HELMHOLTZ
ASSOCIATION
Alliance on Systems Biology

HelmholtzZentrum münchen
German Research Center for Environmental Health

TECHNISCHE
UNIVERSITÄT
MÜNCHEN

Analysis of microRNA function using systemic regulatory features and graph models

Martin Preusse, M.Sc.

March 2016

TECHNISCHE UNIVERSITÄT MÜNCHEN

Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Lehrstuhl M12 (Mathematische Modellierung biologischer Systeme)

Analysis of microRNA function using systemic regulatory features and graph models

Martin Preusse

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 (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende:

Univ.-Prof. Dr. I. Antes

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Dr. F. J. Theis
2. Univ.-Prof. Dr. H.-W. Mewes

Die Dissertation wurde am 29.03.2016 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 05.08.2016 angenommen.

Danksagung

Ich danke vielen Menschen, ohne die ich meine Doktorarbeit bestimmt nicht geschafft hätte. Allen voran natürlich Nikola, die mich nach unerwartetem Auftauchen in die ReNe-Familie aufgenommen hat. Mit viel Geduld hat sie mich danach unterstützt und das ganze Unterfangen in geordnete Bahnen gelenkt. Dabei haben wir noch den ein oder anderen Pokal eingesammelt, das sagt dann wohl alles.

Mein Anfang beim CMB liegt jetzt schon 7 Jahre zurück und darum danke ich Fabian, dass ich mit seiner andauernden Unterstützung als ahnungsloser Biotechnologe mein Plätzchen in der Bioinformatik finden konnte. Ohne Carsten wäre ich vermutlich nicht so lange geblieben und ohne die vielen Diskussionen mit ihm wüsste ich nicht viel von microRNAs und Pathways.

Ich danke Hans-Werner Mewes für die Zweitbetreuung meiner Doktorarbeit. Durch seine Vorlesungen hatte ich meinen ersten Kontakt zur Bioinformatik und auch meine Masterarbeit hat er schon betreut. Iris Antes danke ich für den Vorsitz in der Prüfungskommission.

Am Anfang meiner Doktorarbeit wurde ich von Heiko und dem IDR liebevoll aufgenommen und durfte mich nochmal so richtig an der Pipette austoben. Einen Versuch war es wert. Nach meiner Rückkehr an die Tastatur haben mich die wundervollen Menschen am ICB, und natürlich besonders im Team ReNe, immer wieder aufgebaut und durch inspirierende Diskussionen bereichert.

Ohne meine Eltern wäre ich, offensichtlich, nicht hier. Diese Geduld, dass die Kinder einen gefühlt dreißigjährigen Ausbildungsweg absolvieren, muss man erstmal mitbringen. Danke!

Abstract

MiRNAs are small, ca. 22 nucleotide long endogenous RNAs which regulate gene expression. The landscape of post-transcriptional regulation, and gene expression in general, changed dramatically with their introduction in the early 1990s. Today, it is widely accepted that miRNA-mediated gene regulation influences most mammalian genes and almost all biological processes. In many cases, however, the actual function of miRNA-mediated regulation *in vivo* is not clear.

The network of miRNA-mediated regulation is highly complex and both the quantitative aspects of miRNA-mRNA interactions as well as the cell specific effect size are poorly understood. Several long-standing paradigms of miRNA binding have recently been questioned by new experimental technologies using next-generation sequencing of miRNA-target complexes.

In this thesis, we seek to identify the biological function of miRNAs. In order to deal with the uncertainties in miRNA targeting data and capture the inherent complexity of miRNA regulation, we include three systemic features of miRNA regulation into functional analyses. Firstly, we show that distance-dependent cooperativity of miRNAs is a predictor for their functional impact. Secondly, we demonstrate that miRNA regulation of biological pathways is tissue specific. Thirdly, we identify novel regulatory mechanisms involving co-regulation by miRNAs and miRNA-independent RNA-binding proteins.

The regulatory features are incorporated into three web applications which are developed to support experimental miRNA research: *miRco*, *miTALOS v2* and *simiRa*. They allow to identify candidates with a high biological relevance from lists of potentially interesting miRNAs. Researchers working with miRNAs can thereby generate new hypotheses for functional regulation involving miRNAs.

The functional miRNA analyses presented in this thesis requires integration of a wide range of public data sources. We develop a unified graph data model of the cell which allows integration of molecular data and functional annotations. The noSQL graph database neo4j is used for a reference implementation to demonstrate the advantages of our data model in terms of flexibility and query structure.

In summary, we use systemic features of miRNA-mediated gene regulation to improve functional analyses and share our methods as web applications. A novel graph data model is used to integrate data throughout the individual analyses.

Zusammenfassung

MiRNAs sind kurze, ca. 22 Nukleotide lange RNAs, welche die Genexpression regulieren. Unser Bild der post-transkriptionalen Genregulation wurde revolutioniert, als die lange unbeachteten miRNAs in den 1990ern entdeckt wurden. Heute wird allgemein anerkannt, dass fast alle Gene in vielzelligen Tieren von miRNAs reguliert werden und miRNAs somit an allen wesentlichen biologischen Prozessen beteiligt sind. Ihre tatsächliche Funktion *in vivo* ist jedoch häufig unbekannt.

