted. Chromosomal locations of

all miRNA hairpins were obtained from miRBase. The distances between a miRNA hairpin start position and all promoter start positions were calculated. For each miRNA gene, the promoter located closest to its hairpin sequence was selected. If miRNA genes shared the same promoter and had an inter-gene distance of up to 50 kb, they were proposed to form a transcriptional unit[19]. Promoters located up to 50 kb upstream of a miRNA gene or a transcriptional unit[20] were filtered. Additional promoter regions of intragenic miRNAs located on the same strand and within an intron of a protein-coding gene were considered coincident with the one defined for the host gene[23]. Gene annotations were obtained from Ensembl[118]. All promoter sequences were scanned for vertebrate TF matrix matches using the MatInspector algorithm (matrix family library version 8.4)[262]. ModelInspector (module library version 5.5)[263] was utilized to filter experimentally verified vertebrate modules of transcriptional regulatory units, functional composite elements consisting of at least two TF-binding sites in conserved order and distance. PhastCons[96] scores from 46-way (human) and 30-way (mouse) alignments of vertebrates available through the UCSC Table Browser[93] were used to calculate mean conservation levels of potential TF-binding sites. Each candidate target site was required to correspond to the most conserved nucleotide at 95% of all positions of the TF matrix or to be conserved with an average score of at least 95%. Moreover, all regulatory interactions contained in the literature were extracted by the text-mining tool BioContext[264]. It was required that the associations of two biological entities were organism-specific and the interaction type was included in the set of terms: regulation, positive regulation and negative regulation. BioContext provides a score for each event mirroring the precision of the identified association based on specific event features. This allowed for each TF:target interaction the computation of the prior score based on the TF matrix similarity score $x_{\text{similarity}}$, the conservation score $x_{\text{conservation}}$ and the text-mining score $x_{\text{literature}}$ as follows:

$$\text{prior} = x_{\text{similarity}} + x_{\text{conservation}} + x_{\text{literature}} \tag{5.1}$$

at which $x_{\text{similarity}}$, $x_{\text{conservation}}$ and $x_{\text{literature}}$ are scaled between 0 and 1 by

$$F(x) = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{5.2}$$

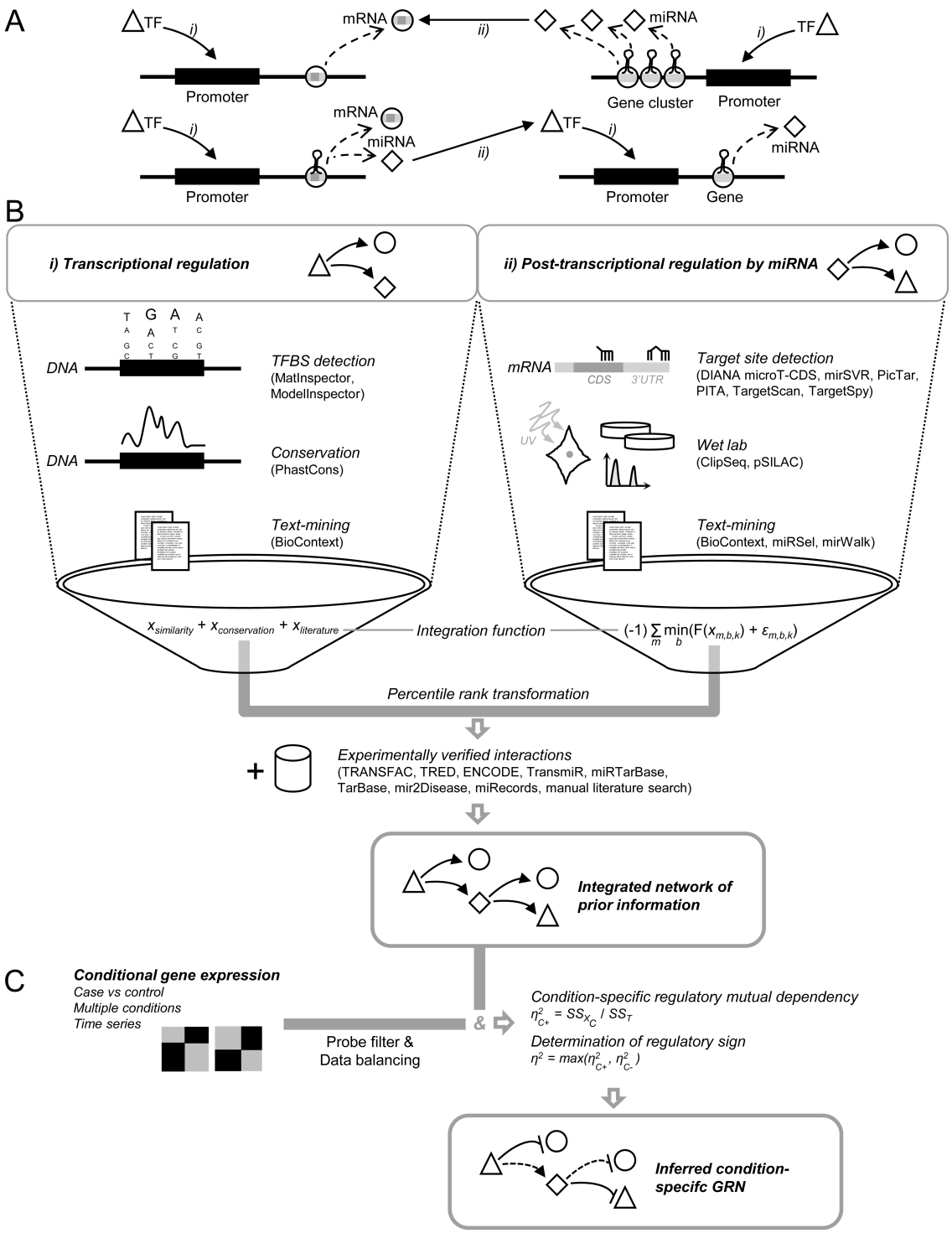


Figure 5.1 | **Overview of the COGERE workflow.** *(continued on next page)*

Figure 5.1 *(previous page)* | **A** | Outline of the biological paradigm of TF and miRNA interplay in gene expression regulation considered by COGERE: (**i**) transcriptional regulation is conducted by TFs binding to sites in promoter regions on the DNA of genes encoding either proteins or non-coding RNAs such as miRNAs. Here, miRNAs can be co-regulated with its protein-coding host gene, within a transcriptional unit (gene cluster), and/or through its own promoter; (**ii**) miRNA regulation takes place post-transcription by binding to sites mainly located on the 3'-UTR and/or the CDS of the target mRNA. The transcriptional and post-transcriptional regulatory pathways are interconnected. **B** | Construction of the prior network by information integration. For transcriptional regulation predicted transcription factor-binding sites (TFBS), their conservation and mined interactions from biomedical text were combined by a linear integration function. For post-transcriptional regulation, individual scores of six miRNA target prediction algorithms and text-mining results were integrated to a unified score weighting the regulatory potential of a miRNA:TG interaction. At this, AGO-bound CLIP-Seq data and proteomics (pSILAC) data was employed. All scores computed by the relevant integration function were normalized to percentile ranks (= prior score). Experimentally verified interactions were added to the prior network (prior score = 1). **C** | Determination of condition-specific regulation. For user-specified normalized and $\log_2$-transformed mRNA and/or miRNA expression data of at least two conditions, COGERE computes for each interaction of the prior network, the strength of the conditional dependency and the condition-specific regulatory sign (stimulation/repression) by deriving the coefficient η^2 with its corresponding P by a two-way ANOVA.

Integration of miRNA-mediated post-transcriptional regulatory interactions

Due to the diverse feature and model selection of miRNA:target prediction approaches[78,108], a set of six current algorithms was selected to cover a wide range of different miRNA targeting characteristics: DIANA-microT-CDS[236], mirSVR[142], PicTar[102], PITA 3/15[49], TargetScan 6.1[50], and TargetSpy[104]. Additionally, predicted interactions from literature mining provided by miRSel[265], miRWalk[266], and BioContext[264] were integrated. For miRSel and miRWalk, each interaction was scored by the number of retrieved documents containing a co-occurrence between the miRNA and its target. To minimize the false positive rate, only the most confident predictions of each tool were obtained as recommended by the authors, respectively (Table C.5). CLIP-Seq data from starBase (version 1.0)[195] was utilized to identify predicted target sites located in an AGO CLIP-Seq peak cluster. Here, each cluster holds a biological complexity score b describing a measure of reproducibility between biological replicates or experiments. Six score vectors $x_{m,b}$ with $\{b = 0, b = 1, b = 2, b = 3, b = 4, b \geq 5\}$ were prepared for each prediction method

$m \in M$. A biological complexity of $b = 0$ denotes target sites not located in any annotated AGO2-binding region. For each miRNA:target interaction the best score was retained. Each prediction score vector $x_{m,b}$ was transformed into an efficiency score vector $y_{m,b}$ of protein downregulation based on miRNA transfection data from Selbach *et al.*[91] (Figure B.5). Finally, the regression function $F(x_{m,b,k})$ for the k-th prediction score and the average log fold-change $y_{m,b,k}$ of all miRNA:target pairs with $x_{m,b,l} \geq x_{m,b,k}$ and a random error $\varepsilon_{m,b,k}$ was computed:

$$y_{m,b,k} = F(x_{m,b,k}) + \varepsilon_{m,b,k} \tag{5.3}$$

Here, the locally weighted least squares method was applied to fit the polynomial function of the predictor[267]. For each miRNA:target pair the prior score was computed:

$$\text{prior} = (-1) \sum_{m \in M} \min_{b}(y_{m,b,k}) \tag{5.4}$$

Transformation of prior scores

The integration of independent sources inherently results in non-identical, heterogeneous prior score distributions. To obtain unified weights for each interaction type, the raw prior scores were converted to percentile rank scores as follows:
Let

$x_i \in \text{prior}_k$ a prior score of an interaction type $k \in \{\text{TF:TG, miRNA:TG}\}$,

n_i the number of equal scored interactions, i.e. $|x_j \in \text{prior}_k : x_j = x_i|$,

m_i the number of lower scored interactions, i.e. $|x_j \in \text{prior}_k : x_j < x_i|$,

r_i the position of x_i in a sorted list of prior_k by decreasing order.

Then, the mean rank for ties was computed:

$$\begin{aligned}
\text{mean}(r_i, r_i+1, ..., r_i+n_i-1) &= \frac{r_i + (r_i+1) + ... + (r_i+n_i-1)}{n_i} \\
&= \frac{n_i r_i + (1 + ... + n_i - 1)}{n_i} \\
&= \frac{n_i r_i + n_i \frac{(n_i-1)}{2}}{n_i} \\
&= r_i + \frac{n_i - 1}{2} \\
&= m_i + 1 + 0.5 n_i - 0.5 \\
&= 0.5 n_i + m_i + 0.5
\end{aligned}$$

To obtain the percentile score, the rank was divided by the size N_k of the prior score distribution:

$$F(x_i \in \text{prior}_k) = \frac{(0.5 n_i + m_i + 0.5)}{N_k} \tag{5.5}$$

This equation allows an intuitive interpretation of the prior scores, e.g. a transformed prior_k score of 0.90 denotes an interaction with a higher regulatory potential than 90% of all interactions of type k contained in the prior model; in return a prior cut-off of 0.90 retains the 10% most reliable integrated regulatory associations.

Integration of verified regulatory interactions

Experimentally verified TF:TG interactions were collected from ENCODE[268], TRED[269], TRANSFAC[270], TransMir[271], and from manual literature search. miRNA:TG interactions were obtained from miRecords[111], miRTarBase[272], miR2Disease[164], and Tarbase[273]. For each interaction contained in one of these sources the prior score was set to 1.0.

5.2.2 Determination of condition-specific regulation by inference

Preprocessing of expression data

Two preprocessing steps were applied to the expression data in order to improve the discriminatory power of the inference approach:

i) *Balancing the data*. To avoid a condition-dependent bias, the sets of microarrays measured under the same condition are pruned to equal size n. The $M_{i,j}$ value was computed for each probe j on microarray i by dividing the intensity of j by the median intensity of the same probe across all microarrays. According to Kauffmann *et al.*[274] and Irizarry *et al.*[275], $M_{i,j}$ can be decomposed to the probe effect z_j (i.e. probe binding affinity), the differential expression effect $\beta_{i,j}$ (i.e. log scale expression level) and an independent identically distributed error term $\varepsilon_{i,j}$. As z_j and $\beta_{i,j}$ are the same across all k samples within one condition, computing the sum of all L_1 distances enabled to filter the n microarrays with minimal technical variation:

$$d_i = \sum_k \sum_j |M_{i,j} - M_{k,j}| \tag{5.6}$$

All samples of each condition were ranked by their increasing order of d_i and the top n microarrays were selected for further processing.

ii) *Filtering of non-present and uninformative transcripts.* In order to assess the context-specific strength of associations, the transcripts of both regulator and target had to meet the following two requirements: the genes needed to be expressed in all samples of interest and to show significant variation across the different conditions. Regarding the latter, TGs whose expression does not alter between the different conditions are unlikely to be under context-specific regulation. All probe sets were filtered which had sufficient expression intensities ($> \log_2(20)$) on more than 5% of the microarrays, and exhibited an adequate variation across samples (probe set expression interquartile range $>$ median expression interquartile range)[276].

Inferring condition-specific regulation by ANOVA

To score regulatory associations of the prior model in terms of condition-specific relevance, the non-parametric, nonlinear correlation coefficient η^2 (eta squared)[277] was utilized.

This variable was derived from a two-way ANOVA and enabled the quantification of the mutual dependency between a regulatory pair based on gene expression profiles over different experimental conditions. Observed expression data was modeled with n replicates and k conditions as responses of two factors X_C (condition) and X_G (RG and TG), their potential interaction $X_C \times X_G$ and the proportion of variation which cannot be explained by the model ε (measurement noise). Variance can be expressed in terms of the sum of squared deviations from the mean (sum of squares, SS)[278]. Accordingly, a two-way ANOVA splits the total sum of squares (SS_T) into four parts:

$$SS_T = SS_{X_C} + SS_{X_G} + SS_{X_C \times X_G} + SS_\varepsilon \tag{5.7}$$

Here, SS_{X_C} reflects the effect of differential gene expression between the conditions, SS_{X_G} is the difference in means of the expression profiles of RG and TG, $SS_{X_C \times X_G}$ denotes the joint effect of both factors and SS_ε quantifies the variation due to inaccuracy of measurement. For each regulatory pair two matrices of size $n \times k$ containing the expression values of the RG and the TG were extracted, respectively. From equation 5.7 the mutual dependence in gene expression between the different conditions was computed for each Z-score standardized expression matrix [1]. It was defined as the fraction of total variation explained by the variation in the data between conditions:

$$\eta^2_{C+} = \frac{SS_{X_C}}{SS_T}, \text{ with } \eta^2_{C+} \in [0,1] \tag{5.8}$$

This value can be interpreted in the same way as common correlation coefficients. It was taken account that η^2 does not explicitly test for negative regulation: the sign of the RG data was reversed and η^2_{C+} was calculated, respectively. The final score was defined as $\eta^2 = \max(\eta^2_{C+}, \eta^2_{C-})$[252]. Interactions with $\eta^2_{C+} < \eta^2_{C-}$ were signed as repression, otherwise as stimulation. The detailed procedure is illustrated in Algorithm 5.1 and Algorithm 5.2.

Regulatory associations showing a strong conditional dependency between RG and TG, i.e. having a high η^2 score, were assumed to be of high relevance. To test this dependency for statistical significance an F-test was conducted. For each η^2 the corresponding F-value

1 $Z\text{-score}(x) = \frac{x_i - \mu}{\sigma}$ with μ is arithmetic mean and σ is the standard deviation of x.

was calculated by dividing the effect variance of factor X_C by the total variance:

$$F_{X_C} = \frac{MS_{X_C}}{MS_T}, \text{ with } MS_i = \frac{SS_i}{\mathrm{df}_i} \text{ and } i \in \{X_C, T\} \tag{5.9}$$

where the degrees of freedom were chosen $\mathrm{df}_{X_C} = k-1$ and $\mathrm{df}_T = 2 \times n \times k - 1$. P were obtained from the F-distribution and adjusted by the Benjamini-Hochberg procedure[279] to control the false discovery rate (FDR).

Algorithm 5.1: Compute summed squared deviation from mean (sum of squares)

Data: Number of biological replicates n, number of conditions k, expression matrix of regulator r and target t, number of gene roles q

Result: η^2

begin
- $SS_T \leftarrow 0$ ▷ Total sum of squares
- $SS_{X_C} \leftarrow 0$ ▷ Sum of squares for factor condition
- $\mu_T \leftarrow \frac{\mu_r + \mu_t}{2}$ ▷ Total arithmetic *mean*
- **for** $i \leftarrow 1$ **to** k **do**
  - $\mu_{X_C} \leftarrow 0$ ▷ Arithmetic *mean* for factor condition
  - **for** $j \leftarrow 1$ **to** n **do**
    - $\mu_{X_C} \leftarrow \mu_{X_C} + \frac{(r_{i,j} + t_{i,j})}{qn}$
    - $SS_T \leftarrow SS_T + (r_{i,j} - \mu_T)^2 + (t_{i,j} - \mu_T)^2$
  - **end for**
  - $SS_{X_C} \leftarrow SS_{X_C} + 2q(\mu_{X_C} - \mu_T)^2$
- **end for**
- $\eta^2 \leftarrow \frac{SS_{X_C}}{SS_T}$
- **return** η^2

end

Algorithm 5.2: Rate condition-specific strength of association

Data: Number of biological replicates n, number of conditions k, expression values of regulator R and target T (r resp. t matrices of size $n \times k$), number of gene roles $q = 2$ (i.e. R or T)

Result: Condition-specific dependency $\eta^2_{\max}$, Sign of interaction s

begin
- $s \leftarrow 0$ ▷ Sign of regulation (-1 = repression, 1 = stimulation)
- $\eta^2_{\max} \leftarrow -1$ ▷ Mutual dependency
- **for** $r \in R$ **do**
  - **for** $t \in T$ **do**
    - $r \leftarrow \frac{r-\mu_r}{\sigma_r}$ ▷ Z-norm by arithmetic mean μ and standard deviation σ
    - $t \leftarrow \frac{t-\mu_t}{\sigma_t}$ ▷ Z-norm by arithmetic μ and standard deviation σ
    - $\bar{r} \leftarrow -r$ ▷ Inversed R expression
    - $\eta^2_+ \leftarrow \text{ANOVA}(n, k, r, t, q)$ ▷ Analysis of variance, refer to Algorithm 5.1
    - $\eta^2_- \leftarrow \text{ANOVA}(n, k, \bar{r}, t, q)$ ▷ Analysis of variance, refer to Algorithm 5.1
    - **if** $\eta^2_+ > \eta^2_-$ **then**
      - **if** $\eta^2_{max} < \eta^2_+$ **then**
        - $\eta^2_{\max} \leftarrow \eta^2_+$ ▷ Positive association
        - $s \leftarrow 1$
      - **end if**
    - **else**
      - **if** $\eta^2 < \eta^2_-$ **then**
        - $\eta^2_{\max} \leftarrow \eta^2_-$ ▷ Negative association
        - $s \leftarrow -1$
      - **end if**
    - **end if**
  - **end for**
- **end for**
- **return** *(η^2_{max}, s)*

end

5.2.3 Evaluation

Comparison to existing tools

For the performance assessment COGERE was compared with the common methods mirConnX[237] and MAGIA2[239].

Performance assessment of the integration function

The set of mRNA expression data was taken from the miRNA transfection study performed by Linsley *et al.*[130]. The data was obtained from the NCBI Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) under accession GSE6838. Expression was measured at 24 h post-transfection featuring maximal mRNA silencing but minimal secondary effects by protein depletion. The expression profiles of HeLa, HCT116 Dicer^{ex5} and DLD-1 Dicer^{ex5} miRNA transfected cells relative to mock-transfected cells were computed. Probe IDs were mapped to NCBI Gene accession numbers. The probe with the lowest log ratio P for each gene was selected. To obtain statistically meaningful results, only experiments for which each prediction tool scored at least 150 interactions were retained. The final set was composed of 18 expression profiles containing 10 miRNAs (miR-106b, miR-16, miR-15a, miR-20a, miR-195, miR-103, let-7c, miR-107, miR-17-5p, miR-103).

COGERE prior scores were computed without the information of validated interactions to make the scores comparable among approaches. To mimic the integration functions of mirConnX[237] and MAGIA2[239], the 6 miRNA target prediction algorithms incorporated in the COGERE prior score were used. The scoring scheme of mirConnX was implemented by weighting a miRNA:TG association by the proportion of algorithms predicting the interaction. For MAGIA2 all 57 possible intersections between the 6 miRNA target prediction algorithms were computed. Spearman's rank correlation was conducted for the observed gene $\log_2$ fold-changes following miRNA transfection versus the scores computed for the miRNA:TG interaction. Further, a precision-recall analysis was performed. Here, the top and bottom 20% of candidate TGs were selected based on their expression changes.

Benchmark of prediction accuracy

An overview of the *in silico* gold standard preparation is shown in figure 5.2A. An *in silico* gold standard of 80 regulatory networks extracted from a human source network composed of 64 029 experimentally verified interactions was generated using GeneNetWeaver (version 3.1)[280]. Each sub-network contained 500 nodes and a varying number of edges (min = 852, median = 1226 and max = 1421) of which 50% were set to occur in a given set of conditions to obtain a balanced test set. Stochastic dynamical models of gene regulation accounting for molecular and experimental noise were applied to simulate matched expression data of mRNA and miRNA. The steady-state expression of all genes

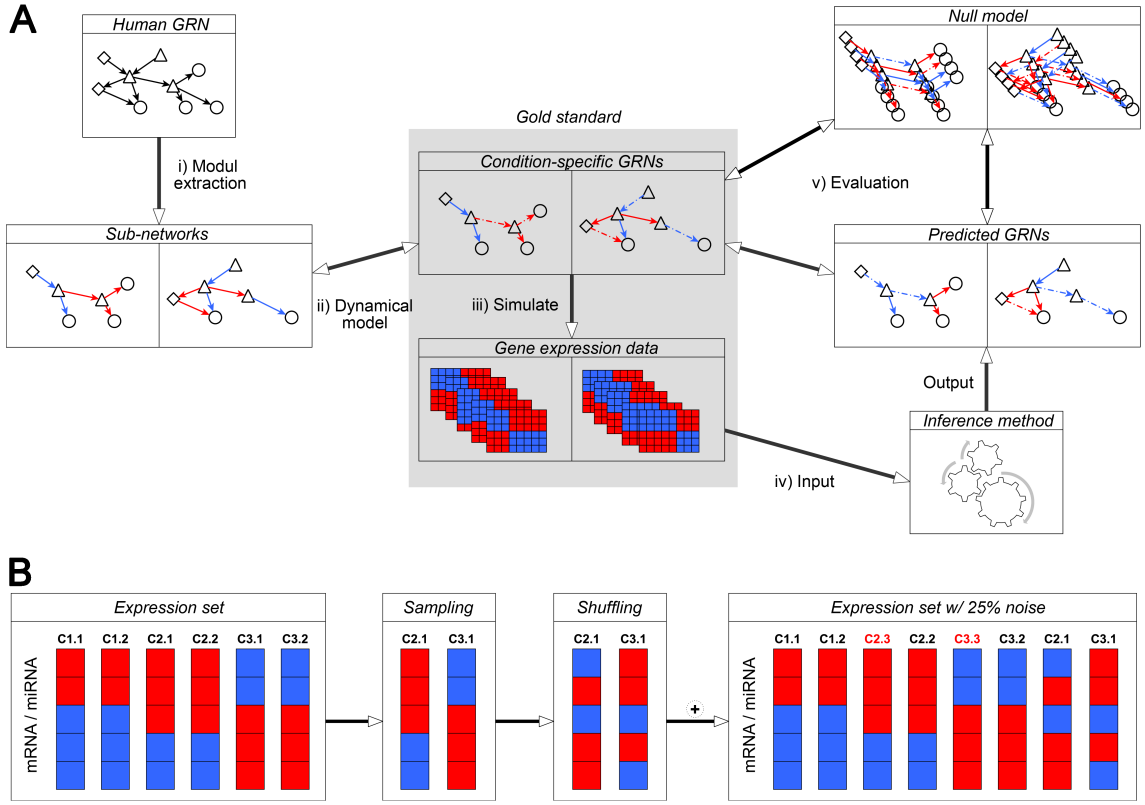


Figure 5.2 | **Outline of the performance assessment. A** | (**i**) Sub-networks were extracted from a known GRN from human. Each edge was randomly assigned a regulatory sign (red = stimulation, blue = repression) and a class: positive, if it was occurring in a given set of conditions, negative otherwise (dashed lines); (**ii**) the condition-specific GRNs were endowed with detailed dynamical models of gene regulation accounting for independent and synergistic interactions, as well as molecular and measurement noise; (**iii**) the set of dynamic sub-networks were simulated to produce steady states of gene and miRNA expression for a variety of conditions; (**iv**) several approaches were queried to infer the condition-specific GRNs from the matched *in silico* expression data; (**v**) the accuracy of the predicted networks was evaluated based on the area under the precision-recall curve metric against the true condition-specific GRNs (gold standard) and compared to random network predictions. **B** | To obtain noisy data sets, a defined number of microarrays was sampled from any condition (here: C1, C2, C3), shuffled and added to the expression data set.

was simulated for 60 conditions [c.f. NCI-60 cancer microarray project[241]], with five replicated measurements each. A precision-recall analysis was applied to determine the accuracy of the inferred condition-specific interactions. Also, a $\text{precision}_{\text{sign}}$-recall analysis

was conducted to estimate the accuracy of the predicted regulatory signs:

$$\text{recall} = \frac{|\Omega \cap G|}{|G|}, \quad \text{precision} = \frac{|\Omega \cap G|}{|\Omega|}, \quad \text{precision}_{\text{sign}} = \frac{|\Omega_s \cap G_s|}{|\Omega \cap G|} \tag{5.10}$$

where Ω is the set of predicted interactions for a given score threshold and G the set of condition-specific interactions in the gold standard. The $\text{precision}_{\text{sign}}$ measure was defined by considering also the sign of an interaction (Ω_s, G_s). By changing the threshold, precision-recall (PR) curves were generated [1]. Subsequently, the area under the curve (AUPR and $\text{AUP}_{\text{sign}}\text{R}$) was computed. To evaluate the predicted models against results from random guessing, a null model was constructed. For each expression set in the gold standard a condition-specific network was computed by sampling the scores and sings from a uniform distribution. This procedure was repeated 100 times and the median AUPR and $\text{AUP}_{\text{sign}}\text{R}$ values were recorded for each expression set. For technical details about the construction of the benchmark suite and the application of the prediction methods refer to Appendix A.

Case study data

The raw total gene signals of the NCI-60 Agilent microarray measurements were taken from the Liu *et al.* study[136]. Six cell samples (MCF7, HCT116, HT29, K562, SK-MEL-2 and CAK1-1) were labeled in quadruplicated and the remaining samples were labeled in duplicate. In accordance with the manufacturer, probe intensities < 5.0 were set to 5.0. Spots were removed if the gene was not detected on the microarray. The data was quantile normalized[135] and $\log_2$ transformed. All probes were assigned a miRBase ID or Entrez Gene ID, respectively. Finally, the set contained 789 miRNA probes measuring 533 genes and 26 091 mRNA probes of 16 651 genes. Processed data from the NCI-60 DTP human tumor cell line screen measuring the activity of 19 941 chemical compounds (drugs) in NCI-60 cell lines were obtained from CellMiner (CellMiner; http://discover.nci.nih.gov/cellminer/; version 1.4, July 2013). Expression values were averaged for replicates of cell lines. As proposed by Liu *et al.*[136], relationships between drug activity and gene expression were

1 For details on PR curves and their relation to the common receiver operating characteristic, please refer to Davis and Goadrich (2006)[281].

scored by Pearson's correlation coefficient.

5.3 Results

5.3.1 Comprehensive information integration

As the prior model of COGERE defines the hypothesis space for the inference of condition-specific regulation, the information integration step has to be extensive. Several sources containing regulatory interaction information were combined to a unique, directed graph constituting a static model of feasible gene regulation. Each interaction was weighted by a prior score computed by a domain-specific integration function. COGERE contained a regulatory network with 5 481 057 interactions for human and 3 472 682 interactions for mouse; thereof, 85 157 human and 18 389 murine interactions had the highest prior score of 1. The prior networks were composed of 22 523 (human) and 21 342 (mouse) genes with at least one interaction; thereof, 2 273 human and 2 007 murine genes were annotated as TFs and 1 028 human and 661 murine genes were annotated as miRNAs.

As a first attempt to globally characterize the topology of a network, the degree distribution is usually analyzed. The number of regulated genes per regulator, i.e. out-degrees, and the number of regulators per regulated gene, i.e. in-degrees, were following an exponential distribution $P_\lambda(k) = \lambda e^{-\lambda k}$ ($R^2_{out} \sim 0.94$, $R^2_{in} \sim 0.76$, Kolmogorov-Smirnov statistic ~ 0.13) rather than a Poisson probability function $P_\lambda(k) = \frac{\lambda^k e^{-\lambda}}{k!}$ ($R^2_{out} \sim 0.02$, $R^2_{in} \sim 0.11$, Kolmogorov-Smirnov statistic ~ 0.53). The latter defines a random network topology[282]. The out-degrees exhibited a broader distribution with a lower γ coefficient ($\gamma \sim 3.8 \times 10^{-4}$) resembling rather a power law, i.e. more regulators have many targets, whereas the in-degree distribution had a narrow exponential decay ($\gamma \sim 5.1 \times 10^{-3}$). This implies that a single TG is less likely to be regulated by a high number of RGs combinatorially. This so-called 'EIPO' topology (exponential in-degree and power-law out-degree) has been suggested to reflect the molecular limits on the number of regulators that can simultaneously exert an effect on the TG expression[283,284]. In any case, however, most nodes of the prior networks exhibited a low connectivity, while very few nodes had a very high degree. It was therefore felt that the common concept of a hub-containing structure with some highly-connected global regulators and many less connected fine-tuners[285] was applicable to the integrated networks. This observation can be more sharpened using

any prior score cut-off.

Comparing the amount of high-confident interactions (prior score > 0.9) to recent data pools, the presented model contained an extensive set of qualitative information: 294 394 TF:TG, 11 258 TF:miRNA and 316 875 miRNA:TG interactions in human. In comparison, ENCODE[268] featured about 27 386 TF:TG, TransmiR[271] 353 TF:miRNA and the recent release of miRTarBase[272] about 45 540 miRNA:TG human regulatory associations. Since there was less data available for mouse, the information gain was even higher: 199 308 TF:TG, 4 105 TF:miRNA and 156 779 miRNA:TG high-confident interactions. In comparison: TRANSFAC[270] had 1 118 TF:TG, TransmiR 16 TF:miRNA and miRTarBase 13 405 murine interactions. The TransmiR database provided regulatory associations for only 9% of human miRNA genes and 2% of murine genes. This shortcoming was substantially improved in the presented study by extensively collecting data from existing studies and carefully predicting promoter sequences. It was considered that miRNA genes can be embedded within a protein-coding host gene, and/or being part of an independent transcriptional unit, and/or can have their own promoter. COGERE predicted transcriptional regulation for 51% of all human and 50% of all murine miRNA genes (as annotated in miRBase 19). This was an increase compared to existing integrative approaches that model transcriptional regulation of about 29% (MAGIA2[239]) to 31% (mirConnX[237]) of human miRNA genes and between 46% and 47% of murine miRNA genes, respectively.

Regulatory interactions were mined from all available biomedical text and from databases, enabling the storage of relevant references. The current prior model contains 141 713 references for 97 816 interactions in human and 44 950 references for 25 142 interactions in mouse.

It was of interest to know whether relevant information can be extracted from such an elaborate collection. Fur this purpose, general co-regulation patterns between TFs and miRNAs were mined exemplary. Here, the null hypothesis that the number of shared targets is not greater than expected by chance was measured by means of the Hypergeometric distribution:

$$P(t_{ij}, t_i, t_j, T) = 1 - \sum_{k=0}^{t_{ij}-1} \frac{\binom{k}{t_j}\binom{t_i-k}{|T|-t_j}}{\binom{t_i}{|T|}} \tag{5.11}$$

where $t \in T$ denotes the number of targets regulated by either RG i or RG j or by both RGs

(t_{ij}). Requiring a minimum prior score of 0.9, the null hypothesis was rejected (Bonferroni corrected $P < 10^{-4}$) for 11 002 synergistic interactions of 211 transcriptional and 273 post-transcriptional regulators in human. Among these the tumor suppressor TP53 and the oncomir mir-21 were found with the strongest evidence to regulate a common set of genes (Bonferroni corrected $P = 1.56 \times 10^{-29}$; $P = 3.57 \times 10^{-09}$ in mouse). This observation suggests that mir-21 overexpression impairs the tumor-suppressive function of the TP53 pathway. Indeed, it has been reported that mir-21 inhibition resulted in mRNA level upregulation of several TP53 TGs which are required for TP53 activity in breast cancer cells[286]. Recently, Ma *et al.*[287] examined the functional interaction between TP53 and mir-21 *in vivo*. They reported that loss of mir-21 has a substantial effect on apoptosis of TP53-deficient cells and concluded that inhibition of mir-21 would be a promising strategy in cancer treatment inducing cell death against TP53-deficient tumors overcoming chemoresistance. Using the prior network, a detailed investigation on the shared targets is feasible. Please note that this analysis did not take context-specific co-expression and co-regulation of biological pathways into account.

5.3.2 Improved weighting of miRNA:TG interactions *a priori*

Since COGERE integrates six miRNA target prediction algorithms into a unique scoring framework under consideration of individual target scores, it was of interested to know whether the integration function improves previous approaches such as the ordinary intersection of several tools.

First, the COGERE prior scores were compared with the prior scores computed by the integration function used by mirConnX[237]. The latter weights each miRNA:TG interaction by the fraction of target prediction tools confirming a potential regulation. The outcome is a prior network with a discrete score distribution composed of $\{0, 1/6, 1/3, 1/2, 2/3, 5/6, 1\}$. To obtain the intrinsic value of how well the weights describe the regulatory potential of an interaction, the overall ranking performance of both scoring schemes was evaluated. For this purpose, Spearman's rank correlation was conducted between the observed $\log_2$ expression change following miRNA transfection and the prior weights of the miRNA:TG interactions, respectively. Figure 5.3A shows that both attempts for combining multiple target prediction tools exhibited a better performance compared to the average performance of all individual tools. At this, the COGERE prior score strongly outperformed the basic

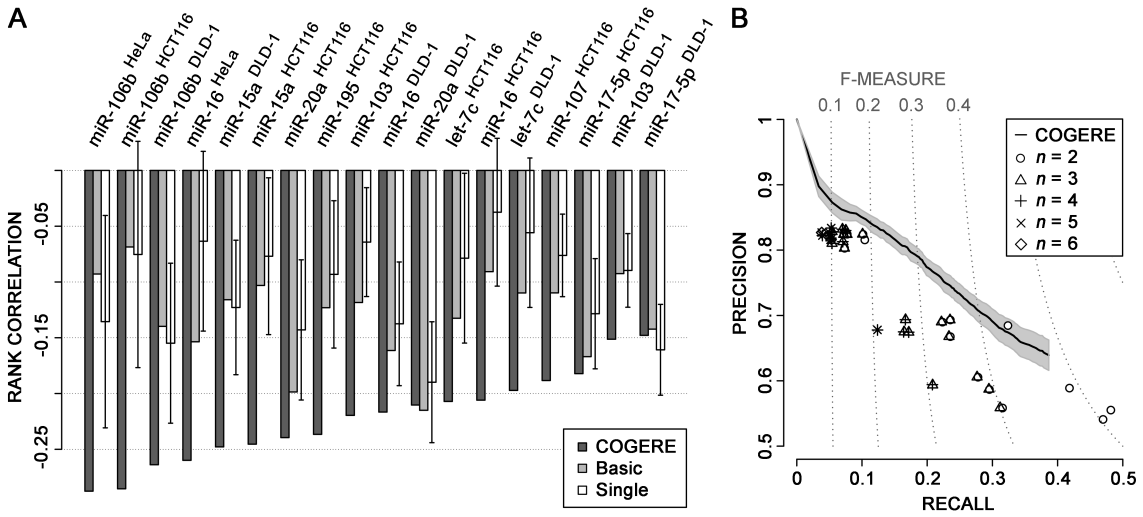


Figure 5.3 | **Evaluation of the prior score of miRNA:TG interactions. A** | Rank correlations (vertical bars) between predicted interaction weights and observed mRNA $\log_2$ expression changes measured post-transfection of 11 miRNAs in three cell lines[130]. The lower the correlation coefficient, the better represents the scoring framework the efficiency of a miRNA-mediated regulation. Weighting miRNA:TG associations using the COGERE prior score outperformed the basic scoring framework applied by mirConnX[237] in 94% of cases. Both scoring frameworks improved the average performance of all single target prediction algorithms. The error bars denote the 95% confidence interval for the mean. **B** | Mean precision-recall curve of the COGERE prior score ranking the top 20% most downregulated targets (positives) and 20% least downregulated targets (negatives) of each transfection data set. Shown are also the mean precision-recall values for all intersections of n miRNA target prediction algorithms. For a given recall the ranking by the prior score yielded an average advantage of 7.5% points in precision compared to the simple tool intersection applied by MAGIA2[239]. The F-measure denotes the harmonic mean between precision and recall. The shaded area indicates the 95% confidence interval for the mean.

weighting used by mirConnX. In contrast to the prior score, the simple combination of target prediction tools was not optimized to describe potential miRNA-induced expression changes. In 16 of the 18 experiments, the performance of the prior score outperformed the basic scoring framework which constituted a significant improvement (paired signed rank test $P = 1.7 \times 10^{-4}$). In all except one case, the COGERE integration function was superior to a blindfolded random selection of a single algorithm.

Second, to analyze the performance of the prior score and the intersection of tools as used in MAGIA2[239], all 57 possible intersections composed of at least two of the six algorithms were generated. The precision metric was defined as the fraction of predictions

that are true positives and recall as the proportion of actual positives that are correctly identified as such. Figure 5.3B shows that the prior score strongly improved the precision of the prior network over almost all values of recall. On average, the ranking by prior scores yielded a significant advantage of 7.5% points in precision (paired signed rank test $P = 8.8 \times 10^{-11}$) compared to any tool intersection. Interestingly, the intersection method was not straightforward and thus did not assure a gain of precision for a higher number of intersected tools on the expense of recall.

5.3.3 Advanced inference of condition-specific interactions

An *in silico* benchmark set (80 networks of size 500 nodes with corresponding steady-state expression data) was generated. The implemented framework was in accordance with the approach proposed by the Dialogue for Reverse Engineering Assessments and Methods (DREAM) competition[280]. This allowed to test COGERE against a known ground truth and to compare it to the common approaches mirConnX[237] and MAGIA2[239]. To measure prediction accuracy, the area under the precision-recall curve (AUPR) and the area under the $\text{precision}_{\text{sign}}$-recall curve ($\text{AUP}_{\text{sign}}\text{R}$) was calculated. At this, the recall metric described the fraction of predicted condition-specific interactions defined by the gold standard, precision denoted the proportion of true condition-specific predictions in the result set, and $\text{precision}_{\text{sign}}$ the fraction of correctly predicted regulatory signs. The performance advancement of each algorithm over the null model (random guessing) was computed and denoted as ΔAUPR and $\Delta\text{AUP}_{\text{sign}}\text{R}$, respectively.

First, it was evaluated how well the algorithms infer condition-specific edges from the expression data. Figure 5.4A shows that all tested algorithms performed better than random guessing predicting the whole condition-specific model ($\Delta\text{AUPR} > 0$ for RG:TG). Here, COGERE (median $\Delta\text{AUPR} = 0.294$) exhibited a significantly higher accuracy (Mann-Whitney U-test $P = 4 \times 10^{-15}$) than mirConnX (median $\Delta\text{AUPR} = 0.079$) and MAGIA2 (median $\Delta\text{AUPR} = 0.054$). COGERE achieved major overall improvements for the prediction of TF:TG as well as miRNA:TG interactions compared to existing tools. The major drawback of mirConnX and MAGIA2 was their low accuracy in predicting transcriptional regulation; both tools had their strength in detecting post-transcriptional regulation by miRNAs ($\Delta\text{AUPR}_{\text{TF:TG}} < \Delta\text{AUPR}_{\text{miRNA:TG}}$).

Second, the accuracy of predicted signs of the regulatory interactions was investigated

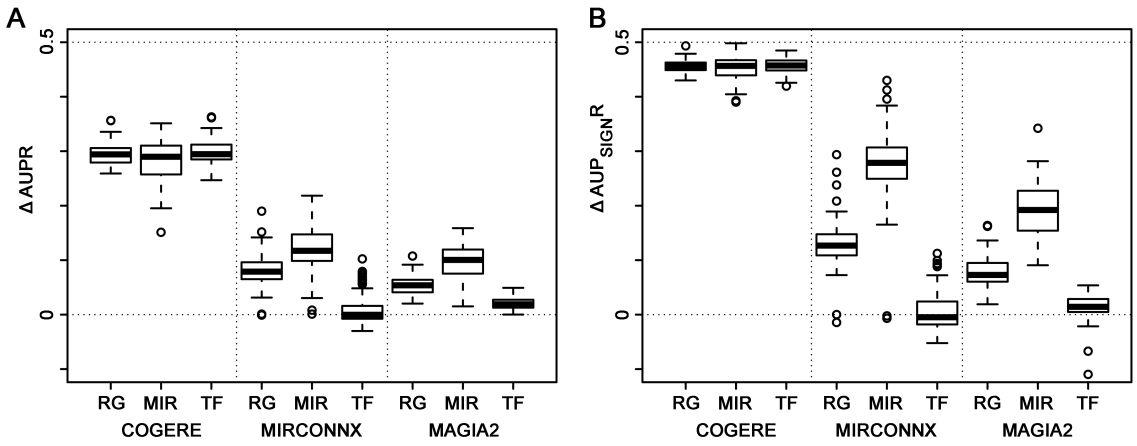


Figure 5.4 | **Accuracy of predicted condition-specific regulation. A** | AUPR values for each inference method for predicting condition-specific interactions. Shown is the deviation Δ from the null model (random guessing). COGERE outperformed mirConnX[237] and MAGIA2[239] on the prediction of condition-specific gene regulation tested against TF- and miRNA-mediated regulation (RG), only miRNA-mediated regulation (MIR), and only transcriptional regulation (TF). **B** | $\text{AUP}_{\text{sign}}\text{R}$ values for each inference method for predicting the condition-specific sign of an interaction. Shown is the deviation Δ from the null model (random guessing). $\text{Precision}_{\text{sign}}$-recall curves were computed to determine the fraction of correctly predicted condition-specific regulatory signs for each value of recall. COGERE exhibited an excellent accuracy tested against TF- and miRNA-mediated regulation (RG), only miRNA-mediated regulation (MIR), and only transcriptional regulation (TF). The accuracy of mirConnX and MAGIA2 in predicting transcriptional regulation was low.

(Figure 5.4B). Again, the $\text{AUP}_{\text{sign}}\text{R}$ values obtained by the tools were higher than the values obtained by the null model, whereat COGERE (median $\Delta\text{AUP}_{\text{sign}}\text{R} = 0.456$) substantially outperformed mirConnX (median $\Delta\text{AUP}_{\text{sign}}\text{R} = 0.127$) and MAGIA2 (median $\Delta\text{AUP}_{\text{sign}}\text{R} = 0.073$). Apparently, COGERE precisely determined the signs for both kinds of regulatory interaction for all values of recall. mirConnX and MAGIA2 exhibited similar lower accuracy profiles for TF:TG interactions compared to miRNA regulatory associations. The $\Delta\text{AUP}_{\text{sign}}\text{R}_{\text{miRNA:TG}}$ values obtained by mirConnX were significantly higher than the values of MAGIA2 (Mann-Whitney U-test $P = 2.2 \times 10^{-12}$).

5.4 Case study: Human cancer GRN

mRNA and miRNA profiles from tumor samples are frequently published. Having only been used to extract tumor-classifying molecular signatures[136] or confirming predicted miRNA:TG interactions[257], these expression data sets contain more information to be exploited. The condition-specific relevance of regulatory interactions was computed for the NCI-60 data panel which involved 60 cell lines originating from prostate cancer, lung cancer, breast cancer, melanoma, ovarian cancer, hematologic cancer, kidney cancer, colorectal cancer and malignant glioma. The top 10% predictions by COGERE (Table 5.1) were considered as highly relevant tumor specific interactions. This network is referred as the cancer GRN. In the following, it will be shown that the inferred GRN enables the systematic analysis of gene regulation in human cancers. This will demonstrate the potential of COGERE to reveal conditional regulatory landscapes.

Table 5.1 | **Network characteristics of the human cancer GRN.** Listed are the network statistics for the full inferred GRN and the subnetwork used in this study: the number of RGs and TGs, the number of their interactions, and the highest predicted P of a condition-specific interaction; prior = 1 denotes the fraction of interactions with a prior score of 1 and reference denotes the proportion of interactions with a reference.

NCI-60 GRN	RG = TF	RG = miRNA	TG	Interactions (prior =1, reference)	max. P
Full	473	251	8 853	634 863 (4%, 5%)	0.67
Study	387	180	5 869	63 486 (5%, 7%)	$< 10^{-5}$

5.4.1 The inferred GRN discovers causal RGs in cancer

To investigate whether the genes contained in the predicted GRN were substantially related to the condition of cancer, 2 760 known gene-cancer associations were extracted from HuGENavigator[288] for all cancer cell lines contained in the NCI-60 data. Altogether, 2 477 cancer-related genes were measured by the NCI-60 microarrays, of which 1 192 were contained in the inferred GRN. This denoted a significant enrichment of cancer-related genes (odds ratio $= 1.2$, Fisher test $P = 2.1 \times 10^{-7}$), consistent with the expectation that the inferred GRN should hold a higher fraction of cancer-related genes than expected by chance. Further, cancer-related TFs with at least one TG were significantly overrepresented

(odds ratio = 2.2, Fisher test $P = 4.1 \times 10^{-10}$). Five of the 10 most highly connected TFs (ELF3, EHF, ETS2, ETV5 and KLF6) have been reported to play a role in carcinogenesis. The enrichment was examined by using all 518 genes listed in the cancer Gene Census database[289]. Again, the GRN showed a significant high content of cancer-related genes (odds ratio = 1.4, Fisher test $P = 7.7 \times 10^{-6}$) and regulatory TFs (odds ratio = 2.1, Fisher test $P = 2.2 \times 10^{-6}$) even without filtering the database for NCI-60 tumors. This result suggests that the inferred GRN can be a valuable resource to extract information regarding cancer-specific gene regulation in general.

Next, it was of interest to know whether the human cancer GRN was able to recapitulate miRNAs that are both, namely dysregulated in malignant cells and at the same time causally linked to specific oncogenic processes. The miRNAs contained in the GRN were compared to entries in PhenomiR[165], a manually curated database of miRNAs that are dysregulated in diseases. All nine cancers of the NCI-60 panel were included. The Disease Ontology resource[290] was used to manually map the NCI-60 cell lines to PhenomiR disease terms (Table C.6). Remarkably, a highly significant enrichment of known dysregulated miRNAs was observed: 164 miRNAs in the inferred GRN were previously shown to be dysregulated in tumors of the NCI-60 data set (odds ratio = 6.0, Fisher test $P = 4.5 \times 10^{-12}$; Table 5.2). To investigate whether the dysregulated miRNAs contained in the human cancer GRN

Table 5.2 | **Enrichment of NCI-60 cancer types.** Listed are the numbers of miRNAs measured by the microarray (column: In database) and contained in the cancer GRN (column: In study) that are known to be dysregulated in a NCI-60 cancer. Also the corresponding odds ratio with its raw and FDR adjusted Fisher Test *P* is shown.

NCI-60 cancer	In database	In study	Odds ratio	*P*	FDR adj. *P*
DOID:10283, prostate cancer	223	127	3.31	1.7×10^{-11}	7.8×10^{-11}
DOID:1324, lung cancer	261	147	4.62	2.4×10^{-15}	2.1×10^{-14}
DOID:1612, breast cancer	296	146	3.42	2.6×10^{-10}	5.9×10^{-10}
DOID:1909, melanoma	286	134	2.50	5.2×10^{-07}	7.8×10^{-07}
DOID:2394, ovarian cancer	205	100	1.99	5.1×10^{-05}	5.7×10^{-05}
DOID:2531, hematologic cancer	211	121	3.11	1.1×10^{-10}	3.2×10^{-10}
DOID:263, kidney cancer	4	3	2.23	2.5×10^{-01}	2.5×10^{-01}
DOID:3070, malignant glioma	124	77	2.46	7.6×10^{-07}	9.8×10^{-07}
DOID:9256, colorectal cancer	172	103	2.80	3.6×10^{-09}	6.5×10^{-09}

were also known to hold a causal influence on cancer phenotypes, the causal relationships annotated in mirR2Disease were manually mapped to PhenomiR. It was striking that 48% of the miRNAs in the GRN that were known to be dysregulated were also annotated to causally affect cancer phenotypes (odds ratio = 1.7, Fisher test $P = 4.3 \times 10^{-3}$). Among the top 10 of the most highly connected miRNAs, all were known to be dysregulated and seven were assigned a known causal relationship (mir-27a, mir-23a, mir-17, mir-21, mir-29a, mir-20a and let-7b); among the top 25, all were dysregulated and 80% causal. In general, the higher the number of predicted condition-specific targets by COGERE, the higher was the probability that a miRNA exhibited a causal relationship to cancer (Figure 5.5A); e.g. of the 5% of miRNAs with the highest number of targets, 67% were causal, whereas for the 5% of miRNAs with the lowest number of regulatory interactions no causal relationship was known.

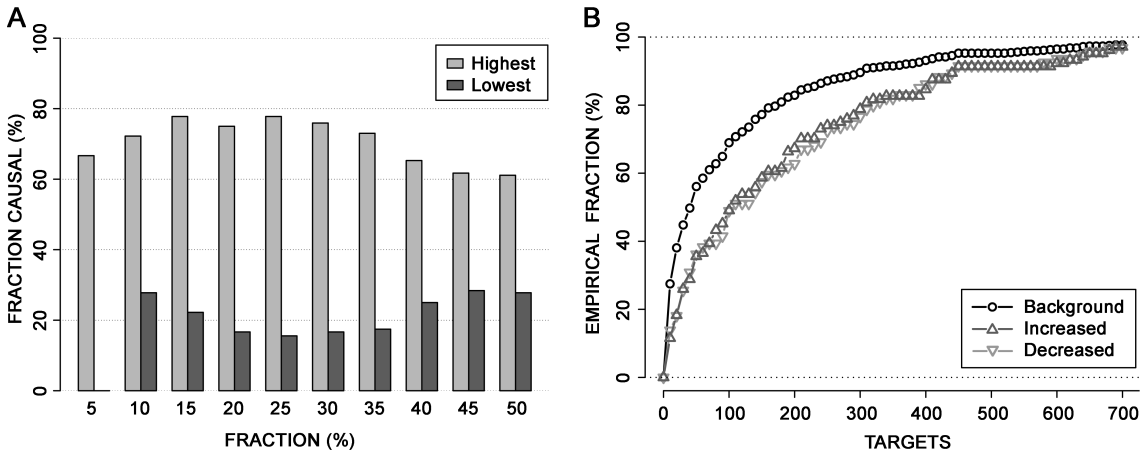


Figure 5.5 | **Degree distributions of the cancer GRN. A** | Fraction of miRNAs that have been reported to be causally linked to specific oncogenic processes (y-axis) for each fraction of miRNAs with the highest (light gray) or lowest (dark gray) number of targets in the cancer GRN (x-axis). miRNAs with a high number of predicted cancer-specific TGs have been more often reported to be causal than miRNAs with a low number of predicted cancer-specific TGs, e.g. 78% of the top 15% of miRNAs with a high out-degree had a causal role in cancer compared to only 22% of the bottom 15% of miRNAs with a low out-degree. **B** | Empirical out-degree distributions of all RGs. In average, RGs with a predicted association with an altered chemosensitivity of cancer cells exhibited 75 (increased drug response) or 81 (decreased drug response) more targets than any RG contained in the cancer GRN (background).