Das regulatorische Netzwerk aus miRNAs und Zielgenen ist sehr komplex und sowohl die Parameter der miRNA-mRNA-Bindung als auch der quantitative Einfluss von miRNAs auf die Genregulation sind umstritten. Einige Paradigmen der miRNA-basierten Regulation sind durch neue experimentelle Methoden basierend auf next-generation sequencing von miRNA-mRNA-Komplexen in Frage gestellt worden.

In dieser Arbeit analysieren wir die biologische Funktion von miRNAs. Um mit der Komplexität von miRNA-Regulation und der Unsicherheit bezüglich miRNA Zielgenen umzugehen, nutzen wir systemische Merkmale der miRNA-basierten Regulation für funktionale Analysen. Zuerst zeigen wir, dass distanzabhängige Kooperativität von miRNAs genutzt werden kann, um ihren funktionalen Einfluss vorherzusagen. Zweitens legen wir dar, dass die Regulation von Pathways durch miRNAs gewebespezifisch ist. Drittens identifizieren wir neue regulatorische Zusammenhänge basierend auf Coregulation durch miRNAs und RNA-Bindeproteine.

Diese regulatorischen Merkmale werden in drei Webanwendungen eingebaut, die experimentelle miRNA-Forschung unterstützen sollen: *miRco*, *miTALOS v2* und *simiRa*. Sie identifizieren Kandidaten für experimentelle Tests aus Listen von potentiell interessanten miRNAs und generieren so neue Hypothesen für regulatorische Zusammenhänge.

Die funktionalen Analysen, die hier vorgestellt werden, nutzen Daten aus vielen verschiedenen öffentlichen Quellen. Wir entwickeln ein allgemeines Graphmodell einer Zelle, mit dem wir molekulare Daten mit Annotationen integrieren. Eine Referenzimplementierung mit der Graphdatenbank neo4j wird genutzt, um die Vorteile des Datenmodells bezüglich Flexibilität und strukturierten Abfragen zu zeigen.

Zusammengefasst nutzen wir in dieser Arbeit systemische Merkmale von miRNAs um ihre funktionale Analyse zu verbessern und entwickeln Webanwendungen aufbauend auf unseren Methoden. Ein neuartiges Graphmodell wird zur Datenintegration über die einzelnen Analysen hinweg genutzt.

Contents

Chapter 1

Introduction

The development and homeostasis of living organisms depends on fine-grained regulation of the complex interactions that occur in its basic building block, the cell. After discovery of cellular structures and the principles of heredity, a fundamental question arose: How is information stored, inherited from parents to progeny and translated into the complex molecular interactions underlying all biological processes.

In the early 20th century, geneticists defined a 'gene' as a unit of inheritance and located them on the chromosomes. Without knowledge of the molecular basis, Thomas Hunt Morgan generated genetic maps of the fruitfly Drosophila melanogaster by crossing wild-type flies with mutants and observing ratios of inheritance. Proteins as distinct biological macromolecules have been known from the beginning of the 19th century. It took until the 1920ies, however, until their pivotal role in biological processes became evident. James B. Sumner demonstrated in 1926 that urease, the enzyme that catalyzes hydrolysis of urea, is a protein [1].

From that point, the molecular nature of genes and inheritance was established. It was unclear, however, how the genotype and the molecular phenotype are related. In 1941, Beadle and Tatum proposed the "one gene one enzyme" hypothesis, stating that a single gene gives rise to a single enzyme which catalyzes a single step in a metabolic process [2]. Consequently, they assumed a linear relationship between genotype and phenotype. This oversimplified view has been extended from the 1950ies, when Watson, Crick and others cracked the "genetic code" and described the "central dogma" of molecular biology: Genes are encoded in the DNA, transcribed into messenger RNA and then translated into proteins [3, 4, 5]. In this process, referred to as gene expression, sequence information is not transferred backwards.

The principles of their discoveries still hold true today. But while there is a flow of information from gene via transcript to protein, there is no linear relationship between transcriptional activity and protein production. On a cellular level, the correlation of transcript abundance to corresponding protein levels is limited [6, 7]. Moreover, the level of correlation varies between different cell types and cell type specific protein levels cannot be explained by changes in gene expression only [7]. A plethora of molecular mechanisms govern the regulation of gene expression and their complex interdependencies are still not fully understood. It is clear, however, that correct regulation of gene expression is essential and abnormalities in this fine-tuned process lead to diseases.

RNA was thought to merely act as a template for proteins. During the last decades, however, a multitude of RNAs with functions beyond encoding of proteins were discovered. The first non-coding RNA was described in 1965 when the structure of transfer RNA was elucidated [8]. This was followed by ribosomal RNAs and RNAs with enzymatic activity in splicing [9, 10]. In the 1990ies, the world of non-coding RNAs was revolutionized when previously overlooked small RNAs were shown to participate in post-transcriptional silencing of genes [11, 12]. Among the small non-coding RNAs, microRNAs (miRNAs) stand out as universal regulators of gene expression. They emerged as a key component in cellular regulation and were shown to play a role in almost all biological processes and pathogenesis of many diseases.