5.4.2 RGs associated to the hallmarks of cancer

The fact that a miRNA or a TF is contained in the inferred cancer GRN did not implicate that this RG plays a role in key oncogenic processes. Hanahan and Weinberg[291,292] proposed 10 traits of cancer that govern the transformation of normal cells to tumor cells: i) self sufficiency in growth signals, ii) insensitivity to antigrowth signals, iii) evading apotosis, iv) limitless replicative potential, v) sustained angiogenesis, vi) tissue invasion and metastasis, vii) genome instability and mutation viii) tumor promoting inflammation, ix) reprogramming energy metabolism, and x) evading immune detection. Plaisier *et al.*[293] prepared a set of Gene ontology[212] biological process terms representing the 10 hallmarks of cancer. This collection was employed to analyze TGs for functional enrichment. 1 393 genes were found involved in key oncologic processes in the cancer GRN which denoted a highly significant over-representation (odds ratio $= 1.3$, Fisher test $P = 6.6 \times 10^{-11}$).

Next, it was of interested to know which RGs in detail interact with these genes and are subsequently associated with the hallmarks of cancer. The functional enrichment analysis of the target sets of each RG recovered 31 miRNAs and 85 TFs that were predicted to regulate at least one process in oncogenesis (FDR adjusted Fisher test $P < 0.05$; Table 5.3).

Table 5.3 | **RGs associated to the hallmarks of cancer.** Listed are the hallmarks of cancer with their corresponding Gene Ontology IDs and their associated RGs (target gene enrichment odds ratio > 1 and FDR adjusted Fisher test $P < 0.05$).

Hallmark	Regulator gene(s)
Evading Apoptosis GO:0043069, GO:0043066, GO:0045768	CTBP2, ELF1, ESR2, let-7B, mir-18B, mir-21, mir-210, mir-23A, mir-23B, mir-24-1, mir-24-2, mir-7-1, JUN, KLF10, KLF2, KLF4, KLF6, NFKB2, SMAD3, TFDP2
Evading Immune Detection GO:0002837, GO:0002418, GO:0002367, GO:0050776	FLI1, mir-181B-1, mir-29A, NFKB2, PATZ1, SIX6, SMAD7
Genome Instability and Mutation GO:0051276, GO:0045005, GO:0006281	PPARD

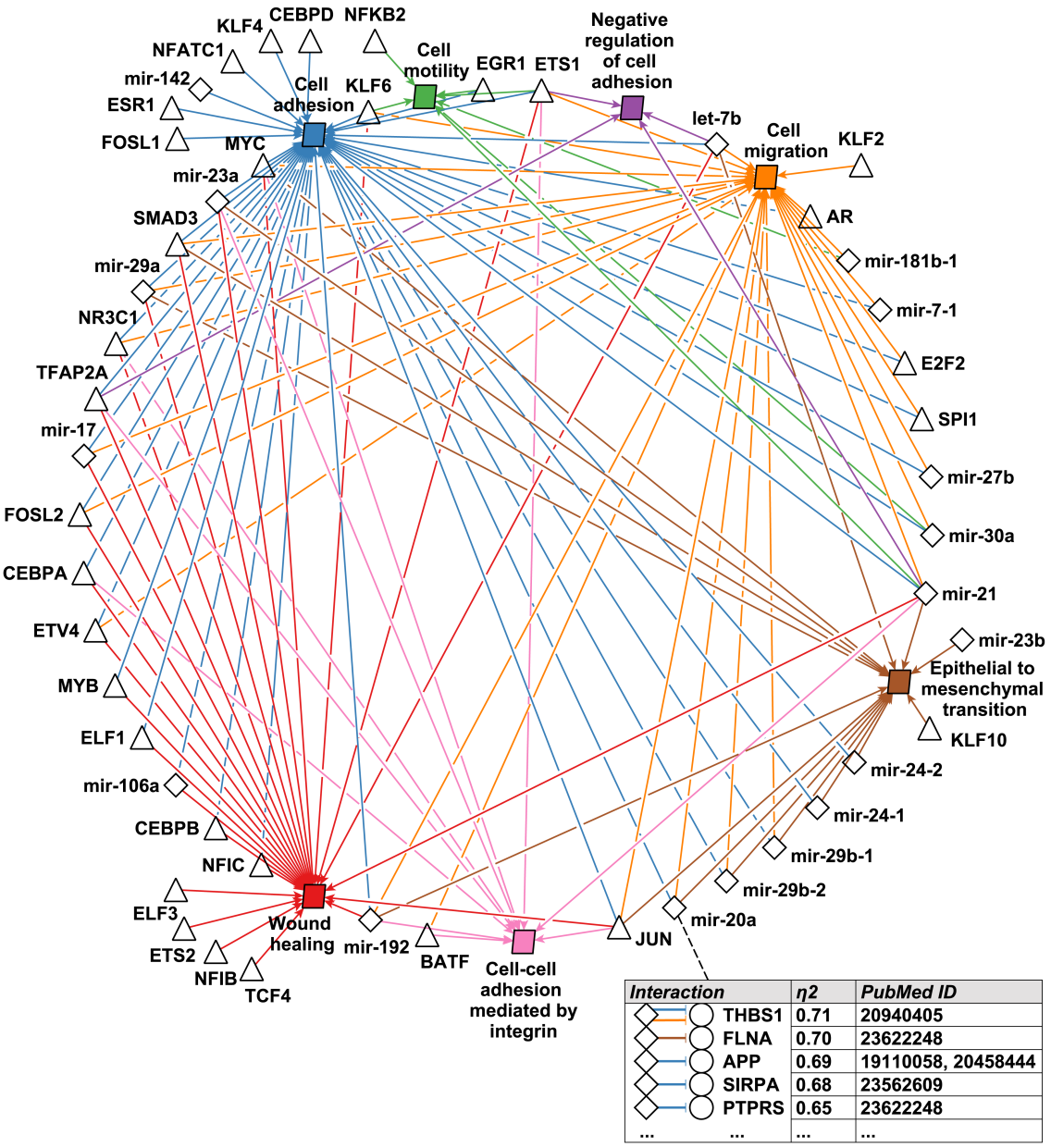
Table 5.3 *(continued)*

Hallmark	Regulator gene(s)
Insensitivity to Antigrowth Signals GO:0009968, GO:0030308, GO:0008285, GO:0045786, GO:0007165	AR, BATF3, CEBPA, CEBPB, CEBPD, CTBP2, E2F2, E2F7, EGR1, EHF, ELF1, ELF3, ELF4, ELK3, ERG, ESR1, ETS1, ETS2, ETV4, ETV6, ETV7, FLI1, FOSL1, FOSL2, GABPB2, GATA6, HIVEP2, HMGA2, HNF1A, let-7b, mir-106a, mir-130a, mir-142, mir-152, mir-17, mir-181a-1, mir-181b-1
Limitless Replicative Potential GO:0001302, GO:0032206, GO:0090398	STAT1
Reprogramming Energy Metabolism GO:0006096, GO:0071456	E2F2, mir-210, mir-23b, JUN
Self Sufficiency in Growth Signals GO:0009967, GO:0030307, GO:0008284, GO:0045787, GO:0007165	AR, BATF3, CEBPD, E2F2, E2F7, EGR1, EHF, ELF1, ELF3, ELF4, ELK3, ERG, ETS1, ETS2, ETV4, ETV5, ETV6, ETV7, FLI1, FOSL1, FOSL2, GATA6, HIVEP2, HMGA2, HNF1A, mir-130a, mir-142, mir-152, mir-17, mir-181a-1, mir-181c, mir-18b, mir-192, mir-19b-1, mir-19b-2, mir-21, mir-22, mir-23a, mir-23b, mir-24-1, mir-24-2, mir-27a, mir-27b, mir-29a, mir-29b-1, mir-29b-2, mir-365a, mir-365b, mir-7-1, JDP2, JUN, KLF10, KLF11, KLF13, KLF15, KLF2, KLF3, KLF4, KLF5, KLF6, KLF7, KLF8, KLF9, MLXIPL, MYB, MYC, MYCN, NCOA3, NFATC1, NFKB2, NPAS1, RARG, RELB, RUNX2, SMAD3, SNAI2, SOX5, SOX9, SP5, SPDEF, SPIB, STAT5A, STAT5B, STAT6, TCF3, TFAP2A, TFDP2, TOX, XBP1
Sustained Angiogenesis GO:0045765, GO:0045766, GO:0030949, GO:0001570	BATF, CEBPB, E2F2, EGR1, ELF1, ELF3, ELK3, ERG, ETS1, ETS2, ETV1, FLI1, FOSL1, FOSL2, let-7b, mir-142, mir-17, mir-181a-1, mir-181b-1, mir-18b, mir-192, mir-21, mir-210, mir-22, mir-23a, mir-23b, mir-24-1, mir-24-2, mir-27a, mir-27b, mir-29a, mir-30a, mir-365a, mir-365b, mir-7-1, JDP2, JUN, KLF10, KLF11, KLF12, KLF15, KLF2, KLF3, KLF4, KLF5, KLF6, KLF7, KLF9, MYB, MYC, NFIB, NPAS1, NR3C1, PPARD, RUNX2, SMAD3, STAT5A, TFDP2, TWIST2

Table 5.3 *(continued)*

Hallmark	Regulator gene(s)
Tissue Invasion and Metastasis GO:0042060, GO:0007162, GO:0033631, GO:0044331, GO:0001837, GO:0016477, GO:0048870, GO:0007155	AR, BATF, BATF3, CEBPA, CEBPB, CEBPD, E2F2, E2F7, EGR1, EHF, ELF1, ELF3, ELF4, ELK3, ERG, ESR1, ESR2, ETS1, ETS2, ETV1, ETV4, ETV5, ETV6, FLI1, FOSL1, FOSL2, GABPB2, GATA6, HIVEP2, let-7b, mir-106a, mir-130a, mir-142, mir-152, mir-17, mir-181b-1, mir-18b, mir-192, mir-20a, mir-21, mir-22, mir-23a, mir-23b, mir-24-1, mir-24-2, mir-27a, mir-27b, mir-29a, mir-29B-1, mir-29b-2, mir-30a, mir-365a, mir-365b, mir-7-1, mir-7-2, JDP2, JUN, KLF10, KLF11, KLF12, KLF15, KLF2, KLF3, KLF4, KLF5, KLF6, KLF7, KLF8, KLF9, MLXIPL, MYB, MYC, MYCN, NFATC1, NFIB, NFIC, NFKB2, NR3C1, RELB, SMAD3, SNAI2, SOX5, SOX9, SP5, SPDEF, SPI1, SPIB, STAT5A, STAT5B, STAT6, TCF3, TCF4, TFAP2A, TFDP2, TOX, TWIST2, XBP1
Tumor Promoting Inflammation GO:0002419, GO:0002420, GO:0002857, GO:0002842, GO:0002367, GO:0050776	FLI1, mir-181b-1, mir-29a, NFKB2, PATZ1, SIX6, SMAD7

Notably, 10 TFs and nine miRNAs were associated with at least five hallmarks of cancer (E2F2, ELF1, FLI1, JUN, KLF2, KLF4, KLF6, KLF10, NFKB2, TFDP2, mir-7-1, mir-18b, mir-21, mir-23a, mir-23b, mir-24-1, mir-24-2, mir-29a and mir-181b-1) suggesting that these genes are promising candidates for follow-up studies. Further, the evasion of inhibition mechanisms blocking proliferation and the metastatic potential of a cell were observed to be under strong control. Together 100 RGs (71 TFs and 29 miRNAs) were predicted to regulate 'insensitivity to antigrowth signals' followed by 97 RGs (71 TFs and 26 miRNAs) associated with 'tissue invasion and metastasis'. The latter hallmark is one of the defining features of malignant tumors making putative regulators excellent biomarker candidates. COGERE proposed a mechanistic explanation of how TFs and miRNAs act together to directly regulate genes involved in metastatic processes (Figure 5.6).



Interaction	η2	PubMed ID
THBS1	0.71	20940405
FLNA	0.70	23622248
APP	0.69	19110058, 20458444
SIRPA	0.68	23562609
PTPRS	0.65	23622248
...	...	...

Figure 5.6 | **Metastatic interplay of TFs and miRNAs.** Nodes are biological processes (colored parallelogram), TFs (triangle) and miRNAs (diamond). Arcs denote an enrichment of RG targets in a metastatic process and are colored, respectively. The top five predicted negative regulations of mir-20a are listed exemplary in the table shown at the lower right corner; e.g. the THBS1 repression by mir-20a which was described by Dews *et al.* [294] and holds a condition-specific regulation score of 0.71. This interaction affects cell adhesion and cell migration (blue and orange arcs). Note that the shown network was filtered by regulatory interactions having at least one literature reference (PubMed ID).

In this example, the cancer model suggested that mir-20a, a member of the miR-17∼92 miRNA cluster, regulates cell adhesion and cell migration in tumor metastasis through direct suppression of thrombospondin 1 (THBS1). An upregulation of miR-17∼92 has been described to promote angiogenesis and tumor growth[294], whereas increased THBS1 expression suppresses growth or metastasis of some tumors *in vivo* and inhibits angiogenesis[295]. The THBS1 downregulation has been observed primarily at the level of mRNA turnover[296] which is probably induced by miRNA-mediated mRNA degradation. These findings return a predicted cancer-specific interaction as an interesting subject for further investigations.

5.4.3 The cancer GRN predicts potential targets for cancer pharmacology

Given a condition-specific GRN, a key next step for the extraction of novel testable hypotheses is the integration of orthogonal information. Drug insensitivity or drug resistance are major obstacles in the successful treatment of cancer. Several studies have suggested that robustly positive or negative correlations between drug activity and gene expression reflect a role in chemosensitivity of cancer cells. A negative correlation may indicate that cancer cells with an increased expression level of mRNA or miRNA are less sensitive to the drug compound than other cells. On the contrary, if the correlation is positive, co-treatment with mRNA or miRNA may be used to enhance drug potency or reduce toxicity[136,297]. The correlation of miRNA and mRNA expression profiles versus drug activities over all NCI-60 cancer cell lines was calculated. First, the informative value of the correlation coefficients was validated by comparing the results to GI_{50} values measuring the growth inhibitory power of the test agent provided by Blower *et al.*[297]. They experimentally tested the activity pattern of 10 drugs following either inhibitor or precursor transfection of three miRNAs (let-7, mir-16 and mir-21) in A549 cell lines. The correlation coefficients were in good agreement with the average $\log_{10}$ fold-changes of GI_{50} values between lowered and raised miRNA levels ($R^2 = 0.38$, $P = 2.7 \times 10^{-4}$; Figure 5.7).

To gain a first broad perspective on the potential roles of the predicted RGs in cancer therapy, the associations of 163 anti-cancer compounds and all genes contained in the cancer GRN were analyzed. The set of drugs was restricted to compounds that were in clinical trial or were approved by the FDA (U.S. Food and Drug Administration). 45 miRNAs and 125 TFs accounting for 105 drug-miRNA and 309 drug-TF correlations

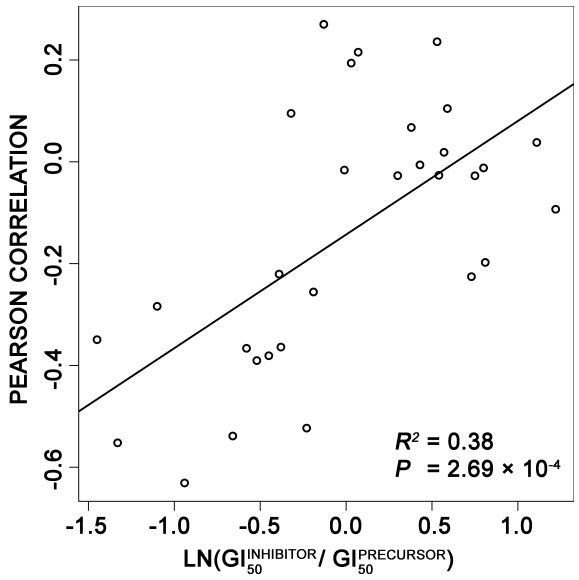


Figure 5.7 | **Comparison of correlation coefficients to GI_{50} values.** Shown are the $\log_{10}$ fold-changes of GI_{50} values between lowered and raised miRNA levels measured by Blower *et al.*[297] (x-axis) and the corresponding correlation coefficients of three miRNA expression profiles and 10 drug activites (y-axis). A good agreement between the experimental measurement and the correlation coefficients was observed ($R^2 = 0.38$, $P = 2.7 \times 10^{-4}$).

were observed reaching the α-level of $P < 10^{-4}$ proposed by Blower *et al.*[297]. This denoted a significant amount of potential drug targets (miRNA odds ratio = 2.0, Fisher test $P = 9.6 \times 10^{-3}$; TF odds ratio = 2.9, Fisher test $P = 1.9 \times 10^{-16}$). Among these, 23 miRNAs and 71 TFs were predicted to decrease the cancer cells' chemosensitivity. This set of chemoresistance factors exhibited in average 1.7 times more targets (factor 2.9 for miRNAs, and factor 1.5 for TFs) than any RG contained in the whole GRN (Mann-Whitney U-test $P = 6.7 \times 10^{-6}$; Figure 5.5B). For example, mir-22 was predicted with the highest amount of negative effects to compound potencies; it had the third most regulatory interactions in the cancer GRN. The aberrant expression of this oncogene has been reported to correlate with poor survival[298] and the results indicate that tumor cells expressing mir-22 are less sensitive to drug treatment. Based on its high number of targets, mir-22 may be an interesting subject for further assessments of its role in resistance to anticancer agents. It remains to be evaluated if mir-22 is suitable as a prognostic biomarker. However, if mir-22 plays a causal role in drug resistance, its inhibition may enhance the response of

malignant cells to cancer drug treatment.

Further, 25 miRNAs and 79 TFs that exhibited a positive correlation coefficient were observed. These were assumed to increase the susceptibility of NCI-60 cells to the action of at least one cancer drug. Interestingly, the proto-oncogene MYC was found as the RG which was predicted to positively affect the potency of the highest number of compounds. This TF is constitutively expressed in many cancers causing augmentation of cell proliferation[299]. To investigate whether this TF plays a substantial role in chemosensitivity, all positive correlated drug-gene associations composed of the 591 predicted MYC targets and the 8 MYC affected compounds were extracted (Figure 5.8). The expression of the MYC targets POLG2, CAMKV, VASH2, and OGFOD2 in cancer cells was predicted to increase the potency of oxaliplatin. Active derivatives of this compound form both inter- and intra-strand DNA cross-links resulting in inhibition of DNA replication and transcription and cell-cycle nonspecific cytotoxicity. POLG2 polymerase promotes DNA synthesis. Oxaliplatin has been described to induce lesions in the human MYC gene[300]. Cancer treatment with oxaliplatin may reduce the positive cancer-specific regulation of POLG2 by MYC. This, in turn, may cause an induced inhibitory effect on DNA synthesis, entailing an enhanced cytotoxic effect of this compound. In addition VASH2 is involved in positive regulation of angiogenesis, a typical process taking place in cancer cells. Loss of induced regulation of this gene may induce a secondary anti-cancer effect. Further, two compounds lowering estrogen levels were found: calusterone and dromostanolone propionate. It has been proposed that the human MYC gen-regulatory region embeds an estrogen-responsive *cis*-acting element[301] inducing rapid MYC expression in presence of estrogen. Further, estrogen repletion is accompanied by significant reduction in leukocyte adhesion[302]. The MYC target ICAM3 was predicted to increase the susceptibility of cancer cells to the action of both anti-estrogen compounds. This gene is a member of the intercellular adhesion molecule family and has been reported to induce cancer cell proliferation, cellular radio-resistance, cancer cell migration, and invasion[303]. Based on the COGERE predictions one can hypothesize, that the reduction in estrogen may reduce MYC expression resulting in reduced ICAM3 function entailing an increased drug potency. Another interesting compound for further investigations may be imexon, a 2-cyanoaziridine derivate with antitumor activity. This compound was predicted to be positively affected by the highest number of MYC targets. These 27 TGs contained amongst others BCL2, a well-known oncogene encoding an anti-apoptotic protein.

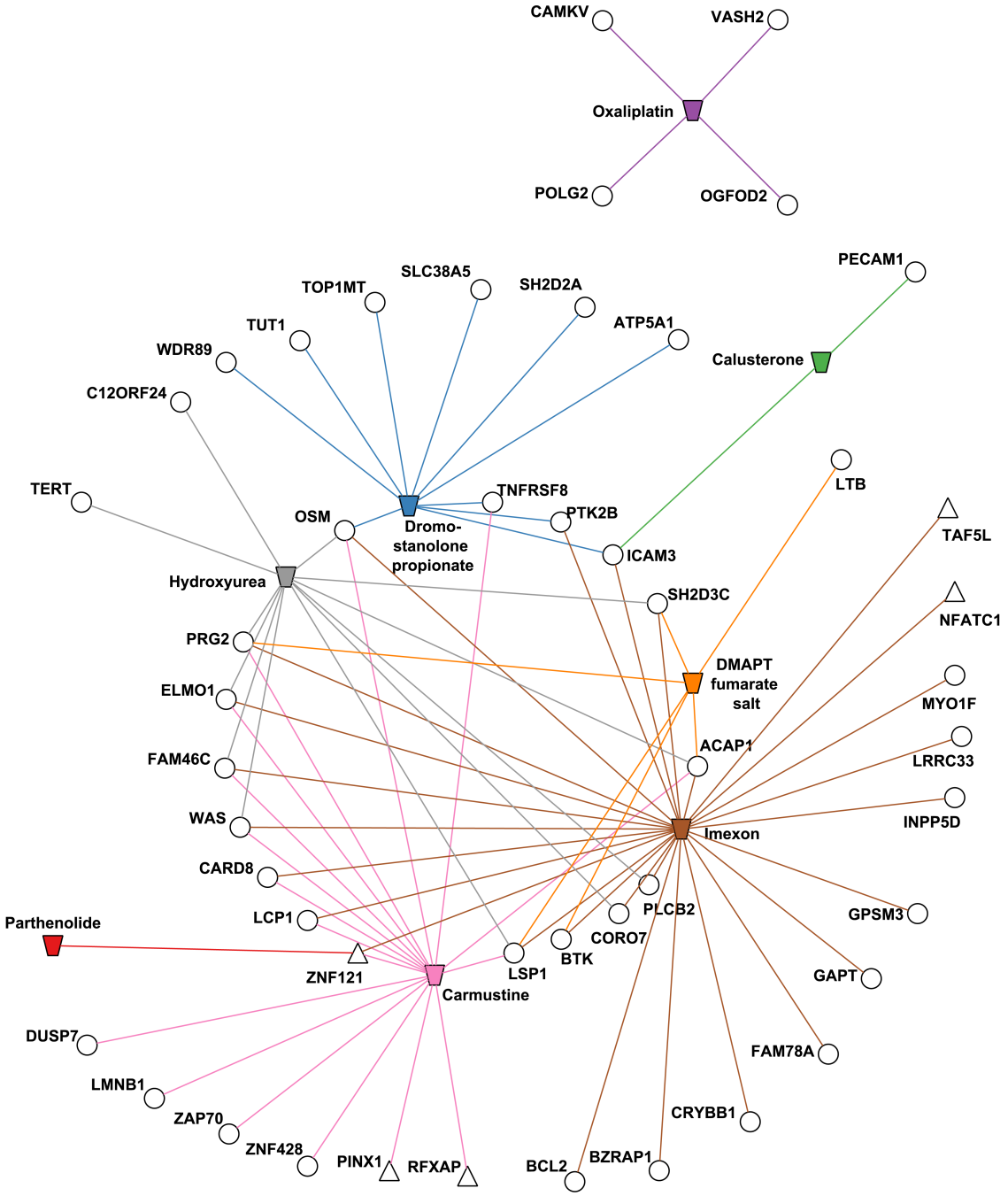


Figure 5.8 | **Drug-gene associations of MYC targets.** MYC was predicted to increase the susceptibility of NCI-60 cells to the action of eight drugs (trapezoids). Shown are the MYC targets contained in the inferred cancer GRN (genes are ellipses, TFs are triangles) which were predicted to positively affect the potency of at least one of the eight compounds. Gene-drug associations are illustrated as edges (colored by compound).

5.5 Conclusion

In contrast to the previous chapters, this study focused on the global modeling of miRNA-mediated gene regulation. The experimentalist is confronted with large data sets of high dimensionality reflecting the interplay of thousands of cellular components. Therefore, it is an imperative computational challenge to develop predictive and actionable models to investigate functionality as well as spatial and temporal behavior of these components. As the availability of experimental evidence in databases and the biomedical literature sharply increases, the systemic integration of existing knowledge to support the analysis of genome-wide molecular expression signatures of complex diseases becomes a bare requirement.

Firstly, a method was presented for the graph-oriented integration of several millions of annotated, literature-mined as well as pure sequence-based miRNA:TG and TF:TG interactions to a uniform scoring framework (prior score) of prior knowledge for human and mouse. It was illustrated that the integrated model comprehensively covers current knowledge provided by common experimental databases, the biomedical literature and computational predictions. The presented comparison to existing attempts revealed that the COGERE prior score constitutes a major improvement in the task of weighting miRNA regulation by their feasible regulatory effect on a TG. A basic combination of multiple prediction tools as conducted by mirConnX[237] performed better than a blindfolded random selection of any individual algorithm. Compared to a sighted systematic selection, this scoring scheme performed effectively worse than several individual tools (Figure 5.9). In contrast, the COGERE prior score improved the accuracy in 78% of all transfection experiments (median rank $= 1$) directly compared to any of the six integrated target prediction algorithms. Further, priors based on the COGERE scoring framework exhibited effectively more accurate information than a simple intersection of tools as used by MAGIA2[239]. The presented evaluation showed that a basic intersection of tools also implies a strong limitation in usability: it remains unclear to the user which tool combination fits best his requirements regarding recall and precision. Despite the current success of the COGERE prior score, ongoing progress in data collection by high-throughput '-omics' techniques will further improve the prior knowledge.

Secondly, to detect condition-specific regulation from mRNA and miRNA expression

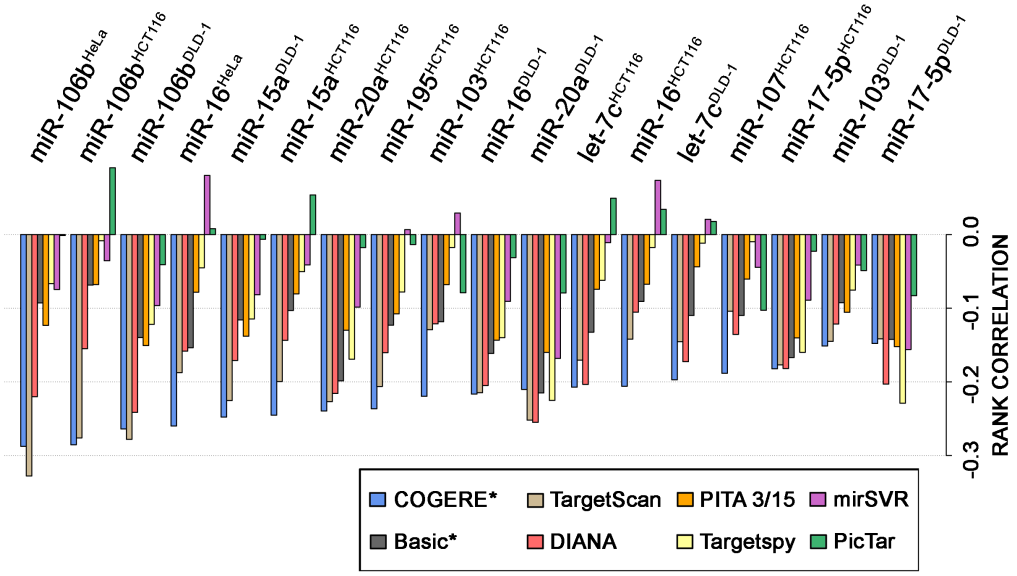


Figure 5.9 | **Comparison of the miRNA:TG prior score to single algorithms.** Shown are the Spearman's rank correlation coefficients between the $\log_2$ mRNA expression fold-change following miRNA transfection and the predicted scores for the regulatory interactions of each tool. Approaches integrating multiple target prediction algorithms are marked with an asterisk. It can be seen that the weighting of regulatory interactions by the prior score of COGERE is better than that of any individual prediction tool in 14 of 18 experiments (median rank = 1). The basic scoring system (median rank = 4) integrating all six sequence-based prediction algorithms is not optimized for the task of ranking the regulatory potential of miRNA:TG interactions and, thus, is not better than individual methods such as TargetScan (median rank = 2). It is of note that some of the individual target prediction tools were not trained on genome-wide expression data (e.g. PicTar) and thus perform worse compared to supervised approaches (e.g. TargetScan) in this assessment.

data, COGERE scores the relevance of prior interactions by measuring the mutual dependency between a RG and its TG. By applying an ANOVA the non-parametric and nonlinear correlation coefficient η^2 and its corresponding FDR adjusted P are derived. Here, neither a discretization of the expression data nor a setup with matching samples is required, increasing the robustness of COGERE. It was shown that COGERE strongly outperforms existing approaches in predicting condition-specific GRNs from synthetic expression data and held an excellent performance for predicting the regulatory sign of an interaction. Notably, the presented analysis denotes a comparative evaluation of MAGIA2 and mirConnX performance for the first time.

COGERE is capable to infer GRNs from unmatched data implying two advantages: i) expression data can be obtained from different studies/measurements with identical experimental setups, ii) detection of signals in at least a subset of experiments increases the robustness of the method against noise. COGERE balances the gene expression data by a condition-specific and individual-independent filtering of microarrays. The discriminatory power of the inference is sharpened as the variation within the conditions (technical variation) is reduced, whereas the differences between the conditions (biological variation) become more pronounced. An increased robustness of accuracy to detect context-specific effects due to differential TF- or miRNA-mediated regulation was observed in a benchmark with noisy expression data (Figure 5.2B and Figure 5.10).

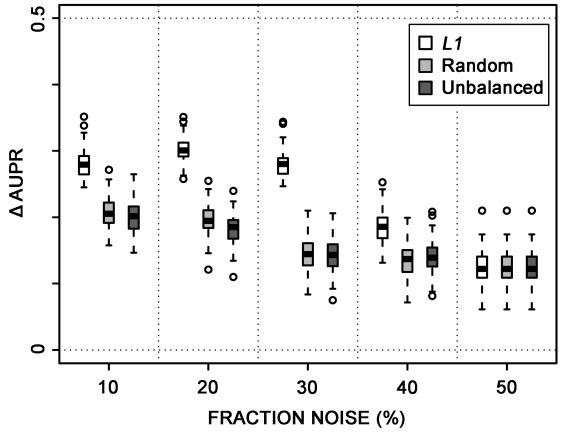


Figure 5.10 | **Robustness analysis of the inference method.** COGERE does not require matched data. Thus, balancing the expression sets, i.e. constituting an equal number of microarray samples for each condition, enables the filtering of appropriate measurements. COGERE computes the L_1 distance between all samples of the same condition to filter the optimal set of measurements of common size having the least sum of distances. The filtered expression data is used for condition-specific regulation inference. Pre-processing the expression data by this method maintains the inference accuracy compared to pre-processing by balancing the expression sets by random selection (Random) and inference without balancing the expression data (Unbalanced).

It should be noted that the performance assessment was based on simulated data. The *in silico* benchmark set was based on sub-networks from a human GRN with known interactions and thus was expected to exhibit similar types of structural properties and regulatory dynamics as realized in biological gene networks. Indeed, the evaluation

represented a simplified model of gene regulation. An *in silico* benchmark does certainly not replace the careful evaluation *in vivo*, but enables a systematically and efficiently performance validation and comparison of prediction methods over multiple networks. Unfortunately, to date an elaborate *in vivo* data set composed of mRNA and miRNA expression for several conditions as well as the corresponding experimentally verified condition-specific GRN is not available for human or mouse. It is likely that methods that do not perform well in a synthetic benchmark will perform even worse with real biological data[304]. In contrast to artificial data, linear correlation between a RG and a TG is a weak indicator of true condition-specific regulatory relationships in real expression measurements. This assumption is supported by a recent comprehensive and comparative evaluation of inference methods rating a two-way ANOVA-based approach best on the prediction of real GRNs from *Escherichia coli* and *Saccharomyces cerevisiae* expression data[252].

The NCI-60 cancer expression study was used to show that COGERE is a valuable resource to promote hypothesis-driven clinical research. It was demonstrated that the GRN inferred by COGERE captured disease-relevant regulation of cancer. A significant reliable proportion of known cancer-related genes and miRNAs were found in the predicted network. At this, causal miRNAs exhibited a higher number of condition-specific targets mirroring their central role in cancerogenesis. A relatively small subset of RGs were identified that play a role in multiple oncogenic processes in cancer. By using the inferred GRN, a mechanistic insight into the TF and miRNA interplay during the regulation of metastatic processes was provided. Since many somatic passenger mutations may also alter expression profiles, it is not expected that all condition-specific correlations are necessarily related to cancer driving processes.

The presented results suggest that the GRN contains novel, testable and interesting hypotheses regarding cancer-specific regulation beyond what is documented in existing databases. Moreover, the network predicted TFs and miRNAs that play a role in the chemosensitivity to approved cancer drugs and made novel predictions regarding the role of 116 RGs mediating the expression of genes associated with oncogenic processes. A predicted strong drug-gene relation may indicate a causal role in drug response[136,297]. If such a relationship proves to be causal, it could be exploited to improve cancer therapy. It was shown that condition-specific GRN information inferred by COGERE enables the analysis of potential drug targets in the context of gene regulation. Based on these

observations, it can be proposed that the predicted GRN contains several hypotheses promoting cancer pharmacogenomics.

In summary, this chapter introduced COGERE, a novel, generalizable approach that boosts signal-to-noise for the modeling of large-scale condition-specific regulatory landscapes in any cellular contexts. COGERE implements a robust inference method together with a concept of high-level data integration. It features the capacity of rational interpretation of expression signals in very large data sets in the context of known regulatory relations driving the discovery of new biology.

5.6 Availability

A web-based user interface was implemented using the Java framework VAADIN (version 6; https://vaadin.com) to enable an easy and fast access to the COGERE application (figure 5.11). COGERE is freely available under http://mips.helmholtz-muenchen.de/cogere .

5.6.1 The cancer GRN

To facilitate reader access and usability all data contained in the predicted NCI-60 cancer GRN was made available for further investigations: two files containing 634 863 ranked regulatory interactions and 1 721 242 scored gene associations to FDA approved compounds. It was aimed to provide cancer researchers a valuable resource to explore the cancer-specific GRN. All data can be downloaded from the COGERE website.

5.6.2 The prior network database

The species-specific prior networks were stored in a normalized MySQL database scheme (MySQL; http://www.mysql.com; Figure B.6). To be able to integrate diverse resources, an elaborate collection of IDs, symbols and synonyms were stored from Entrez Gene[253], Entrez Refseq[193], Ensemble[118], Unigene[305], and miRBase (version 13 to 18)[254].

5.6.3 The COGERE application

To provide experimentalists a tool to infer GRNs for their condition of interest, a standalone application of COGERE was developed. The implementation is based on Java (Java;

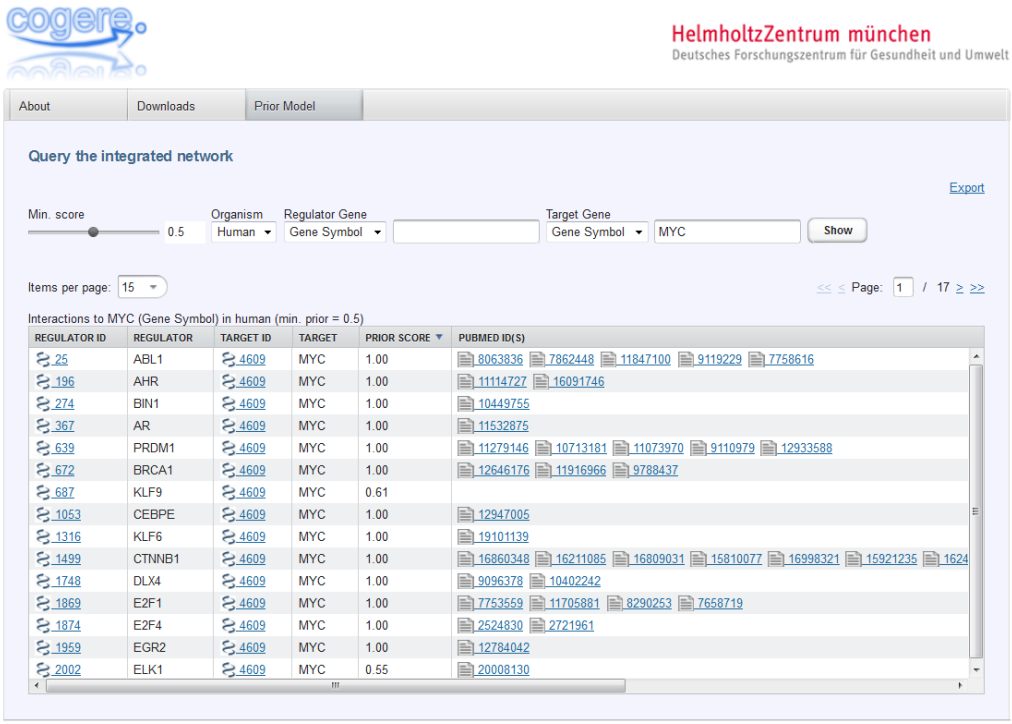


Figure 5.11 | **Web-based user interface of COGERE.** Via the web server the data discussed in this chapter can be accessed: the scored GRN from the NCI-60 expression data with all ranked gene-compound associations and the COGERE stand-alone application (interface section 'Download'). Further, the integrated prior networks are available: shown are exemplary the first 15 regulators of MYC with a minimum prior score of 0.5 in the human network. Listed are all Entrez gene IDs, official gene symbol names, the prior score, and the references. Gene and literature IDs are directly hyperlinked to the corresponding external databases (Entrez Gene and PubMed).

https://www.java.com) and R (R; www.r-project.org) and as such can be run on all major computer architectures. For convenience, a graphical user interface based on the Java SWT widget toolkit (Java SWT; http://www.eclipse.org/swt) is provided to ease the configuration and the usability of COGERE (Figure 5.12). However, modules to use the application as a pure command-line tool (e.g. for remote usage on a server) were provided. As the inference of large-scale GRNs is computationally intensive, COGERE provides the option to use multiple cores for parallel computing. Further, it is fully compatible to Revolution R (Revolution R; http://www.revolutionanalytics.com), a fast enterprise-class big data-big analytics R-based platform. For the sake of performance, the stand-alone application comes with a local copy of the prior information database. COGERE will notify the user

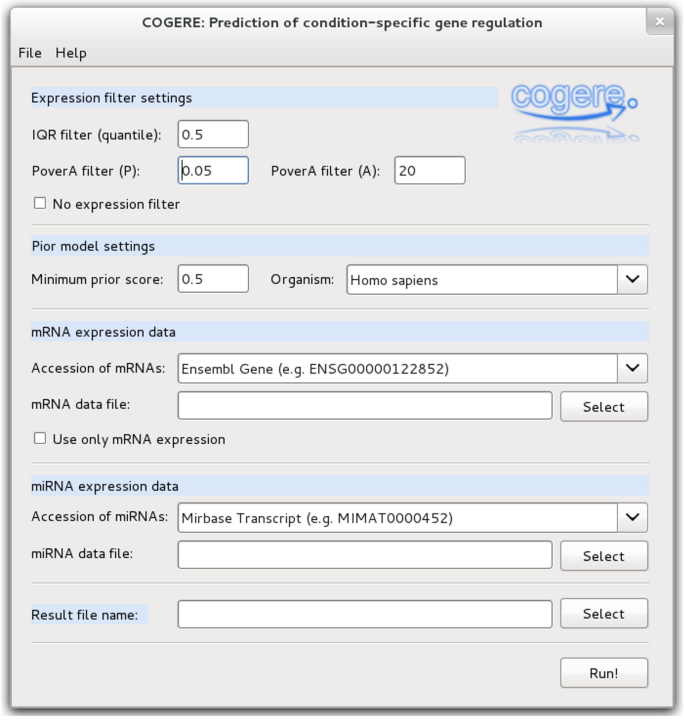


Figure 5.12 | **COGERE application.** Shown is the graphical user interface (runnable on Linux, Mac OS, and Windows; 32 bit and 64 bit versions). Modules checking for new versions, the configuration and the input parameters were implemented. Instructions for the usage, the required input and details on the computed output are provided in the 'Help' section.

if the local version of the prior network is out of date. Since gene regulation is inferred from expression data, COGERE requires processed (background corrected, normalized, and $\log_2$-transformed) mRNA and miRNA expression matrices from at least two different conditions with two samples each (tab-separated file). The first column denotes the probe ID, the first row contains the condition for each sample. Probes may be labeled by either Entrez Gene, Ensembl, Unigene, RefSeq, miRBase accession numbers or symbol names. The output consists of a '.csv' file with the condition-specific scored gene regulatory network (can be opened in any common spreadsheet application) and the execution log file which lists amongst others the filtered samples and probes. A README file is provided for detailed information. The binaries can be downloaded from the COGERE website.

CHAPTER 6

Conclusion and perspectives

The field of small non-coding regulatory RNA significantly evolved in the past two decades. At this, our understanding of post-transcriptional control mediated by miRNAs coupled with protein complexes (miRNPs) has greatly expanded. An important milestone in this field was the experimental capturing of miRNP:mRNA complexes in a cellular context. The advent of the AGO-bound CLIP-Seq protocol enabled, for the first time, the assignment of transcriptome-wide miRNP target sites. Until then, quantitative information on the miRNP:target pairing process was not available. The experimental detection of miRNA targets was mainly either guided by error-prone computational predictions or conducted by differential target expression analysis following miRNA inhibition or overexpression. Verified miRNA target sites were rare and biased, i.e. they were relying on the computational miRNA:mRNA duplex model. Thus, this data was not adequate for the comprehensive analysis of miRNA targeting. Although expression measurements were genome-wide, they were obscured by secondary effects and did not uncover the location of miRNA target sites. Consequently, various studies, including this doctoral thesis, were initiated briefly after the publication of the first two AGO-bound CLIP-Seq measurements in mammals.

Despite the differences in scope and methodology, early research focused on similar and highly relevant topics. Amongst others, they included the motif search in miRNP binding regions (e.g. Chi *et al.*, 2012[306]), the general prediction of RNA:protein interactions (e.g. Muppirala *et al.*, 2011[307]) the identification of determinants of miRNA action (e.g. Wen *et al.*, 2011[120]), and the implementation of novel miRNA target prediction algorithms (e.g. Betel *et al.*, 2010[142]; Liu *et al.*, 2013[308]; Rennie *et al.*, 2014[309]). Some studies

focused on the precise inference of miRNA:mRNA interaction maps integrating specific experimental characteristics (e.g. Corcoran *et al.*, 2011 [310]; Erhard *et al.*, 2013 [311]) and miRNA expression profiles (e.g. Chou *et al.*, 2013 [312]).

The first part of this thesis focused on the investigation of miRNP binding site characteristics. miRNAs act as guide molecule recruiting the miRNP to partial complementary sequences, preferentially on the 3'-UTR, of target mRNAs. Thus, the mining of general sequence patterns in miRNP binding regions that represent the sufficient minimal set of operative response element (MRE) types was of particular interest. The presented study focused on the pairing of the miRNA 5'-terminal seed sequence, since it was designated the highest relevance for target detection [313]. Notably, there were also several other studies elucidating alternative non-canonical modes of miRNA target recognition using AGO-bound CLIP-Seq data, such as the G-bulge site model at position $5-6$ of the miRNA 5'-end by Chi *et al.* in 2012.

I defined a set of six canonical seed types by applying a multi-branched recursion pattern mining strategy on murine and human AGO CLIP-Seq data. Here, the seed size ranged between $6-8$ nt, suggesting that the first eight nucleotides constitute the effective nucleation surface. Concurrent evidence was given by the later structural study of the miRNP by Elkayam *et al.* [38]. They proposed that the first 10 nt of the miRNA 5'-end are preorganized for pairing in an A-form helix conformation by the AGO2 protein. The nucleotides $9-11$ were proposed to face away from incoming target transcripts and, thus, are not available for nucleation [27]. Controversy still exists regarding the pairing of the very 5'-terminal nucleotide of the miRNA seed sequence. It has been hypothesized that its structural conformation makes it unavailable for pairing [27]. However, a large fraction of functional target sites was found complementary to this (α-)position. An explanation may give the AGO2 structure and the target sequence. Backbone atoms of a rigid loop in the middle domain of the AGO2 peptide chain exhibit a higher affinity for the base of uracil monophosphate [39]. This results in a bias of guide miRNA sequences starting with a uracil. The majority of conserved *bona fide* target sites exhibits an 3'-terminal 'adenine anchor' [90]. Thus, using α-seed types may be the favored scenario to encode MREs. However, it remains to be elucidated whether an adenine at the first position of the target site is either presumably recognized by Watson-Crick pairing or directly by a protein of the silencing complex.

Further, I observed that the repressive effect on the target transcript level is positively

related to the length of the consecutive seed-complementary segment. Thus, short seed types probably have only a minor role in direct gene regulation. However, these seed types made up the major fraction (up to 67%) of pairing conformations found in *bona fide* miRNP:mRNA complexes. This raises the question whether the primary role of these seed types is miRNA sequestration. Evidence for this hypothesis is given by the competitive endogenous RNAs (ceRNA) hypothesis proposed by Salmena *et al.* in 2011 [58]. Here, 3'-UTRs are suggested to be not only *cis* regulatory elements controlling the stability of the whole transcript, but also modulate gene expression in *trans*. The more transcripts with *bona fide* MREs are available, the higher the competition for miRNA binding, and ultimately, the lower the effective miRNA activity. This denotes an additional regulation layer by which target transcripts crosstalk by impairing the miRNA-mediated regulation of co-expressed genes. Further evidence comes from the study of Mukherji *et al.* [56] in 2011. They reported that effective miRNA regulation was strongly influenced by the available miRNA concentration, the target mRNA level, and the strength and number of embedded miRNA binding sites at the target sequence. In this context, it appears that long seed-complementary sites may act in *cis* whereas 6mer seed types qualify mRNAs to act as natural miRNA decoys due to their low impact on the stability of the host. Notably, the analysis of target site conservation in mammals showed that the majority of non-conserved sites ($\sim 75\%$) are covered by short seeds. This is in agreement with the general hypothesis that the non-coding transcriptome plays a major role in the greater complexity of higher eukaryotes [59]. Recently, miRNA-target interactions were associated to the evolution of organismal diversity [89].

To date, the ceRNA research is still in its infancy and a lot of questions remain to be answered. First evidences were found that ceRNA networks have implications in the initiation and progression of human diseases [59]. Functionalizing ceRNA interactions will undoubtedly lead to important insights about basic physiology. At this, functional studies assessing to what extend the defined 6mer sites modulate the miRNA function are of particular relevance. Further, the prediction of ceRNA crosstalks is depending on the computational identification of MREs on the relevant transcripts of interest. For this purpose, the majority of existing algorithms are not comprehensive, since these are limited on the detection of long seed matches and focus on conserved sites. This approach condones its lower sensitivity for a higher specificity. Indeed, the prediction of miRNA target sites relying on short seed types is challenging due to the very high false-positive

rate – an issue which has not been solved yet. Future studies are required that elucidate additional rules beyond miRNA:mRNA pairing. In particular, algorithms that are capable of identifying *trans*-acting MREs with high specificity are of interest.

A first step in this direction was taken in Chapter 3 of this thesis. Sequence-based, motif-based, structural and homology-based features of miRNP target sequences were extracted. The AGO-bound CLIP-Seq data supported the discriminative power of the characteristics that have been reported in literature. In addition, a novel feature for target detection was suggested: the asymmetric nucleotide composition between guanine (G) and cytosine (C). Chargaff's second parity rule states that the fraction of adenine (A%) $\sim$ the fraction of tyhmine (T%) and C% $\sim$ G% in polynucleotide chains[138]. 3'-UTR regions not bound by miRNP, in fact, almost perfectly followed this rule. But operative miRNP binding sites exhibited a skewed nucleotide distribution of C% $>$ G%. It has been argued that violations of Chargaff's second parity rule might be caused by RNA intrinsic structural constraints[121,139]. Calculations of the local 3'-UTR structure revealed that a higher C% $>$ G% skew is correlated with a less negative free energy required to unfold this region. Apparently, G nucleotides may be avoided in these regions. A reason may be the unique potential of guanine to pair with cytosine and by wobble base pairing with uracil. Thus, a higher fraction of guanine may induce local stem structures and subsequently lowers target site accessibility. This may be an important determinant of miRNP binding since it has been reported that the RISC is unable to unfold structured RNA[46].

This raises the question whether a good accessibility is already a sufficient condition for unspecific AGO binding or if there also exists characteristic local folds. Several lines of evidence support the second hypothesis. The feature analysis showed that accessible regions are favored by AGO, but this feature by itself is not of high specificity. In general, proteins with RNA-binding capability have a bias towards structurally accessible binding sites[185]. Thus, these segments can be also bound by another RNA-binding protein than the AGO containing miRNP. The 3'-UTR is covered with a variety of RNA-binding proteins[314]. For some of them structured *cis*-acting recognition elements were described. In example, the GAIT complex inhibits the translation of mRNAs that contain a specific stem-loop secondary structure (GAIT hairpins) in their 3'-UTR[185]. Specific RNA secondary structures also play crucial roles in various cellular processes, such as miRNA processing or translation. In example, it has been reported that transcription through GC

skew regions leads to the formation of long R-loop structures *in vivo*[315][1]. Interestingly, the whole miRNP binding region had a higher probability for negative selection than the remaining 3'-UTR. This may indicate an evolutionary pressure to maintain the structure of this region. Further, seed pattern analysis revealed that probably multiple miRNA target sites are embedded within the same segment. Thus, inital miRNP binding may be determined by additional coincidental factors beyond the sequence of the coupled miRNA. Also the results obtained in Chapter 4 spotlight the relevance of a coherent structure in the miRNP:target pairing process. Occuring genetic variation in the miRNA targetome is assumed to be a contributor to complex traits in the human population. An integrative framework was developed to determine trait-associated SNPs in the human miRNA targetome. The set of variants was composed of SNPs reported by GWA studies and proximal SNPs in strong LD. Besides a potential impact of these variants on seed-based MREs, a significant polymorphic structure of the AGO binding region was found.

Therefore, it is reasonable to postulate that specific local RNA structures may i) increase target site accessibility, ii) serve as AGO recognition elements, and/or iii) enhance miRNA:mRNA duplexing. For the latter two cases, the biophysical or biochemical elucidation of the tertiary structure of miRNP binding regions will shed a clear light on the matter. Due to the greater structural diversity of RNAs than proteins as well as the sensitivity of RNA structures to ions, solvent, metabolites and other biomolecules, the computational prediction of the *in vivo* RNA 3D structure has limitations[316]. However, in recent years great advances have been made in this field. Using novel tools will make a first pattern mining of the miRNP structural ensemble a feasible future project. In this connection, also the effect of specific genetic variants on secondary and tertiary RNA conformation is of particular interest. To assess whether specific structures enhance miRNA:mRNA duplexing, the particular guide miRNA sequence has to be known. Currently, this information is not available for AGO-bound CLIP-Seq experiments, but novel protocols addressing this issue, such as CLASH[74], are on their way.

It should be noted that the structural analyses have to consider the interplay between the miRNP and other RNA-binding proteins. It has been shown that RNA-binding proteins might act as switches, either potentiating or antagonizing miRNA-mediated silencing

1 R-loop structures are formed when transcribed G-rich/C-rich RNA strands anneal back to the template C-rich/G-rich DNA strands or *vice versa*.

by altering the local secondary structure of the target sequence[317]. Future work in this field holds promising perspectives to increase our understanding of the topology of post-transcriptional regulation networks.

Another potential future project is the elucidation of the evolutionary mechanism leading to the GC skew in miRNP binding sites. It has been reported that C-to-T deamination is a source of GC skews and has been linked to DNA methylation[318]. The deamination of 5-methylcytosine on the antisense strand induces a G-to-A transition on the complementary strand. Notably, adenine has been assigned a prominent role as anchor nucleotide for miRNA target sites[90]. However, it remains to be clarified whether this epigenetic programming or any alternative evolutionary mechanism contributed to the observed GC imbalance.