Many functional details of miRNA mediated regulation remain poorly understood. While regulation of individual genes has been demonstrated in numerous studies, their system-level effect *in vivo* is often unclear [13]. The main goal of this thesis is to analyze the biological function of miRNAs. Systemic features of miRNA mediated regulation are included in the analysis to capture the complex role of miRNAs in regulation of biological processes.

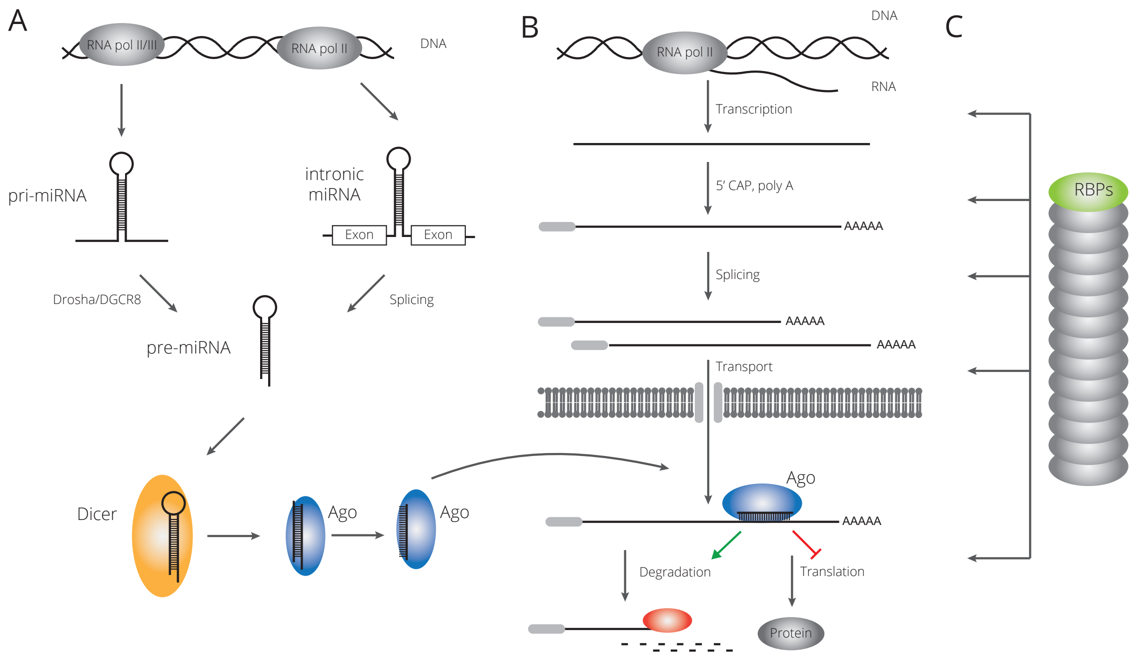


Figure 1.1: Principle of miRNA and RBP mediated regulation of gene expression. A) Pre-miRNAs are either transcribed from individual loci as pri-miRNAs and subsequently cleaved or spliced from host genes as intronic miRNAs. Dicer produces 22 nt long double stranded miRNAs. One of the two strands is then embedded in the AGO protein to form the miRISC while the other is discarded. B) MiRNAs regulate gene expression by binding to mRNAs and either repress translation or promote degradation of the target. C) MiRNA-independent RBPs are involved in regulation of gene expression on all levels of the mRNA life-cycle.

1.1 MicroRNA Mediated Gene Regulation

MiRNAs are small, ca. 22 nucleotide long endogenous RNAs which regulate gene expression [14, 15]. The landscape of post-transcriptional regulation, and gene expression in general, changed dramatically with their introduction in the early 1990s. The first discovery was lin-4, a miRNA which participates in regulation of developmental timing in the nematode *C. elegans* [16, 17]. In 2000, let-7 was found as the second miRNA regulating the same process [18]. Let-7 was also shown to be evolutionary conserved from *C. elegans* to *Drosophila*, zebrafish and human. With this finding, the general relevance of RNA based regulation of gene expression became evident.

MiRNAs have been predominantly described to silence gene expression by repressing translation or promoting degradation of mRNAs [19]. They execute their function by binding to target mRNAs as part of ribonucleoprotein complexes usually referred to as miRNA-induced silencing complex (miRISC) or micro-ribonucleoprotein particles (miRNP). The miRNA is embedded into an Argonaute (AGO) protein, a family of proteins with four conserved family members in mammals (AGO1 to AGO4)[20]. MiRNAs guide the miRISC to its target RNA and mediate sequence specific binding, however, they are not necessary for the regulatory effect [21]. miRISCs also include other proteins which participate in assembly and mediate target regulation [22].