Chapter 3 also presents the application of the extracted miRNP characteristics and the canonical seed types to a real biological use case. By means of a SVM classifier the novel interaction between miR-92a and WISP1 in the progressive fibrotic lung disorder IPF was predicted. Notably, the implemented framework enables the generic selection of the miRNA:target pairing model and allows the extension by novel features in the future. Subsequent experiments provided strong evidence that the predicted interaction is a novel important miRNA-mediated regulation in pulmonary fibrosis. Future projects will address the clinical relevance of this interaction. Since miR-92a transcript levels and WISP1 expression are increased in IPF compared to unaffected controls, the treatment with miRNA mimics is of interest. Very recently a highly compelling work was published heading in this direction[319]. Members of the Kaminski lab, the van Rooij lab, and the biopharmaceutical company miRagen Therapeutics Inc. developed a miRNA-based treatment for IPF. They showed that the intravenous injection of synthetic RNA duplexes resulted in increased target miRNA levels *in vivo* of several days' duration. Further, endogenous miRNA function was restored leading to a blocking and reversing of bleomycin-induced pulmonary fibrosis whereas target gene expression was not affected under basal conditions.

In Chapter 4, polymorphic miRNA-mediated gene regulation was analyzed. As mentioned before, trait-associated variants beyond the seed-pairing region were found which likely affect miRNA efficacy. These were suggested to alter the local structure of the target region or result in alternative spliced transcripts. The former mechanism has been proposed to influence mRNA function in general[201,320], but was not directly related with miRNA-mediated regulation before. One independent study conducted at the same time

(Haas *et al.*[186]) and one very recent work (Day *et al.*[321]) reported similar results and some first experimental evidence for this hypothesis.

Further, a set of 53 trait-associated 3'-UTR SNPs were annotated to potentially impair miRNA activity and were associated to AEI performing an eQTL analysis. Some interesting candidates for further detailed investigation were described. Of note, a *cis*-acting genetic factor for IPF progression was not found.

Certainly, as other initial studies in this vein (e.g. Thomas *et al.*, 2011[322]; Richardson *et al.*, 2011[181]; Bruno *et al.*, 2012[323]), the presented work requires further examination. A major shortcoming is that the analysis was limited to *in silico* mutated reference sequences rather than the actual mature transcripts bearing the trait-associated SNPs. *In vivo* sequence information will shed light on the validity of the predicted mechanisms. In addition, the haplotype block requires a more detailed functional assessment. While the trait-associated SNPs were enriched in the 3'-UTR and related to AEI, other variants in LD may explain the association signal. Recent efforts to annotate genome-wide transcriptional regulatory elements, such as the Encyclopedia of DNA Elements (ENCODE) Project[324], allows to consider alternative compelling mechanisms that are not mediated by miRNAs. Both options, whole transcriptome sequencing (e.g. RNA Sequencing) and LD block dissection, will certainly increase the specificity of this kind of studies. Since Ago-bound CLIP-Seq data is limited to a single transcriptome, future measurements will extend the set of validated miRNP binding sites, and consequential, raise the sensitivity to detect *cis*-acting polymorphisms affecting miRNA regulation. Moreover, novel data from the 1 000 Genomes Project[229] exhibits an augmented map of human genetic variations which increases the number of variants in LD with each GWA study signal by greater than twofold compared with the HapMap resource[188] used in this thesis.

Finally, in Chapter 5, the global miRNA-mediated regulation was modeled using the novel approach COGERE. As the complexity of genetic interactions poses a strict limit to the potential of network inference from expression data single-handedly, the large-scale integration of complementary data, such as prior information from AGO-bound CLIP-Seq data, is of high value. For this purpose, a comprehensive collection of regulatory interactions was extracted from databases and literature containing *in silico*, *in vitro*, and *in vivo* interaction data. A novel data integration framework was presented weighting transcriptional and post-transcriptional interactions by their confidence. Since the robustness of integrative inference methods directly relies on the prior model, the data integration

procedure has to be thoroughly. Evaluation of the COGERE prior scoring scheme exhibits superior performance compared to common integration approaches.

The prior model was created for human and mouse, and permanently stored as data warehouse. Despite the current success of the COGERE prior score, ongoing progress in data collection by high-throughput '-omics' techniques will further improve the prior model. The COGERE database can be easily extended by recent information and any further organisms of interest. It was made accessible via a web-based user interface. To my best knowledge, this resource is the most elaborate collection integrating weighted transcriptional and post-transcriptional interactions yet. It delivers an insight to active and passive potential interactions for a gene of interest. Further, as I have shown exemplary, parameters and characteristics of the topology of miRNA-mediated GRN may be elucidated in future studies. Notably, the network comprise also transcriptional regulatory elements, such as miRNA promoters. Since trait-associated SNPs affecting miRNA biogenesis were rarely found in Chapter 4, complementing the presented analysis by information from the prior database may also denote a perspective to increase sensitivity.

Following information integration, the mutual dependency between a RG and a TG and the corresponding regulatory sign was inferred from gene expression data. For this purpose, the non-parametric, non-linear correlation coefficient η^2 was derived from a two-way ANOVA. The computation of this measure has a good scalability with respect to the number of genes, enabling the inference of large-scale networks. A comparative assessment on simulated and synthetic data showed that COGERE improves state-of-the-art GRN inference approaches. The application of COGERE to a real 60 cell line cancer expression panel demonstrated its potential for biologically meaningful hypothesis generation. COGERE was published as toolkit for academic use to allow the global GRNs inference for any regulatory landscape of interest. While COGERE was developed in the context of microarray analysis, it can be extended to other high-throughput methodologies that measure gene expression levels.

The development of GRN inference methods is still evolving. A major limitation of all studies in this field, is the lack of elaborate *in vivo* gold standards, i.e. expression profiles and the validated global, interconnected view of the system's transcriptional status, respectively. Integrative inference methods, such as COGERE, will be useful in guiding experimental designs to verify condition-specific regulatory interactions of interest.

Another relevant challenge that can be addressed by COGERE, is the identification of

differential gene connectivity that associate to a given phenotype. It has been reported that co-expression patterns rather than single gene expression variation determine phenotypic differences[325]. COGERE computes mutual dependencies between a regulatory pair from the effects of differential gene expression between given conditions. Thus, inferring GRNs for various conditions will enable the identification of connections changing between phenotypes.

COGERE examines the fraction of total variation explained by the variation in the expression data between conditions. Based on this principle, another future project is supposable: the network-assisted clustering of expression measurements. This is highly interesting with regard to, in example, the stratification of cancer subtypes to shed light on tumor heterogeneity, or the spatial and local classification of cells in single cell measurements. The general problem can be stated as follows: given unlabeled or anonymize expression data, i.e. no condition is defined, a finite set of categories has to be found. Indeed, algorithms solving this problem are emerging. However, the majority of existing analyses primarily applies generic clustering algorithms, such as hierarchical clustering[326], k-means[327], or principal component analysis[328]. Recently, it has been reported that considering the network information underlying the gene expression signals during the clustering process raises the biological relevance of the computed classes[329].

COGERE defines the cluster label by the factor condition in the two-way ANOVA. By sampling the factor levels, the global GRNs can be computed for all possible splits of the measurement set. Using the prior information, the updated class labels can be scored by the generated signal strength, e.g. by a function calculating the distance between the inferred model and the prior model. Since an increasing number of measurements induces a quickly growing search space, the grouping problem is NP-hard[330]. Thus, a search heuristic has to be implemented. With respect to the article of Hruschka *et al.*[331], I suppose that adapting the functions (fitness, selection, mutation, and crossover) of a Genetic Algorithm[330] may be appropriate to solve this optimization problem. Finally, the most valuable GRN and the corresponding clustering can be extracted.

In conclusion, since the first publication of an AGO-bound CLIP-Seq dataset, several studies, including this doctoral thesis, have revealed a wealth of novel insights to miRNA-mediated regulation. While it denotes a great progress in this field, the AGO-bound CLIP-Seq techniques do not resolve some impediments.

Firstly, the measurements produced by this protocol are condition-specific, i.e. the

major fraction of the miRNA and mRNA transcriptome is absent. Considering, that the RISC may bind transiently and multiple targets, this deficit becomes even more pronounced. To overcome this drawback, multiple experiments had to be conducted for various cell lines during the last years. At this, the protocol is difficult to perform restricting the number of successful applications. In example, PAR-CLIP requires cell lines pre-incubated with photoreactive ribonucleoside analogs. Such a treatment implies cell-specific limitations in nucleoside uptake, the likely incidence of toxic effects[332] or the occurrence of cross-linking biases (Chapter 3). Further, it has to be noted that CLIP experiments were mainly conducted using RNase T1. This endonuclease has a strong preference to cleave after guanines[116]. Thus, it has to be considered that extensive cleavage may result in sequence reads with a lowered fraction of G nucleotides. However, the amount of AGO CLIP-Seq data is still growing – a fact which raised the need for a central repository (starBase[195,333]). Its current implementation (version 2.0, release of 2014) contains measurements of 18 cell lines in human and 16 cell types in mouse. This enables to address further open questions, e.g. the miRNA targetome diversity across tissue types[334,335].

Secondly, AGO-bound CLIP-Seq data alone gives only information on the miRNP binding region, but no functional information. Thus, the type and strength of the effect caused by RISC binding needs to be measured by complementary experiments.

Lastly and most importantly, the identity of the guiding miRNA is still an unsolved question. Only the extraction of the full RISC:mRNA duplex will enable to display the whole target site details. Notably, while writing these lines, the novel promising miRNA cross-linking and immunoprecipitation (miR-CLIP) protocol was published[336,337]. Comparable to the AGO-bound CLIP-Seq technique, the cross-linked complex is isolated by immunoprecipitation and the purified target RNA fragments are characterized using high-throughput sequencing. The unique feature of this approach is the usage of a synthetic capture miRNA which makes bound mRNAs specific to a single miRNA. This information complements existing approaches, and with applications to a broad range of known miRNAs, a more detailed depiction of the miRNA targetome can be expected in the near future.

Bibliography

[1] Morris KV and Mattick JS. The rise of regulatory RNA. *Nat Rev Genet*, 15(6):423–437, 2014.

[2] Eddy SR. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet*, 2(12):919–929, 2001.

[3] Hoagland MB, Stephenson ML, Scott JF, Hecht LI, and Zamecnik PC. A soluble ribonucleic acid intermediate in protein synthesis. *J Biol Chem*, 231(1):241–257, 1958.

[4] Jacob F and Monod J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3:318–356, 1961.

[5] Britten RJ and Davidson EH. Gene regulation for higher cells: a theory. *Science*, 165(3891):349–357, 1969.

[6] Davidson EH, Klein WH, and Britten RJ. Sequence organization in animal DNA and a speculation on hnRNA as a coordinate regulatory transcript. *Dev Biol*, 55(1):69–84, 1977.

[7] Ohno S. So much "junk" DNA in our genome. *Brookhaven Symp Biol*, 23:366–370, 1972.

[8] Wightman B, Ha I, and Ruvkun G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5):855–862, 1993.

[9] Lee RC, Feinbaum RL, and Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, 1993.

[10] Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, and Ruvkun G. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772):901–906, 2000.

[11] Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Müller P, Spring J, Srinivasan A, Fishman M, Finnerty J, Corbo J, Levine M, Leahy P, Davidson E, and Ruvkun G. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89, 2000.

[12] Grishok A, Pasquinelli AE, Conte D, Li N, Parrish S, Ha I, Baillie DL, Fire A, Ruvkun G, and Mello CC. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*, 106(1):23–34, 2001.

[13] Sen GL and Blau HM. A brief history of RNAi: the silence of the genes. *FASEB J*, 20(9):1293–1299, 2006.

[14] Elbashir SM, Lendeckel W, and Tuschl T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev*, 15(2):188–200, 2001.

[15] Griffiths-Jones S. miRBase: microRNA sequences and annotation. *Curr Protoc Bioinformatics*, Chapter 12:Unit 12.9.1–Unit 12.910, 2010.

[16] Ha M and Kim VN. Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol*, 15(8):509–524, 2014.

[17] Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M, Eaves CJ, and Marra MA. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res*, 18(4):610–621, 2008.

[18] Rogers K and Chen X. Biogenesis, turnover, and mode of action of plant microRNAs. *Plant Cell*, 25(7):2383–2399, 2013.

[19] Rodriguez A, Griffiths-Jones S, Ashurst JL, and Bradley A. Identification of mammalian microRNA host genes and transcription units. *Genome Res*, 14(10A):1902–1910, 2004.

[20] Baskerville S and Bartel DP. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, 11(3):241–247, 2005.

[21] Ozsolak F, Poling LL, Wang Z, Liu H, Liu XS, Roeder RG, Zhang X, Song JS, and Fisher DE. Chromatin structure analyses identify miRNA promoters. *Genes Dev*, 22(22):3172–3183, 2008.

[22] Chang TC, Wentzel EA, Kent OA, Ramachandran K, Mullendore M, Lee KH, Feldmann G, Yamakuchi M, Ferlito M, Lowenstein CJ, Arking DE, Beer MA, Maitra A, and Mendell JT. Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis. *Mol Cell*, 26(5):745–752, 2007.

[23] Monteys AM, Spengler RM, Wan J, Tecedor L, Lennox KA, Xing Y, and Davidson BL. Structure and activity of putative intronic miRNA promoters. *RNA*, 16(3):495–505, 2010.

[24] Winter J, Jung S, Keller S, Gregory RI, and Diederichs S. Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat Cell Biol*, 11(3):228–234, 2009.

[25] Milosevic J, Pandit K, Magister M, Rabinovich E, Ellwanger DC, Yu G, Vuga LJ, Weksler B, Benos PV, Gibson KF, McMillan M, Kahn M, and Kaminski N. Profibrotic role of miR-154 in pulmonary fibrosis. *Am J Respir Cell Mol Biol*, 47(6):879–887, 2012.

[26] Lewis BP, Shih Ih, Jones-Rhoades MW, Bartel DP, and Burge CB. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–798, 2003.

[27] Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, 2009.

[28] Nielsen CB, Shomron N, Sandberg R, Hornstein E, Kitzman J, and Burge CB. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA*, 13(11):1894–1910, 2007.

[29] Babiarz JE, Ruby JG, Wang Y, Bartel DP, and Blelloch R. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes Dev*, 22(20):2773–2785, 2008.

[30] Chong MMW, Zhang G, Cheloufi S, Neubert TA, Hannon GJ, and Littman DR. Canonical and alternate functions of the microRNA biogenesis machinery. *Genes Dev*, 24(17):1951–1960, 2010.

[31] Miyoshi K, Miyoshi T, and Siomi H. Many ways to generate microRNA-like small RNAs: non-canonical pathways for microRNA production. *Mol Genet Genomics*, 284(2):95–103, 2010.

[32] Westholm JO and Lai EC. Mirtrons: microRNA biogenesis via splicing. *Biochimie*, 93(11):1897–1904, 2011.

[33] Xie M, Li M, Vilborg A, Lee N, Shu MD, Yartseva V, Šestan N, and Steitz JA. Mammalian 5'-capped microRNA precursors that generate a single microRNA. *Cell*, 155(7):1568–1580, 2013.

[34] Carthew RW and Sontheimer EJ. Origins and Mechanisms of miRNAs and siRNAs. *Cell*, 136(4):642–655, 2009.

[35] Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, Lin C, Socci ND, Hermida L, Fulci V, Chiaretti S, Foà R, Schliwka J, Fuchs U, Novosel A, Müller RU, Schermer B, Bissels U, Inman J, Phan Q, Chien M, Weir DB, Choksi R, De Vita G, Frezzetti D, Trompeter HI, Hornung V, Teng G, Hartmann G, Palkovits M, Di Lauro R, Wernet P, Macino G, Rogler CE, Nagle JW, Ju J, Papavasiliou FN, Benzing T, Lichter P, Tam W, Brownstein MJ, Bosio A, Borkhardt A, Russo JJ, Sander C, Zavolan M, and Tuschl T. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, 129(7):1401–1414, 2007.

[36] Wu H, Ye C, Ramirez D, and Manjunath N. Alternative processing of primary microRNA transcripts by Drosha generates 5' end variation of mature microRNA. *PLoS One*, 4(10):e7566, 2009.

[37] Kaya E and Doudna JA. Biochemistry. Guided tour to the heart of RISC. *Science*, 336(6084):985–986, 2012.

[38] Elkayam E, Kuhn CD, Tocilj A, Haase AD, Greene EM, Hannon GJ, and Joshua-Tor L. The structure of human argonaute-2 in complex with miR-20a. *Cell*, 150(1):100–110, 2012.

[39] Frank F, Sonenberg N, and Nagar B. Structural basis for 5'-nucleotide base-specific recognition of guide RNA by human AGO2. *Nature*, 465(7299):818–822, 2010.

[40] Ye X, Huang N, Liu Y, Paroo Z, Huerta C, Li P, Chen S, Liu Q, and Zhang H. Structure of C3PO and mechanism of human RISC activation. *Nat Struct Mol Biol*, 18(6):650–657, 2011.

[41] Eulalio A, Huntzinger E, and Izaurralde E. GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay. *Nat Struct Mol Biol*, 15(4):346–353, 2008.

[42] Bail S, Swerdel M, Liu H, Jiao X, Goff LA, Hart RP, and Kiledjian M. Differential regulation of microRNA stability. *RNA*, 16(5):1032–1039, 2010.

[43] Hausser J and Zavolan M. Identification and consequences of miRNA-target interactions–beyond repression of gene expression. *Nat Rev Genet*, 15(9):599–612, 2014.

[44] Rüegger S and Großhans H. MicroRNA turnover: when, how, and why. *Trends Biochem Sci*, 37(10):436–446, 2012.

[45] Winter J and Diederichs S. Argonaute proteins regulate microRNA stability: Increased microRNA abundance by Argonaute proteins is due to microRNA stabilization. *RNA Biol*, 8(6):1149–1157, 2011.

[46] Ameres SL, Martinez J, and Schroeder R. Molecular basis for target RNA recognition and cleavage by human RISC. *Cell*, 130(1):101–112, 2007.

[47] Wilson RC and Doudna JA. Molecular mechanisms of RNA interference. *Annu Rev Biophys*, 42:217–239, 2013.

[48] Brennecke J, Stark A, Russell RB, and Cohen SM. Principles of microRNA-target recognition. *PLoS Biol*, 3(3):e85, 2005.

[49] Kertesz M, Iovino N, Unnerstall U, Gaul U, and Segal E. The role of site accessibility in microRNA target recognition. *Nat Genet*, 39(10):1278–1284, 2007.

[50] Grimson A, Farh KKH, Johnston WK, Garrett-Engele P, Lim LP, and Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, 27(1):91–105, 2007.

[51] Huntzinger E and Izaurralde E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet*, 12(2):99–110, 2011.

[52] Filipowicz W, Bhattacharyya SN, and Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet*, 9(2):102–114, 2008.

[53] Vasudevan S and Steitz JA. AU-rich-element-mediated upregulation of translation by FXR1 and Argonaute 2. *Cell*, 128(6):1105–1118, 2007.

[54] Vasudevan S, Tong Y, and Steitz JA. Switching from repression to activation: microRNAs can up-regulate translation. *Science*, 318(5858):1931–1934, 2007.

[55] Ebert MS and Sharp PA. Roles for microRNAs in conferring robustness to biological processes. *Cell*, 149(3):515–524, 2012.

[56] Mukherji S, Ebert MS, Zheng GXY, Tsang JS, Sharp PA, and van Oudenaarden A. MicroRNAs can generate thresholds in target gene expression. *Nat Genet*, 43(9):854–859, 2011.

[57] Friedman RC, Farh KKH, Burge CB, and Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19(1):92–105, 2009.

[58] Salmena L, Poliseno L, Tay Y, Kats L, and Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, 146(3):353–358, 2011.

[59] Tay Y, Rinn J, and Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. *Nature*, 505(7483):344–352, 2014.

[60] Cohen EEW, Zhu H, Lingen MW, Martin LE, Kuo WL, Choi EA, Kocherginsky M, Parker JS, Chung CH, and Rosner MR. A feed-forward loop involving protein kinase Calpha and microRNAs regulates tumor cell cycle. *Cancer Res*, 69(1):65–74, 2009.

[61] Enright AJ, John B, Gaul U, Tuschl T, Sander C, and Marks DS. MicroRNA targets in Drosophila. *Genome Biol*, 5(1):R1, 2003.

[62] Shalgi R, Lieber D, Oren M, and Pilpel Y. Global and local architecture of the mammalian microRNA-transcription factor regulatory network. *PLoS Comput Biol*, 3(7):e131, 2007.

[63] Hornstein E and Shomron N. Canalization of development by microRNAs. *Nat Genet*, 38 Suppl:S20–S24, 2006.

[64] Mangan S and Alon U. Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci U S A*, 100(21):11980–11985, 2003.

[65] Ørom UA and Lund AH. Experimental identification of microRNA targets. *Gene*, 451(1-2):1–5, 2010.

[66] Zhang L, Ding L, Cheung TH, Dong MQ, Chen J, Sewell AK, Liu X, Yates JR 3rd, and Han M. Systematic identification of C. elegans miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. *Mol Cell*, 28(4):598–613, 2007.

[67] Easow G, Teleman AA, and Cohen SM. Isolation of microRNA targets by miRNP immunopurification. *RNA*, 13(8):1198–1204, 2007.

[68] Hammell M, Long D, Zhang L, Lee A, Carmack CS, Han M, Ding Y, and Ambros V. mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. *Nat Methods*, 5(9):813–819, 2008.

[69] Chodosh LA. UV crosslinking of proteins to nucleic acids. *Curr Protoc Mol Biol*, Chapter 12:Unit 12.5, 2001.

[70] Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, and Tuschl T. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, 2010.

[71] Chi SW, Zang JB, Mele A, and Darnell RB. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–486, 2009.

[72] Broughton JP and Pasquinelli AE. Identifying Argonaute binding sites in Caenorhabditis elegans using iCLIP. *Methods*, 63(2):119–125, 2013.

[73] Helwak A, Kudla G, Dudnakova T, and Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, 153(3):654–665, 2013.

[74] Helwak A and Tollervey D. Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). *Nat Protoc*, 9(3):711–728, 2014.

[75] Esteller M. Non-coding RNAs in human disease. *Nat Rev Genet*, 12(12):861–874, 2011.

[76] Kota J, Chivukula RR, O'Donnell KA, Wentzel EA, Montgomery CL, Hwang HW, Chang TC, Vivekanandan P, Torbenson M, Clark KR, Mendell JR, and Mendell JT. Therapeutic microRNA delivery suppresses tumorigenesis in a murine liver cancer model. *Cell*, 137(6):1005–1017, 2009.

[77] Lanford RE, Hildebrandt-Eriksen ES, Petri A, Persson R, Lindow M, Munk ME, Kauppinen S, and Ørum H. Therapeutic silencing of microRNA-122 in primates with chronic hepatitis C virus infection. *Science*, 327(5962):198–201, 2010.

[78] Ellwanger DC, Büttner FA, Mewes HW, and Stümpflen V. The sufficient minimal set of miRNA seed types. *Bioinformatics*, 27(10):1346–1350, 2011.

[79] Königshoff M, Kramer M, Balsara N, Wilhelm J, Amarie OV, Jahn A, Rose F, Fink L, Seeger W, Schaefer L, Günther A, and Eickelberg O. WNT1-inducible signaling protein-1 mediates pulmonary fibrosis in mice and is upregulated in humans with idiopathic pulmonary fibrosis. *J Clin Invest*, 119(4):772–787, 2009.

[80] Berschneider B, Ellwanger DC, Baarsma HA, Thiel C, Shimbori C, White ES, Kolb M, Neth P, and Königshoff M. miR-92a regulates TGF-β1-induced WISP1 expression in pulmonary fibrosis. *Int J Biochem Cell Biol*, 53:432–441, 2014.

[81] Arnold M, Ellwanger DC, Hartsperger ML, Pfeufer A, and Stümpflen V. Cis-acting polymorphisms affect complex traits through modifications of microRNA regulation pathways. *PLoS One*, 7(5):e36694, 2012.

[82] Ellwanger DC, Leonhardt JF, and Mewes HW. Large-scale modeling of condition-specific gene regulatory networks by information integration and inference. *Nucleic Acids Res*, 42(21):0, 2014.

[83] Bagga S, Bracht J, Hunter S, Massirer K, Holtz J, Eachus R, and Pasquinelli AE. Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell*, 122(4):553–563, 2005.

[84] Guo H, Ingolia NT, Weissman JS, and Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–840, 2010.

[85] Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, and Johnson JM. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–773, 2005.

[86] Hausser J, Landthaler M, Jaskiewicz L, Gaidatzis D, and Zavolan M. Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. *Genome Res*, 19(11):2009–2020, 2009.

[87] Baek D, Villén J, Shin C, Camargo FD, Gygi SP, and Bartel DP. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, 2008.

[88] Farh KKH, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, and Bartel DP. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, 310(5755):1817–1821, 2005.

[89] Xu J, Zhang R, Shen Y, Liu G, Lu X, and Wu CI. The evolution of evolvability in microRNA target sites in vertebrates. *Genome Res*, 23(11):1810–1816, 2013.

[90] Lewis BP, Burge CB, and Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, 2005.

[91] Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, and Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, 2008.

[92] Pruitt KD, Tatusova T, Klimke W, and Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res*, 37(Database issue):D32–D36, 2009.

[93] Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, and Kent WJ. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, 32(Database issue):D493–D496, 2004.

[94] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, and Wheeler DL. GenBank. *Nucleic Acids Res*, 33(Database issue):D34–D38, 2005.

[95] Fisher RA. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85 (1):87–94, 1922.

[96] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, and Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–1050, 2005.

[97] Betel D, Wilson M, Gabow A, Marks DS, and Sander C. The microRNA.org resource: targets and expression. *Nucleic Acids Res*, 36(Database issue):D149–D153, 2008.

[98] Lau NC, Lim LP, Weinstein EG, and Bartel DP. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294(5543):858–862, 2001.

[99] Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, Xu J, Kittler ELW, Zapp ML, Weng Z, and Zamore PD. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. *Science*, 320(5879):1077–1081, 2008.

[100] Ghildiyal M, Xu J, Seitz H, Weng Z, and Zamore PD. Sorting of Drosophila small silencing RNAs partitions microRNA* strands into the RNA interference pathway. *RNA*, 16(1):43–56, 2010.

[101] Stark A, Brennecke J, Bushati N, Russell RB, and Cohen SM. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123(6):1133–1146, 2005.

[102] Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, and Rajewsky N. Combinatorial microRNA target predictions. *Nat Genet*, 37(5):495–500, 2005.

[103] Gaidatzis D, van Nimwegen E, Hausser J, and Zavolan M. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, 8:69, 2007.

[104] Sturm M, Hackenberg M, Langenberger D, and Frishman D. TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics*, 11:292, 2010.

[105] Marín RM and Vanícek J. Efficient use of accessibility in microRNA target prediction. *Nucleic Acids Res*, 39(1):19–29, 2011.

[106] Krüger J and Rehmsmeier M. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res*, 34(Web Server issue):W451–W454, 2006.

[107] Busch A, Richter AS, and Backofen R. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–2856, 2008.

[108] Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, and Hatzigeorgiou AG. Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, 25(23):3049–3055, 2009.

[109] Sevignani C, Calin GA, Siracusa LD, and Croce CM. Mammalian microRNAs: a small world for fine-tuning gene expression. *Mamm Genome*, 17(3):189–202, 2006.

[110] Lai EC. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet*, 30(4):363–364, 2002.

[111] Xiao F, Zuo Z, Cai G, Kang S, Gao X, and Li T. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res*, 37(Database issue):D105–D110, 2009.

[112] Sethupathy P, Megraw M, and Hatzigeorgiou AG. A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat Methods*, 3(11):881–886, 2006.

[113] Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, Lim B, and Rigoutsos I. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, 126(6):1203–1217, 2006.

[114] Rehmsmeier M, Steffen P, Hochsmann M, and Giegerich R. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517, 2004.

[115] Zisoulis DG, Lovci MT, Wilbert ML, Hutt KR, Liang TY, Pasquinelli AE, and Yeo GW. Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. *Nat Struct Mol Biol*, 17(2):173–179, 2010.

[116] Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, and Zavolan M. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods*, 8(7):559–564, 2011.

[117] Fang Z and Rajewsky N. The impact of miRNA target sites in coding sequences and in 3'UTRs. *PLoS One*, 6(3):e18067, 2011.

[118] Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, García-Girón C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kähäri AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ritchie GRS, Ruffier M, Schuster M, Sheppard

D, Sobral D, Taylor K, Thormann A, Trevanion S, White S, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Harrow J, Herrero J, Hubbard TJP, Johnson N, Kinsella R, Parker A, Spudich G, Yates A, Zadissa A, and Searle SMJ. Ensembl 2013. *Nucleic Acids Res*, 41(Database issue):D48–D55, 2013.

[119] Cooper GM, Stone EA, Asimenos G, NISCCSP, Green ED, Batzoglou S, and Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*, 15(7):901–913, 2005.

[120] Wen J, Parker BJ, Jacobsen A, and Krogh A. MicroRNA transfection and AGO-bound CLIP-seq data sets reveal distinct determinants of miRNA action. *RNA*, 17(5):820–834, 2011.

[121] Wen J, Parker BJ, and Weiller GF. In Silico identification and characterization of mRNA-like noncoding transcripts in Medicago truncatula. *In Silico Biol*, 7(4-5):485–505, 2007.

[122] Kiryu H, Terai G, Imamura O, Yoneyama H, Suzuki K, and Asai K. A detailed investigation of accessibilities around target sites of siRNAs and miRNAs. *Bioinformatics*, 27(13):1788–1797, 2011.

[123] Needleman SB and Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, 1970.

[124] Rigoutsos I and Floratos A. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics*, 14(1):55–67, 1998.

[125] Mitchell T. *Machine Learning*. McGraw Hill, 1997.

[126] Alpaydin E. *Introduction to Machine Learning*. MIT Press, 2004.

[127] Cortes C and Vapnik V. Support-vector networks. *Machine Learning*, 20 (3):273, 1995.

[128] Liu H, Yue D, Chen Y, Gao SJ, and Huang Y. Improving performance of mammalian microRNA target prediction. *BMC Bioinformatics*, 11:476, 2010.

[129] Chang C and Lin C. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1, 2011.

[130] Linsley PS, Schelter J, Burchard J, Kibukawa M, Martin MM, Bartz SR, Johnson JM, Cummins JM, Raymond CK, Dai H, Chau N, Cleary M, Jackson AL, Carleton M, and Lim L. Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Mol Cell Biol*, 27(6):2240–2252, 2007.

[131] Fisher R. *Statistical Methods for Research Workers*. Oliver and Boyd (Edinburgh), 1925.

[132] Pandit KV, Corcoran D, Yousef H, Yarlagadda M, Tzouvelekis A, Gibson KF, Konishi K, Yousem SA, Singh M, Handley D, Richards T, Selman M, Watkins SC, Pardo A, Ben-Yehudah A, Bouros D, Eickelberg O, Ray P, Benos PV, and Kaminski N. Inhibition and role of let-7d in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med*, 182(2):220–229, 2010.

[133] Cho JH, Gelinas R, Wang K, Etheridge A, Piper MG, Batte K, Dakhallah D, Price J, Bornman D, Zhang S, Marsh C, and Galas D. Systems biology of interstitial lung diseases: integration of mRNA and microRNA expression changes. *BMC Med Genomics*, 4:8, 2011.

[134] Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, and Smyth GK. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23(20):2700–2707, 2007.

[135] Bolstad BM, Irizarry RA, Astrand M, and Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[136] Liu H, D'Andrade P, Fulmer-Smentek S, Lorenzi P, Kohn KW, Weinstein JN, Pommier Y, and Reinhold WC. mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. *Mol Cancer Ther*, 9(5):1080–1091, 2010.

[137] Livak KJ and Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods*, 25(4):402–408, 2001.

[138] Chargaff E. Structure and function of nucleic acids as cell constituents. *Fed Proc*, 10(3):654–659, 1951.

[139] Xia T, SantaLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, and Turner DH. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37(42):14719–14735, 1998.

[140] Varani G and McClain WH. The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep*, 1(1):18–23, 2000.

[141] Ascano M, Hafner M, Cekan P, Gerstberger S, and Tuschl T. Identification of RNA-protein interaction networks using PAR-CLIP. *Wiley Interdiscip Rev RNA*, 3(2):159–177, 2012.

[142] Betel D, Koppal A, Agius P, Sander C, and Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol*, 11(8):R90, 2010.

[143] Raghu G, Collard HR, Egan JJ, Martinez FJ, Behr J, Brown KK, Colby TV, Cordier JF, Flaherty KR, Lasky JA, Lynch DA, Ryu JH, Swigris JJ, Wells AU, Ancochea J, Bouros D, Carvalho C, Costabel U, Ebina M, Hansell DM, Johkoh T, Kim DS, King TE Jr, Kondoh Y, Myers J, Müller NL, Nicholson AG, Richeldi L, Selman M, Dudden RF, Griss BS, Protzko SL, Schünemann HJ, and ATSRSRSLATCoIPF. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med*, 183(6):788–824, 2011.

[144] Baroke E, Gauldie J, and Kolb M. New treatment and markers of prognosis for idiopathic pulmonary fibrosis: lessons learned from translational research. *Expert Rev Respir Med*, 7(5):465–478, 2013.

[145] du Bois RM. Strategies for treating idiopathic pulmonary fibrosis. *Nat Rev Drug Discov*, 9(2):129–140, 2010.

[146] Holgate ST, Davies DE, Lackie PM, Wilson SJ, Puddicombe SM, and Lordan JL. Epithelial-mesenchymal interactions in the pathogenesis of asthma. *J Allergy Clin Immunol*, 105(2 Pt 1):193–204, 2000.

[147] Fernandez IE and Eickelberg O. The impact of TGF-β on lung fibrosis: from targeting to biomarkers. *Proc Am Thorac Soc*, 9(3):111–116, 2012.

[148] Königshoff M and Eickelberg O. WNT signaling in lung disease: a failure or a regeneration signal? *Am J Respir Cell Mol Biol*, 42(1):21–31, 2010.

[149] Lam AP and Gottardi CJ. β-catenin signaling: a novel mediator of fibrosis and potential therapeutic target. *Curr Opin Rheumatol*, 23(6):562–567, 2011.

[150] Xiao J, Meng XM, Huang XR, Chung AC, Feng YL, Hui DS, Yu CM, Sung JJ, and Lan HY. miR-29 inhibits bleomycin-induced pulmonary fibrosis in mice. *Mol Ther*, 20(6):1251–1260, 2012.

[151] Yang S, Banerjee S, de Freitas A, Sanders YY, Ding Q, Matalon S, Thannickal VJ, Abraham E, and Liu G. Participation of miR-200 in pulmonary fibrosis. *Am J Pathol*, 180(2):484–493, 2012.

[152] Liu G, Friggeri A, Yang Y, Milosevic J, Ding Q, Thannickal VJ, Kaminski N, and Abraham E. miR-21 mediates fibrogenic activation of pulmonary fibroblasts and lung fibrosis. *J Exp Med*, 207(8):1589–1597, 2010.

[153] Sime PJ, Xing Z, Graham FL, Csaky KG, and Gauldie J. Adenovector-mediated gene transfer of active transforming growth factor-beta1 induces prolonged severe fibrosis in rat lung. *J Clin Invest*, 100(4):768–776, 1997.

[154] Karlin S and Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*, 87(6):2264–2268, 1990.

[155] Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, 1990.

[156] Blaauboer ME, Emson CL, Verschuren L, van Erk M, Turner SM, Everts V, Hanemaaijer R, and Stoop R. Novel combination of collagen dynamics analysis and transcriptional profiling reveals fibrosis-relevant genes and pathways. *Matrix Biol*, 32(7-8):424–431, 2013.

[157] Ventura A, Young AG, Winslow MM, Lintault L, Meissner A, Erkeland SJ, Newman J, Bronson RT, Crowley D, Stone JR, Jaenisch R, Sharp PA, and Jacks T. Targeted deletion reveals essential and overlapping functions of the miR-17 through 92 family of miRNA clusters. *Cell*, 132(5):875–886, 2008.

[158] Lu Y, Thomson JM, Wong HYF, Hammond SM, and Hogan BLM. Transgenic over-expression of the microRNA miR-17-92 cluster promotes proliferation and inhibits differentiation of lung epithelial progenitor cells. *Dev Biol*, 310(2):442–453, 2007.

[159] Heinrich EM, Wagner J, Krüger M, John D, Uchida S, Weigand JE, Suess B, and Dimmeler S. Regulation of miR-17-92a cluster processing by the microRNA binding protein SND1. *FEBS Lett*, 587(15):2405–2411, 2013.

[160] Dakhlallah D, Batte K, Wang Y, Cantemir-Stone CZ, Yan P, Nuovo G, Mikhail A, Hitchcock CL, Wright VP, Nana-Sinkam SP, Piper MG, and Marsh CB. Epigenetic regulation of miR-17 92 contributes to the pathogenesis of pulmonary fibrosis. *Am J Respir Crit Care Med*, 187(4):397–405, 2013.

[161] Avraham R and Yarden Y. Regulation of signalling by microRNAs. *Biochem Soc Trans*, 40(1):26–30, 2012.

[162] Ambros V. The functions of animal microRNAs. *Nature*, 431(7006):350–355, 2004.

[163] Ventura A and Jacks T. MicroRNAs and cancer: short RNAs go a long way. *Cell*, 136(4):586–591, 2009.

[164] Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, and Liu Y. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res*, 37(Database issue):D98–104, 2009.

[165] Ruepp A, Kowarsch A, and Theis F. PhenomiR: microRNAs in human diseases and biological processes. *Methods Mol Biol*, 822:249–260, 2012.

[166] Gupta SK, Bang C, and Thum T. Circulating microRNAs as biomarkers and potential paracrine mediators of cardiovascular disease. *Circ Cardiovasc Genet*, 3(5):484–488, 2010.

[167] Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, and Golub TR. MicroRNA expression profiles classify human cancers. *Nature*, 435(7043):834–838, 2005.

[168] Rosenfeld N, Aharonov R, Meiri E, Rosenwald S, Spector Y, Zepeniuk M, Benjamin H, Shabes N, Tabak S, Levy A, Lebanony D, Goren Y, Silberschein E, Targan N, Ben-Ari A, Gilad S, Sion-Vardy N, Tobar A, Feinmesser M, Kharenko O, Nativ O, Nass D, Perelman M, Yosepovich A, Shalmon B, Polak-Charcon S, Fridman E, Avniel A, Bentwich I, Bentwich Z, Cohen D, Chajut A, and Barshack I. MicroRNAs accurately identify cancer tissue origin. *Nat Biotechnol*, 26(4):462–469, 2008.

[169] Ryan BM, Robles AI, and Harris CC. Genetic variation in microRNA networks: the implications for cancer research. *Nat Rev Cancer*, 10(6):389–402, 2010.

[170] Consortium IH, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JTF, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y,

Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PIW, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CDM, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, and Stewart J. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, 2007.

[171] Schork NJ, Murray SS, Frazer KA, and Topol EJ. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev*, 19(3):212–219, 2009.

[172] Shields R. Common disease: are causative alleles common or rare? *PLoS Biol*, 9(1):e1001009, 2011.

[173] Baker M. Genomics: The search for association. *Nature*, 467(7319):1135–1138, 2010.

[174] Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S,

Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, and Stamatoyannopoulos JA. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, 2012.

[175] Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, and Cheung VG. Common genetic variants account for differences in gene expression among ethnic groups. *Nat Genet*, 39(2):226–231, 2007.

[176] Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, Morgan TM, Mathews CA, Pauls DL, Rasin MR, Gunel M, Davis NR, Ercan-Sencicek AG, Guez DH, Spertus JA, Leckman JF, Dure LS 4th, Kurlan R, Singer HS, Gilbert DL, Farhi A, Louvi A, Lifton RP, Sestan N, and State MW. Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science*, 310(5746):317–320, 2005.

[177] Saunders MA, Liang H, and Li WH. Human polymorphism at microRNAs and microRNA target sites. *Proc Natl Acad Sci U S A*, 104(9):3300–3305, 2007.

[178] Duan R, Pak C, and Jin P. Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA. *Hum Mol Genet*, 16(9):1124–1131, 2007.

[179] Wu M, Jolicoeur N, Li Z, Zhang L, Fortin Y, L'Abbe D, Yu Z, and Shen SH. Genetic variations of microRNAs in human cancer and their effects on the expression of miRNAs. *Carcinogenesis*, 29(9):1710–1716, 2008.

[180] Meola N, Gennarino VA, and Banfi S. microRNAs and genetic diseases. *Pathogenetics*, 2(1):7, 2009.

[181] Richardson K, Lai CQ, Parnell LD, Lee YC, and Ordovas JM. A genome-wide survey for SNPs altering microRNA seed sites identifies functional candidates in GWAS. *BMC Genomics*, 12:504, 2011.

[182] Georges M. The long and winding road from correlation to causation. *Nat Genet*, 43(3):180–181, 2011.

[183] Bennett CL, Brunkow ME, Ramsdell F, O'Briant KC, Zhu Q, Fuleihan RL, Shigeoka AO, Ochs HD, and Chance PF. A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA–>AAUGAA) leads to the IPEX syndrome. *Immunogenetics*, 53(6):435–439, 2001.

[184] Walters RW, Bradrick SS, and Gromeier M. Poly(A)-binding protein modulates mRNA susceptibility to cap-dependent miRNA-mediated repression. *RNA*, 16(1):239–250, 2010.

[185] Li X, Quon G, Lipshitz HD, and Morris Q. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, 16(6):1096–1107, 2010.

[186] Haas U, Sczakiel G, and Laufer SD. MicroRNA-mediated regulation of gene expression is affected by disease-associated SNPs within the 3'-UTR via altered RNA structure. *RNA Biol*, 9(6):924–937, 2012.

[187] Yang JO, Kim WY, and Bhak J. ssSNPTarget: genome-wide splice-site Single Nucleotide Polymorphism database. *Hum Mutat*, 30(12):E1010–E1020, 2009.

[188] Consortium IH. The International HapMap Project. *Nature*, 426(6968):789–796, 2003.

[189] Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, and Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*, 42(Database issue):D1001–D1006, 2014.

[190] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, and Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, 106(23):9362–9367, 2009.

[191] Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, and de Bakker PIW. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, 24(24):2938–2939, 2008.

[192] Bush WS and Moore JH. Chapter 11: Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822, 2012.

[193] Pruitt KD, Tatusova T, Brown GR, and Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res*, 40(Database issue):D130–D135, 2012.

[194] Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Barber GP, Haussler D, and Kent WJ. The UCSC Genome Browser database: update 2010. *Nucleic Acids Res*, 38(Database issue):D613–D619, 2010.

[195] Yang JH, Li JH, Shao P, Zhou H, Chen YQ, and Qu LH. starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Res*, 39(Database issue):D202–D209, 2011.

[196] Lee JY, Yeh I, Park JY, and Tian B. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res*, 35(Database issue):D165–D168, 2007.

[197] Beaudoing E, Freier S, Wyatt JR, Claverie JM, and Gautheret D. Patterns of variant polyadenylation signal usage in human genes. *Genome Res*, 10(7):1001–1010, 2000.

[198] Reese MG, Eeckman FH, Kulp D, and Haussler D. Improved splice site detection in Genie. *J Comput Biol*, 4(3):311–323, 1997.

[199] Schwarz JM, Rödelsperger C, Schuelke M, and Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods*, 7(8):575–576, 2010.

[200] Hofacker IL. Vienna RNA secondary structure server. *Nucleic Acids Res*, 31(13):3429–3431, 2003.

[201] Halvorsen M, Martin JS, Broadaway S, and Laederach A. Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet*, 6(8):e1001074, 2010.

[202] Onwuegbuzie AJ, Daniel L, and Leech NL. *Encyclopedia of Measurement and Statistics*, chapter Pearson Product-Moment Correlation Coefficient, pages 751–756. SAGE Publications, Inc., 2007.

[203] Consortium IH, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PIW, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Gibbs RA, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarrol SA, Nemesh J, Dermitzakis E, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Bonnen PE, Gibbs RA, Gonzaga-Jauregui C, Keinan A, Price AL, Yu F, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Schaffner SF, Zhang Q, Ghori MJR, McGinnis R, McLaren W, Pollack S, Price AL, Schaffner SF, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, and McEwen JE. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.

[204] Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, Travers M, Potter S, Grundberg E, Small K, Hedman AK, Bataille V, Tzenova Bell J, Surdulescu G, Dimas AS, Ingle C, Nestle FO, di Meglio P, Min JL, Wilk A, Hammond CJ, Hassanali N, Yang TP, Montgomery SB, O'Rahilly S, Lindgren CM, Zondervan KT, Soranzo N, Barroso I, Durbin R, Ahmadi K, Deloukas P, McCarthy MI, Dermitzakis ET, Spector TD, and MHERC. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet*, 7(2):e1002003, 2011.

[205] Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, Sekowska M, Smith GD, Evans D, Gutierrez-Arcelus M, Price A, Raj T, Nisbett J, Nica AC, Beazley C, Durbin R, Deloukas P, and Dermitzakis ET. Patterns of cis regulatory variation in diverse human populations. *PLoS Genet*, 8(4):e1002639, 2012.

[206] Myers JL and Well AD. *Research Design and Statistical Analysis*. Lawrence Erlbaum Assoc Inc., 2nd edition, 2003.

[207] Yang TP, Beazley C, Montgomery SB, Dimas AS, Gutierrez-Arcelus M, Stranger BE, Deloukas P, and Dermitzakis ET. Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*, 26(19):2474–2476, 2010.

[208] Kaern M, Elston TC, Blake WJ, and Collins JJ. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet*, 6(6):451–464, 2005.

[209] Zhang R and Su B. MicroRNA regulation and the variability of human cortical gene expression. *Nucleic Acids Res*, 36(14):4621–4628, 2008.

[210] Lu J and Clark AG. Impact of microRNA regulation on variation in human gene expression. *Genome Res*, 22(7):1243–1254, 2012.

[211] Becker KG, Barnes KC, Bright TJ, and Wang SA. The genetic association database. *Nat Genet*, 36(5):431–432, 2004.

[212] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, 2000.

[213] Fogarty MP, Xiao R, Prokunina-Olsson L, Scott LJ, and Mohlke KL. Allelic expression imbalance at high-density lipoprotein cholesterol locus MMAB-MVK. *Hum Mol Genet*, 19(10):1921–1929, 2010.

[214] Hirschfield GM, Liu X, Xu C, Lu Y, Xie G, Lu Y, Gu X, Walker EJ, Jing K, Juran BD, Mason AL, Myers RP, Peltekian KM, Ghent CN, Coltescu C, Atkinson EJ, Heathcote EJ, Lazaridis KN, Amos CI, and Siminovitch KA. Primary biliary cirrhosis associated with HLA, IL12A, and IL12RB2 variants. *N Engl J Med*, 360(24):2544–2555, 2009.

[215] Padgett KA, Lan RY, Leung PC, Lleo A, Dawson K, Pfeiff J, Mao TK, Coppel RL, Ansari AA, and Gershwin ME. Primary biliary cirrhosis is associated with altered hepatic microRNA expression. *J Autoimmun*, 32(3-4):246–253, 2009.

[216] Onn I, Aono N, Hirano M, and Hirano T. Reconstitution and subunit geometry of human condensin complexes. *EMBO J*, 26(4):1024–1034, 2007.

[217] Heale JT, Ball AR Jr, Schmiesing JA, Kim JS, Kong X, Zhou S, Hudson DF, Earnshaw WC, and Yokomori K. Condensin I interacts with the PARP-1-XRCC1 complex and functions in DNA single-strand break repair. *Mol Cell*, 21(6):837–848, 2006.

[218] Zindy P, Andrieux L, Bonnier D, Musso O, Langouët S, Campion JP, Turlin B, Clément B, and Théret N. Upregulation of DNA repair genes in active cirrhosis associated with hepatocellular carcinoma. *FEBS Lett*, 579(1):95–99, 2005.

[219] Szwergold B, Manevich Y, Payne L, and Loomes K. Fructosamine-3-kinase-related-protein phosphorylates glucitolamines on the C-4 hydroxyl: novel substrate specificity of an enigmatic enzyme. *Biochem Biophys Res Commun*, 361(4):870–875, 2007.

[220] Delpierre G, Veiga-da Cunha M, Vertommen D, Buysschaert M, and Van Schaftingen E. Variability in erythrocyte fructosamine 3-kinase activity in humans correlates with polymorphisms in the FN3K gene and impacts on haemoglobin glycation at specific sites. *Diabetes Metab*, 32(1):31–39, 2006.

[221] Köttgen A, Albrecht E, Teumer A, Vitart V, Krumsiek J, Hundertmark C, Pistis G, Ruggiero D, O'Seaghdha CM, Haller T, Yang Q, Tanaka T, Johnson AD, Kutalik Z, Smith AV, Shi J, Struchalin M, Middelberg RPS, Brown MJ, Gaffo AL, Pirastu N, Li G, Hayward C, Zemunik T, Huffman J, Yengo L, Zhao JH, Demirkan A, Feitosa MF, Liu X, Malerba G, Lopez LM, van der Harst P, Li X, Kleber ME, Hicks AA, Nolte IM, Johansson A, Murgia F, Wild SH, Bakker SJL, Peden JF, Dehghan A, Steri M, Tenesa A, Lagou V, Salo P, Mangino M, Rose LM, Lehtimäki T, Woodward OM, Okada Y, Tin A, Müller C, Oldmeadow C, Putku M, Czamara D, Kraft P, Frogheri L, Thun GA, Grotevendt A, Gislason GK, Harris TB, Launer LJ, McArdle P, Shuldiner AR, Boerwinkle E, Coresh J, Schmidt H, Schallert M, Martin NG, Montgomery GW, Kubo M, Nakamura Y, Tanaka T, Munroe PB, Samani NJ, Jacobs DR Jr, Liu K, D'Adamo P, Ulivi S, Rotter JI, Psaty BM, Vollenweider P, Waeber G, Campbell S, Devuyst O, Navarro P, Kolcic I, Hastie N, Balkau B, Froguel P, Esko T, Salumets A,

Khaw KT, Langenberg C, Wareham NJ, Isaacs A, Kraja A, Zhang Q, Wild PS, Scott RJ, Holliday EG, Org E, Viigimaa M, Bandinelli S, Metter JE, Lupo A, Trabetti E, Sorice R, Döring A, Lattka E, Strauch K, Theis F, Waldenberger M, Wichmann HE, Davies G, Gow AJ, Bruinenberg M, LCS, Stolk RP, Kooner JS, Zhang W, Winkelmann BR, Boehm BO, Lucae S, Penninx BW, Smit JH, Curhan G, Mudgal P, Plenge RM, Portas L, Persico I, Kirin M, Wilson JF, Mateo Leach I, van Gilst WH, Goel A, Ongen H, Hofman A, Rivadeneira F, Uitterlinden AG, Imboden M, von Eckardstein A, Cucca F, Nagaraja R, Piras MG, Nauck M, Schurmann C, Budde K, Ernst F, Farrington SM, Theodoratou E, Prokopenko I, Stumvoll M, Jula A, Perola M, Salomaa V, Shin SY, Spector TD, Sala C, Ridker PM, Kähönen M, Viikari J, Hengstenberg C, Nelson CP, CARDIRAMC, DIAGRAMC, ICBPC, MAGICC, Meschia JF, Nalls MA, Sharma P, Singleton AB, Kamatani N, Zeller T, Burnier M, Attia J, Laan M, Klopp N, Hillege HL, Kloiber S, Choi H, Pirastu M, Tore S, Probst-Hensch NM, Völzke H, Gudnason V, Parsa A, Schmidt R, Whitfield JB, Fornage M, Gasparini P, Siscovick DS, Polašek O, Campbell H, Rudan I, Bouatia-Naji N, Metspalu A, Loos RJF, van Duijn CM, Borecki IB, Ferrucci L, Gambaro G, Deary IJ, Wolffenbuttel BHR, Chambers JC, März W, Pramstaller PP, Snieder H, Gyllensten U, Wright AF, Navis G, Watkins H, Witteman JCM, Sanna S, Schipf S, Dunlop MG, Tönjes A, Ripatti S, Soranzo N, Toniolo D, Chasman DI, Raitakari O, Kao WHL, Ciullo M, Fox CS, Caulfield M, Bochud M, and Gieger C. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat Genet*, 45(2):145–154, 2013.

[222] Feig DI, Mazzali M, Kang DH, Nakagawa T, Price K, Kannelis J, and Johnson RJ. Serum uric acid: a risk factor and a target for treatment? *J Am Soc Nephrol*, 17(4 Suppl 2):S69–S73, 2006.