In metazoans, miRNAs arise from two genetic origins: They are either transcribed by RNA Polymerase II from individual loci, resulting in pri-miRNA molecules, or as parts of introns of protein coding genes (Figure 1.1A). Pri-miRNAs are further processed by the Drosha/DGCR8 protein complex to produce smaller, ca. 80 nucleotides long precursor miRNAs (pre-miRNAs). Pre-miRNAs usually have a hairpin structure and the mature miRNA sequence is located in the stem region. Intronic miRNAs are processed into pre-miRNAs during splicing. The pre-miRNA is transported to the cytoplasm where it is cleaved by Dicer and subsequently embedded in an AGO protein upon formation of the miRISC [23]. Dicer is indispensable for miRNA maturation and, consequently, a knock-out of the Dicer gene results in a complete loss of all miRNA function. Early in miRNA research, Dicer knock-outs have been used to show that miRNAs are necessary for almost all biological processes.

Many details of the mechanism behind miRNA-mediated gene regulation remain unclear. It has been shown that mammalian AGO2 is able to endonucleolytically cleave mRNA [24, 25] if there is perfect base pairing between miRNA and target mRNA. However, base pairing is usually incomplete whereupon AGO proteins exert their function through recruitment of effector proteins. This leads to either deadenylation of the mRNA and subsequent degradation or inhibition of translation (Figure 1.1B). The process of deadenylation is initiated through interaction of AGO with glycine-tryptophan protein of 182 kDa (GW182), which, in turn, recruits the CCR4–NOT deadenylating complex. The molecular mechanisms behind translational repression are poorly understood and different modes of action have been proposed. This includes repression of translation at the initiation stage as well as elongation.

Generally, the binding properties of miRNA-mRNA interactions are not fully understood. Two points remain widely discussed: Firstly the miRNA-mRNA hybridization structure of effectual binding sites and secondly the localization of binding sites on the target mRNA. Early studies based on known lin-4 and let-7 binding sites and computational analysis of evolutionary conservation suggested that complementary binding of nucleotide 2 to 8 from the 5' end of the miRNA nucleates the binding [26, 27, 28]. Later, this was supported by analyses of the crystal structures of AGO-RNA complexes. These studies showed that nucleotides 2 to 10 of the miRNA are stacked within the AGO protein. Nucleotides 2 to 6 are oriented for nucleation with the target mRNA while nucleotides at position 10 and 11 are less likely to hybridize [29, 30, 31]. The nucleotides 2 to 8 have hence been named 'seed region' and binding sites with seed complementarity are referred to as 'canonical'. Today, seed complementarity is widely accepted as one of the major determinants of functional repression of target mRNAs [32]. Lin-4 and let-7 have the vast majority of their binding sites in the 3' UTR of their target mRNAs [18] and binding sites in the 3' UTR were shown to be under strong evolutionary selection [32, 33]. Based on these observation, functional miRNA binding sites were thought to be preferably located in the 3' UTR of protein coding genes.

Since the early days of miRNA research, computational methods based on seed complementarity in the 3' UTR have been developed for *de novo* prediction of miRNA target genes. Many methods also include free energy of miRNA-mRNA binding and binding site accessibility [34]. Beyond that, individual target prediction methods employ various additional features to improve their performance. TargetScan, a well-known target prediction tool, uses evolutionary conservation of the binding site as well as additional sequence features such as A-U content and location at the end of the 3' UTR [26, 28, 35, 36]. Various other target prediction tools have been extensively reviewed [37, 34, 38]. However, the results of different prediction methods are often inconsistent and their performance can only be assessed based on the small set of known miRNA-mRNA interactions [39, 40]. Moreover, all common prediction methods produce high numbers of false positive results [41].

More recently, biochemical methods to elucidate miRNA-mRNA interactions were developed. They are based on sequencing of cross-linked AGO-RNA complexes (CLIP-seq) and allow to generate maps of all miRNA binding sites on their target mRNAs. Several modifications of this approach have been reported: HITS-CLIP [42, 43], PAR-CLIP [44], iCLIP [45], CLASH [46] and CLEAR-CLIP [47]. All methods use UV cross-linking which forms irreversible, covalent bonds between proteins and nucleic acids in close proximity and does not cross-link proteins. Subsequently, the protein-RNA complex of interest is immunoprecipitated with an antibody and the bound RNA is extracted, transcribed into cDNA and sequenced. Binding sites of the protein are recovered by mapping the sequences to a reference genome. PAR-CLIP extends this approach by introducing the photoreactive ribonucleoside analog 4-thiouridine, which is incorporated into all transcripts. After cross-linking, characteristic sequence changes from T to C are detected in the sequenced cDNA. These changes mark binding sites and allow to discriminate RNA which is precipitated but not bound. CLASH and CLEAR-CLIP ligate the miRNA and the bound mRNA and sequence these hybrid. This allows to directly identify the binding site, while other CLIP methods map to the genome by assuming complementarity in binding of miRNA and mRNA. In summary, these studies complemented the aforementioned interaction paradigms and suggested that binding occurs in all parts of the target mRNA and that the majority of binding sites is non-canonical, i.e. they do not show full complementarity of the miRNA seed region [46]. This finding contradicts the basic assumptions used for computational target prediction methods and questions their utility for functional miRNA research.