[223] Hasselbacher P and Schumacher HR. Immunoglobulin in tophi and on the surface of monosodium urate crystals. *Arthritis Rheum*, 21(3):353–361, 1978.

[224] Campion EW, Glynn RJ, and DeLabry LO. Asymptomatic hyperuricemia. Risks and consequences in the Normative Aging Study. *Am J Med*, 82(3):421–426, 1987.

[225] Clop A, Marcq F, Takeda H, Pirottin D, Tordoir X, Bibé B, Bouix J, Caiment F, Elsen JM, Eychenne F, Larzul C, Laville E, Meish F, Milenkovic D, Tobin J, Charlier

C, and Georges M. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat Genet*, 38(7):813–818, 2006.

[226] Delay C, Calon F, Mathews P, and Hébert SS. Alzheimer-specific variants in the 3'UTR of Amyloid precursor protein affect microRNA function. *Mol Neurodegener*, 6:70, 2011.

[227] Carlini DB and Genut JE. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J Mol Evol*, 62(1):89–98, 2006.

[228] Black DL. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, 72:291–336, 2003.

[229] Consortium GP, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, and McVean GA. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[230] Visscher PM and Montgomery GW. Genome-wide association studies and human disease: from trickle to flood. *JAMA*, 302(18):2028–2029, 2009.

[231] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, and Visscher PM. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[232] Benayoun BA, Caburet S, and Veitia RA. Forkhead transcription factors: key players in health and disease. *Trends Genet*, 27(6):224–232, 2011.

[233] Mendell JT and Olson EN. MicroRNAs in stress signaling and human disease. *Cell*, 148(6):1172–1187, 2012.

[234] Keene JD. RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet*, 8(7):533–543, 2007.

[235] Pillai RS. MicroRNA function: multiple mechanisms for a tiny RNA? *RNA*, 11(12):1753–1761, 2005.

[236] Reczko M, Maragkakis M, Alexiou P, Grosse I, and Hatzigeorgiou AG. Functional microRNA targets in protein coding sequences. *Bioinformatics*, 28(6):771–776, 2012.

[237] Huang GT, Athanassiou C, and Benos PV. mirConnX: condition-specific mRNA-microRNA network integrator. *Nucleic Acids Res*, 39(Web Server issue):W416–W423, 2011.

[238] Sales G, Coppe A, Bisognin A, Biasiolo M, Bortoluzzi S, and Romualdi C. MAGIA, a web-based tool for miRNA and Genes Integrated Analysis. *Nucleic Acids Res*, 38(Web Server issue):W352–W359, 2010.

[239] Bisognin A, Sales G, Coppe A, Bortoluzzi S, and Romualdi C. MAGIA2: from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update). *Nucleic Acids Res*, 40(Web Server issue):W13–W21, 2012.

[240] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, and Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7, 2006.

[241] Weinstein JN. Spotlight on molecular profiling: "Integromic" analysis of the NCI-60 cancer cell lines. *Mol Cancer Ther*, 5(11):2601–2605, 2006.

[242] Muniategui A, Pey J, Planes FJ, and Rubio A. Joint analysis of miRNA and mRNA expression data. *Brief Bioinform*, 14(3):263–278, 2013.

[243] Liang Z, Zhou H, He Z, Zheng H, and Wu J. mirAct: a web tool for evaluating microRNA activity based on gene expression data. *Nucleic Acids Res*, 39(Web Server issue):W139–W144, 2011.

[244] Conover WJ. Practical Nonparametric Statistics. *Hoboken, NJ: John Wiley & Sons, Inc*, 1980.

[245] Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, Hughes TR, Blencowe BJ, Frey BJ, and Morris QD. Using expression profiling data to identify human microRNA targets. *Nat Methods*, 4(12):1045–1049, 2007.

[246] Muniategui A, Nogales-Cadenas R, Vázquez M, Aranguren XL, Agirre X, Luttun A, Prosper F, Pascual-Montano A, and Rubio A. Quantification of miRNA-mRNA interactions. *PLoS One*, 7(2):e30766, 2012.

[247] Zhou Y, Ferguson J, Chang JT, and Kluger Y. Inter- and intra-combinatorial regulation by transcription factors and microRNAs. *BMC Genomics*, 8:396, 2007.

[248] Tran DH, Satou K, Ho TB, and Pham TH. Computational discovery of miR-TF regulatory modules in human genome. *Bioinformation*, 4(8):371–377, 2010.

[249] Chen CY, Chen ST, Fuh CS, Juan HF, and Huang HC. Coregulation of transcription factors and microRNAs in human transcriptional regulatory network. *BMC Bioinformatics*, 12 Suppl 1:S41, 2011.

[250] Nam S, Li M, Choi K, Balch C, Kim S, and Nephew KP. MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Res*, 37(Web Server issue):W356–W362, 2009.

[251] Sokal R and Rohlf F. Biometry: the Principles and Practice of Statistics in Biological Research. *W H Freeman and Co, New York*, 3rd edn., 1995.

[252] Küffner R, Petri T, Tavakkolkhah P, Windhager L, and Zimmer R. Inferring gene regulatory networks by ANOVA. *Bioinformatics*, 28(10):1376–1382, 2012.

[253] Maglott D, Ostell J, Pruitt KD, and Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, 39(Database issue):D52–D57, 2011.

[254] Kozomara A and Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, 39(Database issue):D152–D157, 2011.

[255] Fujita S and Iba H. Putative promoter regions of miRNA genes involved in evolutionarily conserved regulatory systems among vertebrates. *Bioinformatics*, 24(3):303–308, 2008.

[256] Zhou X, Ruan J, Wang G, and Zhang W. Characterization and identification of microRNA core promoters in four model species. *PLoS Comput Biol*, 3(3):e37, 2007.

[257] Wang X, Xuan Z, Zhao X, Li Y, and Zhang MQ. High-resolution human core-promoter prediction with CoreBoost_HM. *Genome Res*, 19(2):266–275, 2009.

[258] Corcoran DL, Pandit KV, Gordon B, Bhattacharjee A, Kaminski N, and Benos PV. Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. *PLoS One*, 4(4):e5279, 2009.

[259] Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, Guenther MG, Johnston WK, Wernig M, Newman J, Calabrese JM, Dennis LM, Volkert TL, Gupta S, Love J, Hannett N, Sharp PA, Bartel DP, Jaenisch R, and Young RA. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, 134(3):521–533, 2008.

[260] Chien CH, Sun YM, Chang WC, Chiang-Hsieh PY, Lee TY, Tsai WC, Horng JT, Tsou AP, and Huang HD. Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data. *Nucleic Acids Res*, 39(21):9345–9356, 2011.

[261] Down TA and Hubbard TJP. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res*, 12(3):458–461, 2002.

[262] Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, and Werner T. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, 21(13):2933–2942, 2005.

[263] Klingenhoff A, Frech K, Quandt K, and Werner T. Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics*, 15(3):180–186, 1999.

[264] Gerner M, Sarafraz F, Bergman CM, and Nenadic G. BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. *Bioinformatics*, 28(16):2154–2161, 2012.

[265] Naeem H, Küffner R, Csaba G, and Zimmer R. miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature. *BMC Bioinformatics*, 11:135, 2010.

[266] Dweep H, Sticht C, Pandey P, and Gretz N. miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *J Biomed Inform*, 44(5):839–847, 2011.

[267] Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 79:829–836, 1979.

[268] Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, Wu L, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, and Snyder M. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, 2012.

[269] Jiang C, Xuan Z, Zhao F, and Zhang MQ. TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res*, 35(Database issue):D137–D140, 2007.

[270] Wingender E, Dietze P, Karas H, and Knüppel R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res*, 24(1):238–241, 1996.

[271] Wang J, Lu M, Qiu C, and Cui Q. TransmiR: a transcription factor-microRNA regulation database. *Nucleic Acids Res*, 38(Database issue):D119–D122, 2010.

[272] Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH, Chu CF, Huang HY, Lin CM, Ho SY, Jian TY, Lin FM, Chang TH, Weng SL, Liao KW, Liao IE, Liu CC, and Huang HD. miRTarBase update 2014: an information resource for exper-

imentally validated miRNA-target interactions. *Nucleic Acids Res*, 42(Database issue):D78–D85, 2014.

[273] Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, Gerangelos S, Koziris N, Dalamagas T, and Hatzigeorgiou AG. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res*, 40(Database issue):D222–D229, 2012.

[274] Kauffmann A, Gentleman R, and Huber W. arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–416, 2009.

[275] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, and Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[276] Hackstadt AJ and Hess AM. Filtering for increased power for microarray data analysis. *BMC Bioinformatics*, 10:11, 2009.

[277] Cohen J. Eta-Squared and Partial Eta-Squared in Fixed Factor Anova Designs. *Educational and psychological measurement*, 33:107–112, 1973.

[278] Miller RG. Beyond ANOVA: Basics of Applied Statistics. *Chapman & Hall/CRC*, 1997.

[279] Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57:289–300, 1995.

[280] Schaffter T, Marbach D, and Floreano D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011.

[281] Davis J and Goadrich M. The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[282] Barabási AL and Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, 2004.

[283] Guelzim N, Bottani S, Bourgine P, and Képès F. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet*, 31(1):60–63, 2002.

[284] Cheng C, Yan KK, Hwang W, Qian J, Bhardwaj N, Rozowsky J, Lu ZJ, Niu W, Alves P, Kato M, Snyder M, and Gerstein M. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput Biol*, 7(11):e1002190, 2011.

[285] Babu MM, Luscombe NM, Aravind L, Gerstein M, and Teichmann SA. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol*, 14(3):283–291, 2004.

[286] Frankel LB, Christoffersen NR, Jacobsen A, Lindow M, Krogh A, and Lund AH. Programmed cell death 4 (PDCD4) is an important functional target of the microRNA miR-21 in breast cancer cells. *J Biol Chem*, 283(2):1026–1033, 2008.

[287] Ma X, Choudhury SN, Hua X, Dai Z, and Li Y. Interaction of the oncogenic miR-21 microRNA and the p53 tumor suppressor pathway. *Carcinogenesis*, 34(6):1216–1223, 2013.

[288] Yu W, Gwinn M, Clyne M, Yesupriya A, and Khoury MJ. A navigator for human genome epidemiology. *Nat Genet*, 40(2):124–125, 2008.

[289] Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, and Stratton MR. A census of human cancer genes. *Nat Rev Cancer*, 4(3):177–183, 2004.

[290] Schriml LM, Arze C, Nadendla S, Chang YWW, Mazaitis M, Felix V, Feng G, and Kibbe WA. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res*, 40(Database issue):D940–D946, 2012.

[291] Hanahan D and Weinberg RA. The hallmarks of cancer. *Cell*, 100(1):57–70, 2000.

[292] Hanahan D and Weinberg RA. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[293] Plaisier CL, Pan M, and Baliga NS. A miRNA-regulatory network explains how dysregulated miRNAs perturb oncogenic processes across diverse cancers. *Genome Res*, 22(11):2302–2314, 2012.

[294] Dews M, Fox JL, Hultine S, Sundaram P, Wang W, Liu YY, Furth E, Enders GH, El-Deiry W, Schelter JM, Cleary MA, and Thomas-Tikhonenko A. The myc-miR-17 92 axis blunts TGFbeta signaling and production of multiple TGFbeta-dependent antiangiogenic factors. *Cancer Res*, 70(20):8233–8246, 2010.

[295] Roberts DD. Regulation of tumor growth and metastasis by thrombospondin-1. *FASEB J*, 10(10):1183–1191, 1996.

[296] Janz A, Sevignani C, Kenyon K, Ngo CV, and Thomas-Tikhonenko A. Activation of the myc oncoprotein leads to increased turnover of thrombospondin-1 mRNA. *Nucleic Acids Res*, 28(11):2268–2275, 2000.

[297] Blower PE, Chung JH, Verducci JS, Lin S, Park JK, Dai Z, Liu CG, Schmittgen TD, Reinhold WC, Croce CM, Weinstein JN, and Sadee W. MicroRNAs modulate the chemosensitivity of tumor cells. *Mol Cancer Ther*, 7(1):1–9, 2008.

[298] Song SJ, Ito K, Ala U, Kats L, Webster K, Sun SM, Jongen-Lavrencic M, Manova-Todorova K, Teruya-Feldstein J, Avigan DE, Delwel R, and Pandolfi PP. The oncogenic microRNA miR-22 targets the TET2 tumor suppressor to promote hematopoietic stem cell self-renewal and transformation. *Cell Stem Cell*, 13(1):87–101, 2013.

[299] Dang CV. MYC, metabolism, cell growth, and tumorigenesis. *Cold Spring Harb Perspect Med*, 3(8), 2013.

[300] Woynarowski JM, Chapman WG, Napier C, Herzig MC, and Juniewicz P. Sequence- and region-specificity of oxaliplatin adducts in naked and cellular DNA. *Mol Pharmacol*, 54(5):770–777, 1998.

[301] Dubik D and Shiu RP. Mechanism of estrogen activation of c-myc oncogene expression. *Oncogene*, 7(8):1587–1594, 1992.

[302] Santizo R and Pelligrino DA. Estrogen reduces leukocyte adhesion in the cerebral circulation of female rats. *J Cereb Blood Flow Metab*, 19(10):1061–1065, 1999.

[303] Park JK, Park SH, So K, Bae IH, Yoo YD, and Um HD. ICAM-3 enhances the migratory and invasive potential of human non-small cell lung cancer cells by inducing MMP-2 and MMP-9 via Akt and CREB. *Int J Oncol*, 36(1):181–192, 2010.

[304] Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, and Stolovitzky G. Revealing strengths and weaknesses of methods for gene network inference. *Proc Natl Acad Sci U S A*, 107(14):6286–6291, 2010.

[305] Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, and Yaschenko E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 36(Database issue):D13–D21, 2008.

[306] Chi SW, Hannon GJ, and Darnell RB. An alternative mode of microRNA target recognition. *Nat Struct Mol Biol*, 19(3):321–327, 2012.

[307] Muppirala UK, Honavar VG, and Dobbs D. Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics*, 12:489, 2011.

[308] Liu C, Mallick B, Long D, Rennie WA, Wolenc A, Carmack CS, and Ding Y. CLIP-based prediction of mammalian microRNA binding sites. *Nucleic Acids Res*, 41(14):e138, 2013.

[309] Rennie W, Liu C, Carmack CS, Wolenc A, Kanoria S, Lu J, Long D, and Ding Y. STarMir: a web server for prediction of microRNA binding sites. *Nucleic Acids Res*, 42(Web Server issue):W114–W118, 2014.

[310] Corcoran DL, Georgiev S, Mukherjee N, Gottwein E, Skalsky RL, Keene JD, and Ohler U. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol*, 12(8):R79, 2011.

[311] Erhard F, Dölken L, Jaskiewicz L, and Zimmer R. PARma: identification of microRNA target sites in AGO-PAR-CLIP data. *Genome Biol*, 14(7):R79, 2013.

[312] Chou CH, Lin FM, Chou MT, Hsu SD, Chang TH, Weng SL, Shrestha S, Hsiao CC, Hung JH, and Huang HD. A computational approach for identifying microRNA-target interactions using high-throughput CLIP and PAR-CLIP sequencing. *BMC Genomics*, 14 Suppl 1:S2, 2013.

[313] Yekta S, Shih IH, and Bartel DP. MicroRNA-directed cleavage of HOXB8 mRNA. *Science*, 304(5670):594–596, 2004.

[314] Szostak E and Gebauer F. Translational control by 3'-UTR-binding proteins. *Brief Funct Genomics*, 12(1):58–65, 2013.

[315] Skourti-Stathaki K and Proudfoot NJ. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes Dev*, 28(13):1384–1396, 2014.

[316] Laing C and Schlick T. Computational approaches to RNA structure prediction, analysis, and design. *Curr Opin Struct Biol*, 21(3):306–318, 2011.

[317] Kedde M, van Kouwenhove M, Zwart W, Oude Vrielink JAF, Elkon R, and Agami R. A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility. *Nat Cell Biol*, 12(10):1014–1020, 2010.

[318] McLean MA and Tirosh I. Opposite GC skews at the 5' and 3' ends of genes in unicellular fungi. *BMC Genomics*, 12:638, 2011.

[319] Montgomery RL, Yu G, Latimer PA, Stack C, Robinson K, Dalby CM, Kaminski N, and van Rooij E. MicroRNA mimicry blocks pulmonary fibrosis. *EMBO Mol Med*, 6(10):1347–1356, 2014.

[320] Shen LX, Basilion JP, and Stanton V Jr. Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc Natl Acad Sci U S A*, 96(14):7871–7876, 1999.

[321] Day L, Abdelhadi Ep Souki O, Albrecht AA, and Steinhöfel K. Accessibility of microRNA binding sites in metastable RNA secondary structures in the presence of SNPs. *Bioinformatics*, 30(3):343–352, 2014.

[322] Thomas LF, Saito T, and Sætrom P. Inferring causative variants in microRNA target sites. *Nucleic Acids Res*, 39(16):e109, 2011.

[323] Bruno AE, Li L, Kalabus JL, Pan Y, Yu A, and Hu Z. miRdSNP: a database of disease-associated SNPs and microRNA target sites on 3'UTRs of human genes. *BMC Genomics*, 13:44, 2012.

[324] Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[325] Hudson NJ, Dalrymple BP, and Reverter A. Beyond differential expression: the quest for causal mutations and effector molecules. *BMC Genomics*, 13:356, 2012.

[326] Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, and Pietenpol JA. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest*, 121(7):2750–2767, 2011.

[327] Bullinger L, Döhner K, Bair E, Fröhling S, Schlenk RF, Tibshirani R, Döhner H, and Pollack JR. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N Engl J Med*, 350(16):1605–1616, 2004.

[328] Durruthy-Durruthy R, Gottlieb A, Hartman BH, Waldhaus J, Laske RD, Altman R, and Heller S. Reconstruction of the mouse otocyst and early neuroblast lineage at single-cell resolution. *Cell*, 157(4):964–978, 2014.

[329] Liu Y, Gu Q, Hou JP, Han J, and Ma J. A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. *BMC Bioinformatics*, 15:37, 2014.

[330] Falkenauer E. *Genetic Algorithms and Grouping Problems*. John Wiley & Sons, Inc., 1998.

[331] Hruschka E, Campello R, AA F, and de Carvalho A. A Survey of Evolutionary Algorithms for Clustering. 39(2):133–155, 2009.

[332] Lozzio CB and Wigler PW. Cytotoxic effects of thiopyrimidines. *J Cell Physiol*, 78(1):25–32, 1971.

[333] Li JH, Liu S, Zhou H, Qu LH, and Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res*, 42(Database issue):D92–D97, 2014.

[334] Erhard F, Haas J, Lieber D, Malterer G, Jaskiewicz L, Zavolan M, Dölken L, and Zimmer R. Widespread context dependency of microRNA-mediated regulation. *Genome Res*, 24(6):906–919, 2014.

[335] Clark PM, Loher P, Quann K, Brody J, Londin ER, and Rigoutsos I. Argonaute CLIP-Seq reveals miRNA targetome diversity across tissue types. *Sci Rep*, 4:5947, 2014.

[336] Imig J, Brunschweiger A, Brümmer A, Guennewig B, Mittal N, Kishore S, Tsikrika P, Gerber AP, Zavolan M, and Hall J. miR-CLIP capture of a miRNA targetome uncovers a lincRNA H19-miR-106a interaction. *Nat Chem Biol*, 11(2):107–114, 2015.

[337] Grimson A. Noncoding RNA: Linking microRNAs to their targets. *Nat Chem Biol*, 11(2):100–101, 2015.

APPENDIX A

Supplemental text

In silico benchmark of inference methods

A set of known regulatory interactions (experimentally verified in any condition) in human composing a GRN with 14 768 nodes and 64 029 edges (the source network) was created. Depending on the node type, the nodes were labeled by either the mRNA Entrez Gene ID or the miRNA miRBase gene ID. Auto-regulatory loops were removed. Each interaction was randomly assigned a regulation sign. GeneNetWeaver (version 3.1)[280] was used to extract 80 modules of size 500 nodes (twice the size of the benchmark suite A proposed by Schaffter *et al.*[280]) from the source network as follows: the parameter *seed* was set to *random vertex* and *neighbor selection* was set to *random among top 50%*; networks holding less than 33% regulator genes were discarded to avoid structures with many genes not regulating any other gene. To obtain a balanced condition-specific gold standard, for each sub-network 50% of its edges were randomly chosen to occur in the simulated conditions (positive instances); the 50% of regulatory associations that take not place in the simulated conditions constitute the set of negative instances. Note that extracted sub-networks had identical numbers of nodes but the number of edges varied.

GeneNetWeaver was applied to endow each network contained in the gold standard with a detailed kinetic model considering both, independent and synergistic gene regulation. *Stochastic differential equations* (Langevin equations with *coefficient* $= 0.05$) were selected to model internal noise in the dynamics of the network and added experimental noise to the gene expression data sets by applying the *model of noise in microarrays* (similar to a

mix of normal and log-normal noise). Synthetic gene expression profiles of 60 conditions (c.f. NCI-60 cancer microarray project[241]) were produced by simulating steady-states of *multifactorial perturbations* (variation of the network steady state) to the original network. Replicates were generated by executing the stochastic simulation of the identical perturbations (condition) for 5 times, i.e. 5 samples per condition were obtained. Note that expression values of replicates of the same condition differ due to the intrinsic and experimental noise. For each network contained in the gold standard, all measurements were combined to a matched data set of mRNA and miRNA gene expression.

Finally, 80 sets of expression profiles for 60 conditions and their corresponding true condition-specific interaction graph (order = 500, size: $min = 852$, $median = 1\,226$, $max = 1\,421$) was obtained. Note that the *in silico* expression data is on linear scale, whereas actual microarray data is $\log_2$-scaled. This transformation step was passed as the dynamic model of GeneNetWeaver produces data that lends itself well to a linear scale and thus does not vary over several orders of magnitude like raw microarray data. Positive as well as negative signs for the regulation by TFs and miRNAs were considered as all three tools (COGERE, mirConnX, MAGIA2) predict repression and stimulation for any class of RG.

All condition-specific interactions were predicted with the stand-alone version of COGERE. All filters (sample distance, probe intensity, and probe variance) were switched off. The parameters for mirConnX[237] as well as MAGIA2[239] were chosen as proposed by the authors in the respective references. First, the simulated expression data was uploaded to mirConnX and the condition-specific regulatory network was computed as follows: *organism = human hg19 (GRCh37) 20111109*, *gene ID = Entrez Gene ID*, *microRNA ID = accession*, *association measure* = *Pearson*, *prior weight* = 0.3, *integration function = weighted sum*.The regulation threshold was set to *0* to reach the maximum possible sensitivity. As mirConnX labels the nodes by the internal symbol names (stored in its database), the symbols were translated to the corresponding official Entrez Gene or miRBase ID. At this, synonyms were considered and ambiguous as well as not assignable cases were resolved manually (e.g. systematic gene name errors such as 1-SEP, 1-MAR or NaOG, MT-CO2). Second, all expression matrices of the gold standard were uploaded to the MAGIA2 web-server. The *ID Type* was set to *Entrez Gene* for the mRNA expression data; for miRNA MAGIA2 allows only transcript symbols. The variability filter was skipped and *Pearson Correlation* was selected as inference method and the intersection of TargetScan and DIANA-microT was chosen as prior.

APPENDIX B

Supplemental figures

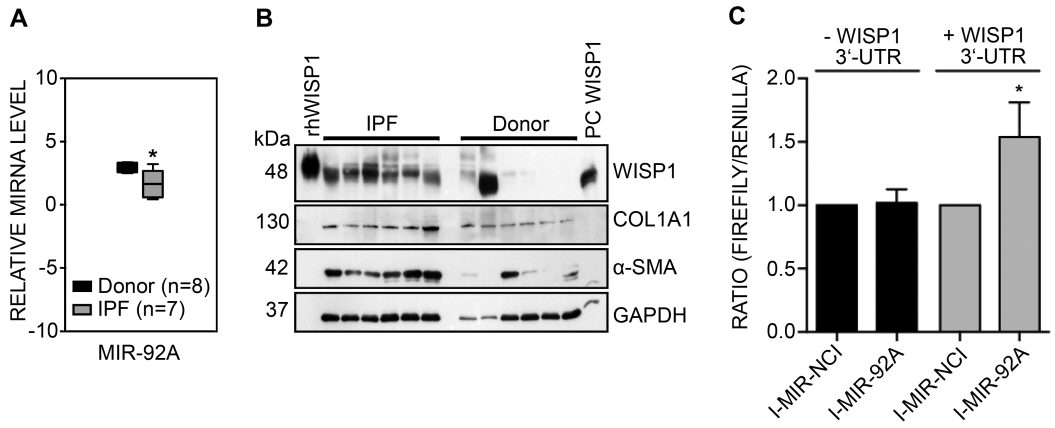


Figure B.1 | **Experimental confirmation of miR-92a regulation of WISP1. A |** Expression of miR-92a in donor (from unaffected lung tissue biopsies) and human IPF lung homogenate relative to RUN43. miR-92a is significantly downregulated in IPF (Wilcoxon rank sum test $P < 0.05$). **B |** Western blot analysis of WISP1 and pro-fibrotic markers COL1A1 and α-SMA in donor and human IPF lung homogenate. GAPDH was used as loading control. Recombinant human WISP1 protein (rhWISP1) and A549 cell lysates overexpressing WISP1 (PC WISP1) were used as loading controls. It can be seen that WISP1 is upregulated on protein level in IPF. **C |** Luciferase reporter assays. Black bars illustrate normalized ratios to reporter construct without WISP1 3'-UTR, miR-92a inhibitor (I-MIR-92A) and the negative control (I-MiR-NCI); grey bars illustrate normalized ratios to reporter construct with WISP1 3'-UTR. The activity of the reporter gene with the 3'-UTR was significantly increased following miR-92a inhibition. Experiments were performed by and figures were adapted from Berschneider *et al.*, 2014[80] with permission of Elsevier (license number 3573240996630); * $P < 0.05$.