In CLIP-seq data analysis, the exact binding site and specific miRNA have to be determined computationally. Next to the straight forward approach of sequence alignment to the reference genome, several computational models were developed for the prediction of binding site properties from CLIP-seq data. "PAR-CLIP miRNA assignment" (PARma) integrates sequence position and characteristics of the experiments, such as the specific processing signature of the nuclease used in sequencing library preparation, into a generative model that scores the most likely miRNAs [48]. MicroMUMMIE combines various additional features such as the type of miRNA seed, evolutionary conservation, sequence composition and positioning of the binding site within the peaks of CLIP-seq reads in a framework that outperforms sequence-only methods in prediction of the specific miRNA [49]. Lastly, MIRZA uses a biophysical model of miRNA-mRNA interactions that does not assume seed complementarity or other binding paradigms [50]. The energy parameters of the interaction model were inferred from CLIP-seq data. Interestingly, without assuming seed complementarity, this model identified even more non-canonical binding sites than previous methods.

Next to silencing of target genes, there is also few evidence for positive regulation of gene expression by miRNAs [51, 52, 53]. Possible explanations for these effects are subsumed under the concept of competing endogenous RNA (ceRNA) [54] which describes indirect regulatory effects between RNAs. There can be more potential miRNA binding sites than available miRNA molecules and target genes thus compete for binding to shared miRNAs. Up-regulation of a transcript with many binding sites for a given miRNA can decrease or abolish the effect on other target genes by acting as a miRNA sponge. These regulatory networks have been demonstrated to participate in formation of various cancer types [55, 56, 57, 58]. The ceRNA hypothesis was also considered to explain tissue specific regulatory effects which have been demonstrated for many miRNAs. Next to gene silencing, other miRNA functions have been suggested. In 2004 already, miRNAs were shown to play a role in DNA methylation [59] and recently the evidence for an involvement in epigenetic regulation accumulated.

Ambiguity of individual miRNA-mRNA interactions and complex regulatory relationships severely hinder functional miRNA analyses. Despite recent advances in CLIP-seq technologies, the results still suffer from a lack of reproducibility [60]. Moreover, miRNA-mediated regulation identified by *in vitro* over-expression and loss-of-function experiments is often not supported *in vivo* [13].

1.2 MicroRNA-independent RNA-binding proteins

AGO, as a key component of miRISC, binds to the mRNA and can thus be categorized as a RNA-binding protein (RBP). However, next to AGO there are a multitude of miRNA independent RBPs. Throughout their life-cycle of transcription, processing, transport, translation and decay, mRNAs are bound and accompanied by various different proteins in messenger-ribonucleoprotein (mRNP) complexes (Figure 1.1C) [61, 62]. In a recent effort to catalogue RBPs, more than 1500 proteins have been identified [63].

Since the early 1990ies, biochemical in vitro methods have been used to study RNA-protein interactions. In particular, systematic evolution of ligands by exponential enrichment (SELEX) has been employed to identify RNA-binding elements of proteins [64]. Based upon this in vitro data, many new RBPs were identified through sequence homology and prediction of protein domains [65]. Later on, methods based on immunoprecipitation of cross-linked RNA-protein complexes followed by microarray analysis of the RNA component allowed further insights into RNA-protein interactions [66].

The universe of RBPs was greatly expanded with new high throughput methods based on next-generation sequencing and mass spectrometry. Interactome capture methods were developed to map the complete RNA-bound proteome. Here, RNAs and proteins are cross-linked with UV light. Subsequently, complexes of polyadenylated RNAs and proteins are extracted and the RNA-binding proteins are analyzed with mass spectrometry [67, 68, 69]. In a reciprocal approach, CLIP-seq methods have been used to identify the targets of individual RBPs [70, 71]. All methods not including a ligation step are suitable for all miRNA-independent RBPs.

Despite advances in experimental technologies to identify RNA-protein interactions, similar issues exist as described for miRNA-mediated gene regulation. There is currently no way to reliably predict all target mRNAs of an RBP and the prospective effect of the RNA-protein interaction. While many binding-motifs are known [72] and sequence-independent binding determinants such as mRNA structure were discovered [73], the RNA-protein interaction network is not fully deciphered. Moreover, RBPs are very diverse in their function, ranging from the splicing machinery to post-transcriptional regulation.

1.3 Functional Analysis of MicroRNAs

Today, the number of genes, transcripts and proteins can be be reasonably estimated, at least in human and mouse [74, 75]. As such, we have an overview of the molecular components of a cell and hence the nodes of the cellular interaction network. However, it is often unclear how these molecular components interact in order to carry out a specific biological process. The size and structure of the cellular interactome remains ellusive. Consequently, functional genomics approaches employ high-throughput technologies to map the regulation, interaction and function of genes and their products.

On DNA level, ChIP-seq experiments provide insights in transcription factors binding to DNA and epigenetic modifications regulating gene activity. On the transcript level, the aforementioned CLIP-seq experiments yield RNAs which are bound by miRNAs and RBPs, interactome capture methods identify sets of RNA-bound proteins and RNA-seq experiments find regulated transcripts. On the protein level, mass spectroscopy based methods are able to identify regulated proteins or proteins undergoing modifications. These high-throughput methods, collectively named *omics* technologies, often generate lists of genes or proteins relevant in the context under investigation. Ever since microarrays have been widely adopted to measure gene expression, a central question arose in high-throughput biology: What is the biological function of the identified genes, transcript or proteins?

The biological function itself is often difficult to capture and the description of a function is a matter of perspective and the designated level of detail. Mitogen-activated protein kinases (MAPK), for example, are a class of proteins that phosphorylate other proteins. As such, their function can be described as ’protein phosphorylation’. They carry out their function in a signaling pathway consisting of a cascade of protein phosphorylations. On this level, the function can be described as ’MAPK signaling pathway’. The pathway is involved in regulation of cell proliferation and differentiation. Consequently, the function of MAPK can be characterized as ’regulation of proliferation’.

The two main categories used for classification of gene sets are biological pathways and functional ontologies. A plethora of computational approaches were developed to associate gene sets to these categories in order to describe their function.

1.3.1 Biological Pathways

A biological cell requires tight regulation in order to maintain its function. The elements of a cell, that is genes, macromolecules such as RNAs and proteins, and metabolites, interact to process external signals, transport information and control changes of cellular processes. In order to structure the cellular regulation map, the large and complex interaction network was sub-divided into distinct smaller modules, referred to as biological pathways. They subsume all components of a cell which interact to perform or change a specific process. Accordingly, pathways are used as a functional category to study the systemic effects and functional role of gene sets such as miRNA targets.

While there is no consistent definition of a pathway, its size or members, they can be categorized in different classes. Signal transduction pathways transfer information, such as external stimuli, and lead to a change in behavior. The Wnt-signaling pathway, for example, transduces the binding of secreted glycolipoproteins of the Wnt family to cell surface receptors and is involved in embryogenesis and cell proliferation [76]. Metabolic pathways describe enzymatic processing of metabolites, such as glycolysis and gluconeogenesis. Next to classification based on function, pathways can be defined by other properties such as association to a disease [77].

Various public databases compile, curate and maintain biological pathways [78]. They focus on different aspects of cellular regulation with varying levels of detail, ranging from presence of an interaction to quantitative chemical reaction rates. Among the most popular are KEGG [79], Reactome [80] and WikiPathways [81]. KEGG was established in 1995 as the first pathway database. It includes both signaling and metabolic pathways with curated, easily interpretable visualizations. However, it gives a high-level overview of the interactions and contains limited metadata. Reactome provides a more detailed view on the individual interactions within a pathway. Enzymatic reactions are described with all input and output components. WikiPathways, on the other hand, is a community effort for curation of pathways. A plethora of other databases collects pathways with a focus on various aspects of cellular regulation. Chowdhury et. al provide a comprehensive, recent overview [82].

1.3.2 Functional Ontologies

Ontologies provide a formal, controlled vocabulary for a body of knowledge in various domains in biology and medicine. They have been developed to describe biological processes, diseases, cellular components or experimental setups. Ontologies increase interoperability of data and allow standardized description of biological topics. Similar to pathways, they are used as a functional categories to classify genes.

One of the key applications of ontologies in biology is to annotate genes, proteins or other molecular components. The most widely used biological ontology and annotation project is the Gene Ontology (GO) [83]. GO consists of three independent ontologies for biological processes (BP), molecular function (MF) and cellular component (CC) which currently contain more than 40.000 terms. They are used to describe processes such as "cell cycle", specific molecular functions such as "protein phosphorylation" and localization within the cell such as "cytoplasm". Each ontology has a single root node and terms are hierarchical, that is child terms are generally more specific than parent terms. However, child terms can have multiple parents and can be connected to parents with multiple relationships. The GO consortium compiles and maintains annotation of gene products, providing the most comprehensive collection of functional annotations. In 2015, more than 50 million gene products for more than 400,000 organisms were annotated. The vast majority of annotations was generated automatically, only 300,000 were manually curated [84]. The annotations are classified by evidence codes which describe the source of the information and, consequently, can be used to filter for more reliable data points.

The number of ontologies for biomedical research is constantly growing [85]. Projects such as the OBO foundry have been established to collect, coordinate and harmonize ontologies [86]. The Ontology Lookup Service [87] and BioPortal [88] integrate a wide range of ontologies and provides a centralized interface.

A plethora of other ontologies and associated annotation projects exist. Disease Ontology for human diseases [89], BRENDA Tissue Ontology for tissues and cell types [90] and the Ontology for Biomedical Investigations for the description of experimental research [91]. The pathway databases described above can be considered as controlled vocabularies as well. They define terms for individual pathways and provide annotation of genes and proteins. They usually do not use all concepts of structured ontologies and do not adhere to data formats commonly used for ontologies.