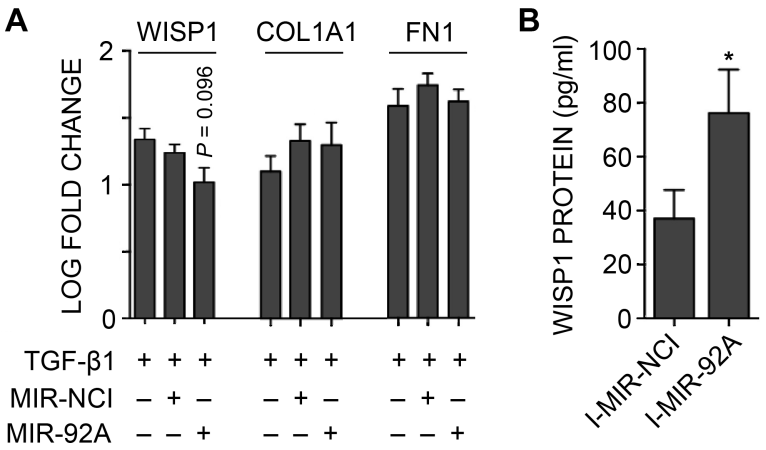


Figure B.2 | **miR-92a affects TGF-β1-induced WISP1 expression. A |** Human pFB were treated with TGF-β1 and transfected miR-92a mimics. Normalized log fold-changes of WISP1, COL1A1 and FN1 were computed. The housekeeper HPRT and untreated, non-transfected control cells after 24 h served as reference for normalization. Transfection with MiR-NCI denotes the negative control. Transfection with the miRNA mimic lowers the WISP1 level. **B |** WISP1 enzyme-linked immunosorbent assay from pFB supernatants transfected with a miR-92a inhibitor. Suppression of miR-92a significantly increased the WISP1 concentration. Experiments were performed by and figures were adapted from Berschneider *et al.*, 2014[80] with permission of Elsevier (license number 3573240996630); * $P < 0.05$.

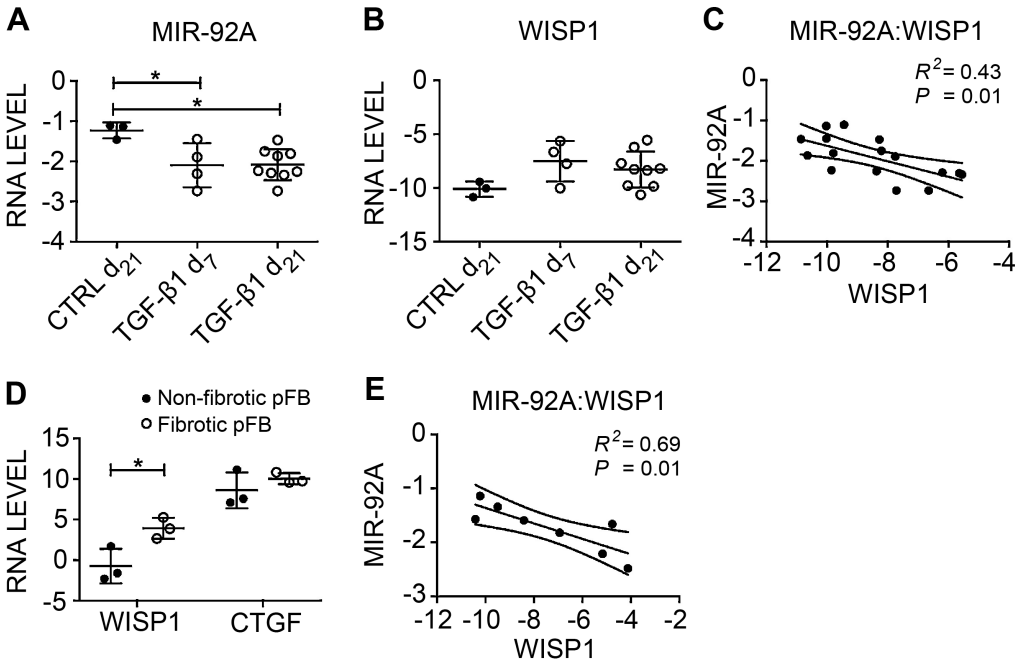


Figure B.3 | **Correlation of miR-92a and WISP1 levels *in vivo* and *ex vivo*.** RNA levels of WISP1 (relative to GAPDH, **A**) and miR-92a (relative to RUN6, **B**) in rat whole lung homogenates at day 7 and day 14 after infection with TGF-β1 bearing adenoviruses. WISP1 levels are increasing while miR-92a concentration declines *in vivo*. Further, WISP1 was also found highly expressed in non-fibrotic and fibrotic human pFB from IPF patients (RNA level relative to HPRT, **D**). Regression analysis reveals a strong negative correlation between WISP1 and miR-92a in whole rat lung homogenates (**C**) as well as in human non-fibrotic and fibrotic pFB (**E**). Lines above and below the linear regression lines denote the 95% confidence interval. RNA levels were measured by RT-qPCR. Experiments were performed by and figures were adapted from Berschneider *et al.*, 2014[80] with permission of Elsevier (license number 3573240996630); * $P < 0.05$.

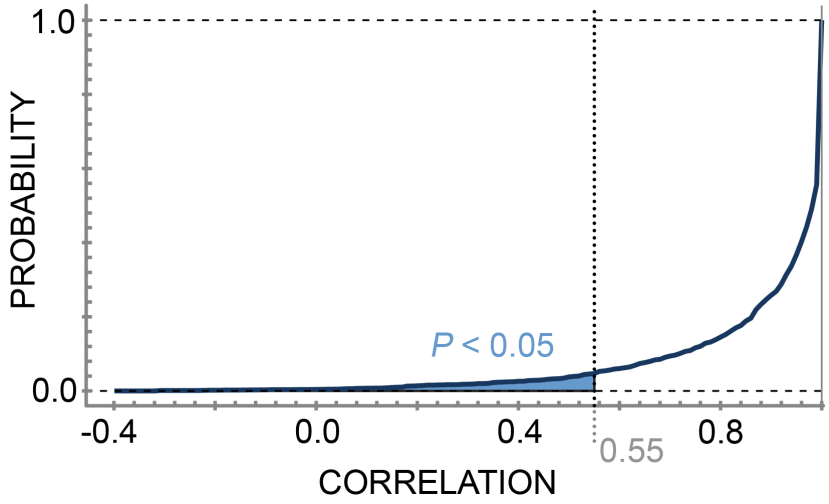


Figure B.4 | **Probability distribution of correlation coefficients.** Shown are the linear relationships between the reference and mutated structures of miRNP binding regions. The lower the coefficient the severer the structural shift. Below a coefficient of 0.55 the probability to observe a chance of RNA secondary structure of this extent by change is less than 5%.

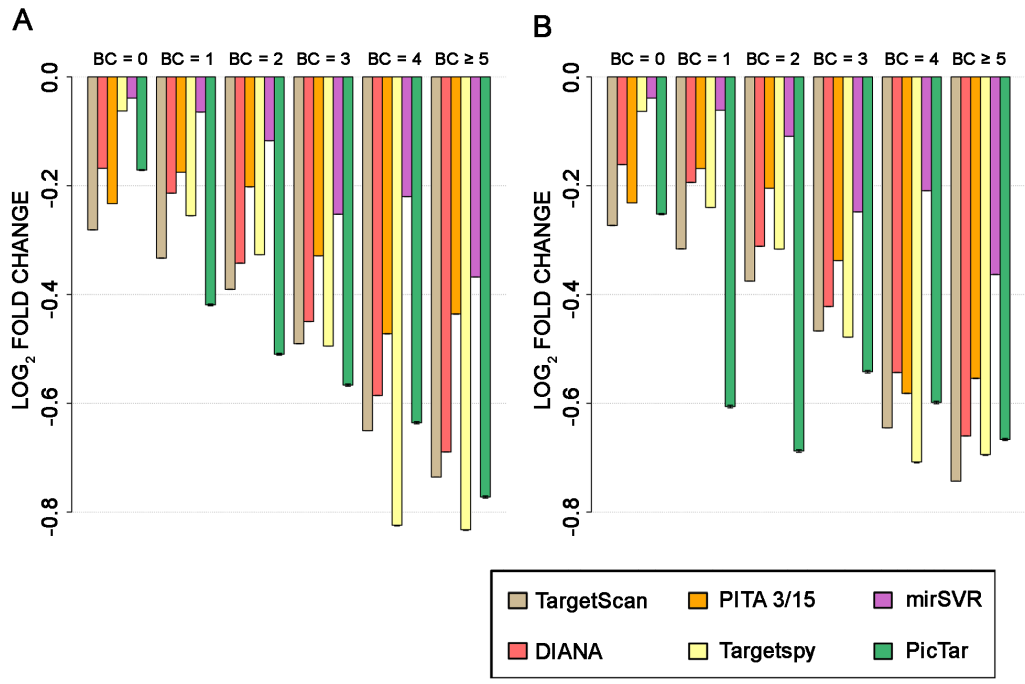


Figure B.5 | **Contribution of individual prediction algorithms to the *prior* score.** Based on pulsed SILAC data[91], individual scores of six miRNA target prediction algorithms were transformed to a unified score weighting the regulatory potential of a miRNA:TG interaction, i.e. the expected $\log_2$ expression fold-change of the TG in human (**A**) and mouse (**B**). Additional CLIP-Seq data was utilized to identify predicted targets located in a known AGO2 binding region. BC denotes the biological complexity of the AGO2 binding region, i.e. a measure of reproducibility between biological replicates or experiments; BC $= 0$ denotes target sites not located in any known AGO2 binding region. Shown is the average for each transformed score distribution for each biological complexity. The error bars denote the 95% confidence interval for the mean. COGERE scores each miRNA:TG interaction by the sum of the maximum predicted fold-change of each tool. Thus, the vertical bars denote the average contribution of each tool to the final score, e.g. if a human miRNA target site of TargetSpy is located in an AGO2 region with BC $= 5$, then TargetSpy will contribute the highest fraction to the final miRNA:TG score.

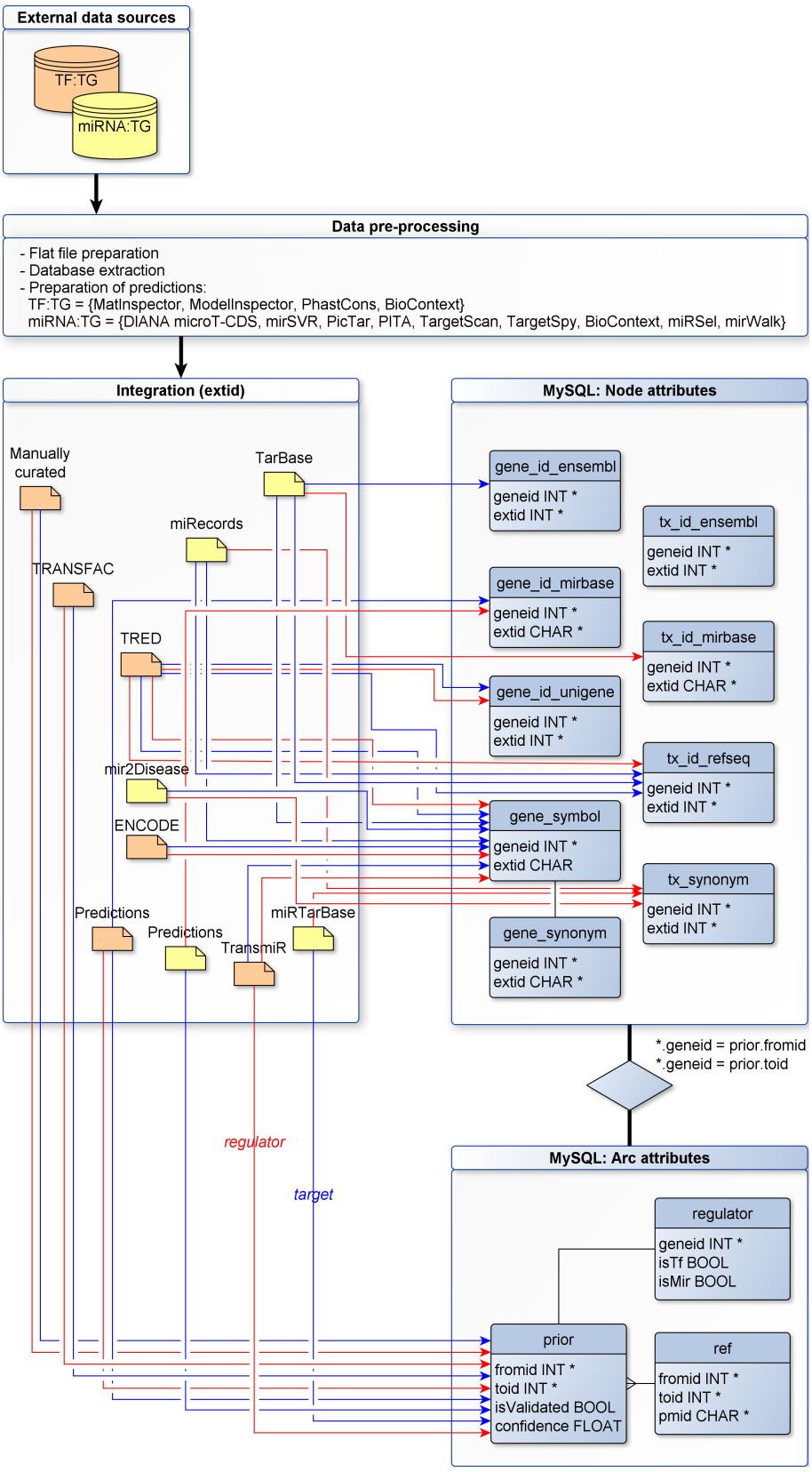


Figure B.6 | **Integration framework and database scheme.** *(continued on next page)*

Figure B.6 *(previous page)* | TF:TG (orange) and miRNA:TG (yellow) regulatory interactions were extracted from several heterogenous data sources. All validated data was manually pre-processed to remove flat file inconsistencies, and multiple target predictions were integrated to a single unified set of interactions. Since all resources use varying gene and transcript (tx) accession numbers, all external IDs (extid) of a regulator (red arcs) and a target (blue arcs) were mapped to its unique Entrez Gene ID (geneid, fromid, toid). All used MySQL tables are illustrated as blue boxes. The interactions with their confidence, i.e. the prior score, are stored as directed edges (arcs) in the prior table. In addition, the type of a regulator (miRNA or TF) and the references (table ref) are stored.

APPENDIX C

Supplemental tables

Table C.1 | **SNPs predicted to disrupt/dampen existing MREs.** The first column lists the SNP rs-numbers, in the second column the respective transcripts are given, the fourth column indicates the conservation of the locus in mammals, and the fifth column lists the miRNAs for which MREs were disrupted/dampened.

SNP	RefSeq	Conservation	miRNA
rs4564	NM_000108	✓	miR-323-3p
rs6706	NM_003302		miR-509-3p; miR-219-5p; miR-508-3p
rs7089	NM_024107; NM_177441; NM_001076674		miR-545
rs7118	NM_133458		miR-512-3p; miR-218; miR-455-5p
rs7119	NM_018200		miR-571
rs8523	NM_017770	✓	miR-583; miR-1276; miR-203; miR-539
rs9253	NM_022756	✓	miR-545
rs10923	NM_005496; NM_001002800		miR-299-5p
rs12439	NM_013943		miR-338-5p
rs699779	NM_024408		miR-381; miR-300; miR-1284
rs835575	NM_024408		miR-559; miR-106b; miR-20a; miR-340; miR-142-5p
rs835576	NM_024408		miR-218

Table C.1 *(continued)*

SNP	RefSeq	Conservation	miRNA
rs1045100	NM_001190266; NM_001190267; NM_030803; NM_017974; NM_198890		miR-190b; miR-190
rs1047440	NM_153223; NM_001166226		miR-503
rs1058588	NM_003761		miR-573
rs1379659	NM_004787	✓	miR-602
rs2229302	NM_002145		miR-886-5p
rs3816661	NM_025240; NM_001024736		miR-1278; miR-152; miR-148a; miR-148b; miR-152; miR-148a; miR-148b
rs4770433	NM_014363	✓	miR-361-5p
rs7350928	NM_001193466; NM_015443; NM_001193465		miR-483-5p; miR-184
rs11713355	NM_001134367; NM_003043		miR-487b
rs17574361	NM_001193466; NM_015443; NM_001193465	✓	miR-583; miR-1276; miR-203; miR-488

Table C.2 | **SNPs predicted to create/enhance MREs.** The first column lists the SNP rs-numbers, in the second column the respective transcripts are given, the fourth column indicates the conservation of the locus in mammals, and the fifth column lists the miRNAs for which MREs are created/enhanced.

SNP	RefSeq	Conservation	miRNA
rs1121	NM_015027		hsa-miR-892b; hsa-miR-647; miR-149; miR-1254; miR-550; miR-515-3p; miR-661
rs7089	NM_024107; NM_177441; NM_001076674		miR-640
rs7097	NM_015972		miR-335
rs7118	NM_133458		miR-556-5p
rs8523	NM_017770	✓	miR-548c-3p
rs9927	NM_002862		miR-634; miR-1226
rs10923	NM_005496; NM_001002800		miR-636
rs11700	NM_001950	✓	miR-328; miR-1291
rs12439	NM_013943		miR-421; miR-1324
rs12916	NM_000859; NM_001130996		miR-1909; miR-1262; miR-342-5p; miR-608; miR-1207-5p
rs42039	NM_001145306; NM_001259		miR-509-5p; miR-509-3-5p; miR-330-5p; miR-942; miR-544; miR-1205; miR-593
rs232775	NM_001085487		miR-1909; miR-1266; miR-342-5p; miR-608
rs823136	NM_003929; NM_001135664; NM_001135663; NM_001135662		miR-384
rs835575	NM_024408		miR-513a-3p; miR-587
rs835576	NM_024408		miR-1208; miR-210; miR-141; miR-200a; miR-1914; miR-892a
rs1045100	NM_001190266; NM_001190267; NM_030803; NM_017974; NM_198890		miR-597
rs1045407	NM_178549		miR-508-5p; miR-766; miR-490-5p; miR-136
rs1047440	NM_153223; NM_001166226		miR-34c-5p; miR-449b; miR-34a; miR-449a
rs2032933	NM_152308		miR-215

Table C.2 *(continued)*

SNP	RefSeq	Conservation	miRNA
rs2071518	NM_002514		miR-649
rs2229302	NM_002145		miR-542-5p; miR-769-3p
rs2293578	NM_152264; NM_001128225		miR-575; miR-7; miR-335
rs4819388	NM_015259		miR-1915
rs7350928	NM_001193466; NM_015443; NM_001193465		miR-1308; miR-1262
rs8176751	NM_020469		miR-1287; miR-370; miR-34a; miR-449a; miR-1207-5p
rs11067231	NM_052845		miR-624
rs11713355	NM_001134367; NM_003043		miR-1267
rs17574361	NM_001193466; NM_015443; NM_001193465	✓	miR-185

Table C.3 | **SNPs predicted to affect 3'-UTR splicing.** The first column lists the SNP rs-numbers, in the second column the respective transcripts are given, the third column indicates the conservation of the locus in mammals, the fourth column denotes the distance to a reference splice site (negative: 5' upstream of exon junction site; positive: 3' downstream of exon border of mRNA), the sixth column contains the type of the gained splice site (acceptor gain: Acc+, donor gain: Don+), the score column contains the likelihood of NNSplice, and the last column provides the fraction of lost miRNP binding sites.

SNP	RefSeq	Conservation	Effect	Distance	Score	Loss
rs7371	NM_006496	✓	Acc+	341	0.53	0.18
rs42038	NM_001145306; NM_001259		Acc+	-881	0.85	0.29
rs699779	NM_024408		Acc+	-4013	0.51	1.00
rs2244967	NM_001031746		Acc+	-3054	0.90	0.50
rs4973768	NM_003615	✓	Don+	1799	0.98	0.24
rs6722332	NM_018256		Acc+	-334	0.98	1.00
rs7528419	NM_001408		Acc+	550	0.76	0.17

Table C.4 | **SNPs predicted to affect 3'-UTR secondary structure.** The first column lists the SNP rs-numbers, in the second column the respective transcripts are given, the third column indicates the conservation of the locus in mammals, and the fourth column lists the correlation between the reference and the mutated structure of the respective transcripts.

SNP	**RefSeq**	**Conservation**	ρ
rs7444	NM_003347		0.49
rs12956	NM_012234	✓	0.20
rs13099	NM_006827		0.29
rs42038	NM_001145306		0.46
rs835575	NM_024408		0.30
rs1045407	NM_178549		0.04
rs1046917	NM_024619		0.54
rs2077579	NM_004397	✓	0.53
rs2282301	NM_006912	✓	-0.10
rs2564921	NM_052859		0.19
rs3821301	NM_001145909		0.44
rs4819388	NM_015259		0.05
rs10892082	NM_002572	✓	-0.11
rs11542478	NM_001077710		0.46

Table C.5 | **Features of miRNA target prediction algorithms.** List of tools combined into a unified scoring-framework with each feature set explicitly considered by the respective algorithm: finding target sites in the coding-sequence (CDS) and 3'-UTR, requiring a Watson-Crick pairing between the miRNA seed sequence and the target sequence (Seed), computing thermodynamic features such as free energy of the miRNA:mRNA duplex (Energy), calculation of context features such as the local AU content of the target site (Context), conservation of the target site (Cons.), usage of expression data to weight feature scores based on the target fold-change (FC) and mining for miRNA:target pairs in biomedical text (Text). The last column contains the selected parameters for each tool to predict targets.

Algorithm	**CDS**	**Seed**	**Energy**	**Context**	**Cons.**	**FC**	**Text**	**Parameters**
DIANA [236]	✓	✓	✓	✓	✓			Score > 0.6
miRSVR [142]		✓	✓	✓	✓	✓		Score < -0.1, Cons. > 0.566
PicTar [102]		✓		✓	✓			Only conserved
PITA 3/15 [49]		✓	✓		✓			Seed $>$ 6mer, No gaps in seed, Cons. > 0.9
TargetScan 6.1 [50]		✓		✓	✓	✓		Only conserved
TargetSpy [104]			✓	✓				No seed match, high specificity
miRSel [265]							✓	Only human/mouse
miRWalk [266]							✓	Only human/mouse
Biocontext [264]							✓	Only human/mouse

Table C.6 | **Disease term mapping.** Shown are the NCI-60 cell lines with their Disease Ontology ID (DOID) and their corresponding PhenomiR and mir2Disease terms.

NCI-60 cancer	PhenomiR term	Mir2Disease term
DOID:10283, Prostate cancer	Prostate cancer	Prostate cancer
DOID:1324, Lung cancer	Lung cancer	Lung cancer; Non-small cell lung cancer
DOID:1612, Breast cancer	Breast cancer	Breast cancer
DOID:1909, Melanoma	Melanoma, cutaneous malignant, 2; Melanoma and neural system tumor syndrome	Malignant melanoma; Melanoma
DOID:2394, Ovarian cancer	Ovarian cancer	Epithelial ovarian cancer; Ovarian cancer; Recurrent ovarian cancer; Serous ovarian cancer
DOID:2531, Hematologic cancer	Leukemia, acute myeloid; Leukemia, chronic lymphatic, susceptibility to; Leukemia, chronic myeloid; Multiple myeloma; Non-Hodgkin lymphoma, somatic	Acute myeloid leukemia; Acute promyelocytic leukemia; Chronic lymphocytic leukemia; Chronic myeloid leukemia; Multiple myeloma; Myeloproliferative disorder; Follicular lymphoma; Acute lymphoblastic leukemia; T-cell leukemia
DOID:263, Kidney cancer	Renal cell carcinoma	Kidney cancer; Renal clear cell carcinoma
DOID:3070, Malignant glioma	Glioblastoma multiforme, somatic	Glioblastoma; Glioblastoma multiforme
DOID:9256, Colorectal cancer	Colorectal cancer; Adenomas, multiple colorectal	Colorectal cancer

APPENDIX D

Teaching activities

Lectures

During my PhD course, I was involved in the preparation and execution of the following lectures.

2014 Course 'Introduction to bioinformatics I: Exercises',
Technical University Munich

2014 Course 'Introduction to bioinformatics II: Exercises',
Technical University Munich

2014 Practical course 'Disease-oriented Bioinformatics',
Technical University Munich

2014 Practical course 'Genome-oriented bioinformatics',
Technical University Munich

2014 Lecture 'Non-coding and regulatory RNAs',
Course 'Advanced Bioinformatics',
Technical University Munich

2013 Course 'Introduction to bioinformatics I: Exercises',
Technical University Munich

2013 Course 'Introduction to bioinformatics II: Exercises',
Technical University Munich

2013 Practical course 'Applied Bioinformatics',
Technical University Munich

2013 Practical course 'Disease-oriented Bioinformatics',
Technical University Munich

2013 Practical course 'Genome-oriented bioinformatics',
Technical University Munich

2013 Lecture 'Non-coding and regulatory RNAs',
Course 'Advanced Bioinformatics',
Technical University Munich

2012 Course 'Introduction to bioinformatics I: Exercises',
Technical University Munich

2012 Course 'Introduction to bioinformatics II: Exercises',
Technical University Munich

2012 Practical course 'Applied Bioinformatics',
Technical University Munich

2012 Practical course 'Disease-oriented Bioinformatics',
Technical University Munich

2012 Practical course 'Genome-oriented bioinformatics',
Technical University Munich

2012 Lecture 'MicroRNAs: Small actors with a big role in the play of gene regulation.',
12th Bioinformatics Spring School of the Helmholtz Zentrum München,
Hainburg, Austria

2012 Lecture 'Phylogenetics',
Course 'Introduction to Bioinformatics II',
Technical University Munich

2012 Lecture 'Non-coding and regulatory RNAs',
Course 'Advanced Bioinformatics',
Technical University Munich

2011 Lecture 'MicroRNAs in systems biology',
Course 'Systems Biology of Diseases and Drug Treatment',
Technical University Munich

2011 Lecture 'Hidden Markov Models',
Course 'Introduction to Bioinformatics I',
Technical University Munich

2010 Course 'Introduction to bioinformatics I: Exercises',
Technical University Munich

2010 Course 'Introduction to bioinformatics II: Exercises',
Technical University Munich

2009 Lecture 'Biological networks',
Practical course 'Genome-oriented bioinformatics',
Technical University Munich

Theses

Further, I supervised the following theses.

2011 'Simulation of miRNA-mediated gene regulatory systems via Kauffman networks with memory', Master thesis in bioinformatics, Göksel Kaya.

2011 'Modeling of a diet-induced non-alcoholic fatty liver disease system from mRNA and miRNA expression profiles', Bachelor thesis in bioinformatics, Alice Meier.