1.3.3 Functional Enrichment Methods

A multitude of different methods have been developed to associate the lists of genes produced in high-throughout experiments with pathways and GO terms. Here, pathways and GO terms are treated as functional categories and as indicator of biological function. Many of these methods detect an enrichment of the genes of interest in pathways and GO terms [92]. They generally assume that a functional category is relevant if it contains more genes from the analyzed list than expected by chance. The enrichment is usually described with an odds ratio of the genes within a category compared to a background. The significance of the enrichment is determined with a hypergeometric test or the equivalent Fisher's exact test [93]. The simple enrichment approach of using fixed lists of genes has been extended by gene set enrichment analysis (GSEA). Here, the complete result of a high-throughput experiment, such as all genes with fold-changes in a gene expression study, are used to rank functional categories [94]. GO terms are structured hierarchically and if a gene is annotated to a term it is also annotated to all parent terms. This leads to a large number of related terms with overlapping annotations. Recently, probabilistic methods were developed which include all functional categories at once and account for dependencies between the categories [95].

While functional enrichment analyses provide valuable insights into the biological context of gene sets, they suffer from major drawbacks. Firstly, the enriched categories cannot be validated. There is no comprehensive catalogue of validated gene-category associations and thus no ground truth to assess the accuracy of the enrichment method. Comparing associations to limited curated sets of known associations is therefore subject to a large bias in manually curated data. Secondly, the mere association of a gene list to functional categories does not implicate an actual impact on the process. Over-representation of genes in a signaling pathway does not determine the regulatory effect, that is if the pathway signal is activated, silenced or not affected at all. Thirdly, associated categories need further interpretation. If a pathway or GO term is too broadly defined, the relation between functional categories and the biological context under investigation are difficult to deduce.

1.3.4 Functional Enrichment with miRNAs

MiRNAs are important components of the cellular regulation network and, consequently, their impact on biological pathways has long been under investigation. Specifically, miRNA mediated regulation of signal transduction pathways was studied extensively. These pathways generally translate an external signal from a cell-surface receptor into gene expression. The strength of this signal and the specific genes which are affected are modulated through signal processing by the pathway. Among the first observations was that miRNAs preferentially target downstream components of signal transduction pathways [96]. Also, specific regulatory motifs involving miRNAs were uncovered. Both Feedback and feed-forward loops where miRNAs and their target genes are either positively or negatively co-regulated by an upstream factor have been shown to be a dominant factor of miRNA-mediated regulation in signal transduction [97].

Signaling pathways are able to activate target genes, however, in absence of the signal, the target genes have to be switched off. MiRNAs were shown to participate in the repression of signaling targets and realize the 'default repression' of those genes. In a reciprocal case, miRNAs also participate in 'default activation' of genes which are later switched off in a response through external signals [98]. In addition to regulating the default behavior of signaling pathway targets, miRNAs participate in context-dependent processing of signaling. That is, the binding of a ligand to a cell surface receptor can have different results in different cell types. Here, different expression levels of miRNAs confer the cell type specific response. This effect has been demonstrated for both TGFβ-signaling [99] and the Wnt/β-catenin pathway [100]. The effects of miRNAs on signal transduction were extended to signaling robustness, that is the ability of the pathway to retain its function under perturbations from internal or external sources. MiRNAs were shown to confer robustness and participate in accurate timing of cellular signaling and precise regulation of biological changes such as cell-fate decisions in development [101].

Functional enrichment of miRNA target genes provides a way to categorize the functional role of the miRNA. However, the uncertainty in miRNA targeting data poses a challenge. The complete set of effectual miRNA-mRNA interactions is not known and target prediction methods and biochemical approaches for target identification yield contradictory results. Also, the effect size of miRNA-mediated regulation of target genes is ambiguous. It has been shown that miRNAs mostly have mild effects on the protein level of their target genes [102]. Hence, even if a miRNA is thought to target genes of a specific pathway, the pathway as a whole might not be affected at all.

The genome wide distribution of miRNA target genes is unclear. The breadth of research discovering miRNA regulation of biological processes lead to the general assumption that all genes are potential miRNA targets. A general preference for particular target genes might question the applicability of functional enrichment methods for miRNA target genes. Indeed, recent studies suggested a general bias of miRNA targets towards gene with high transcriptional noise [103] and uncovered a bias in functional enrichment analyses [104].

Several computational methods were developed to mitigate problems in functional enrichment of miRNAs. "Functional assignment of miRNAs via enrichment" (FAME), for example, considers the predicted effect sizes of miRNA-mRNA interactions to improve the relevance of associated functional categories [105]. FAME utilizes a weighted, directed bipartite graph of miRNAs and their target genes. Both the number of miRNA binding sites and the predicted strength of the repression, as represented by the TargetScan score, were thus considered. Statistical significance is calculated through comparison to degree-preserving permutations of the miRNA-mRNA graph. FAME detects specific functions for miRNAs with similar seed sequence which are often overlooked by common enrichment methods.

1.4 Data Storage for Functional Analyses

Functional analyses rely on data generated by recent high-throughput *omics* technologies. They measure different types of molecules and thus capture multiple levels of cellular regulation. The key challenge in analyzing regulatory processes in a systemic way is to integrate data from different *omics* levels. However, the combined analysis of multiple *omics* datasets from different sources is difficult. In order to integrate them and draw conclusions, biological knowledge about interactions between the *omics* levels as well as functional annotation such as GO terms and pathways have to be included.

For example, to asses the impact of a transcription factor on gene expression, ChIP-seq experiments have been combined with complementary RNA-seq approaches [106]. The ChIP-seq experiment yields genomic positions of the transcription factors' binding sites while the RNA-seq experiment quantifies the amount of mRNA. To answer the biological question which genes are regulated by this transcription factor, however, additional information about transcribed genes, splice variants and gene promotors have to be considered. We assume a functional interaction only when the transcription factor binds within or near the promoter of a transcribed gene and the mRNA products of this gene can be mapped backed to their genomic origin.

This simple problem comprising two regulatory levels is well studied and can be handled with standardized methods of gene expression analysis [107]. If we add more regulatory levels, however, the required amount of prior knowledge increases exponentially. For example, miRNA expression can be quantified in addition to mRNA from an RNA-seq experiments [108, 109]. To include miRNAs in the analysis, we need information about miRNA target mRNAs. To cover feedback loops, we need to know if the analyzed transcription factor also regulates expression of miRNAs and if the expressed miRNAs regulate the transcription factor. The complexity increases further, if we also include ChIP experiments to analyze the epigenomic landscape of histone modifications, as is done in recent functional epigenetic analyses [110, 111].

In summary, the growing variety of experimental technologies and the complexity of the interactions between measured molecules are key challenges in the effort to answer functional biological questions on a systemic level by integrating omics data sets.

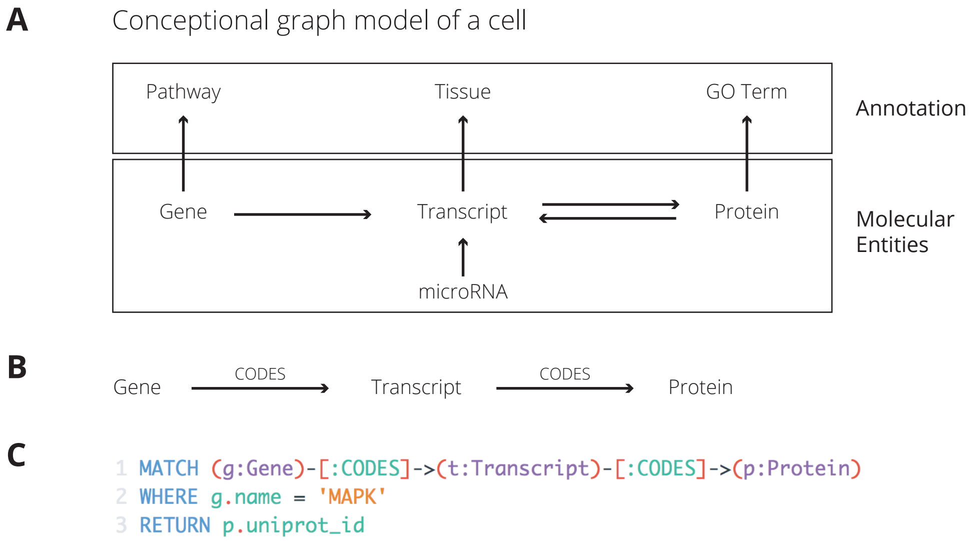
1.4.1 Graph Databases

Relational database management systems (RDBMS) such as MySQL or PostgreSQL have served as the primary means to store all types of data. They generally adhere to the relational data model and use SQL as query language for data retrieval. All data is stored in strongly typed tables where rows represent records and columns denote attributes. Relationships between records are implemented with key attributes.

In recent years, a wide range of new database technologies were developed in answer to the growing challenges of big data. While big data lacks a universal definition, it is usually associated with the growth in data volume, increased velocity of data input and output, and large variety of data types. These new technologies are collectively called 'noSQL' databases, short for 'not only SQL'. They extend the relational data model and provide new paradigms for storage and retrieval of data. Usually, they are classified in four groups: I) Key-value stores provide a simple data model by storing only key-value pairs. II) Wide-column stores implement table like structures which can be describes as nested hash maps. They are geared towards scalability and distributed deployments. III) Document stores use a semi-structured data model to deal with diverse data. IV) Graph databases use a network model to store data as nodes connected by edges.

NoSQL databases have been adopted for the storage and retrieval of biological data sets, especially in the field of next-generation sequencing [112]. Various noSQL data models have been developed which show an increased query performance and scalability compared to relational databases [113].

Graph databases are particularly promising for high-dimensional biological data sets. They store data in a property graph model, that is nodes are connected by relationships and



```
MATCH (g:Gene)-[:CODES]->(t:Transcript)-[:CODES]->(p:Protein)
WHERE g.name = 'MAPK'
RETURN p.uniprot_id
```

Figure 1.2: (A) Simplified graph representation of molecular entities in a cell with annotation elements such as pathways and GO terms. (B) Detailed example of the direct modeling of molecular entities and their relationships. (C) A query example from the graph database neo4j selecting the gene 'MAPK' and returning the proteins encoded by the gene.

key-value properties stored on both. They allow to directly model biological interaction networks which consist of molecular entities and their interactions. Figure 1.2A shows a simplified general graph model of the molecular components of a cell (e.g. genes, transcripts, miRNAs) and associated annotation data (e.g. pathways, tissue expression). Relationships between the nodes in a graph database can be denoted with meaningful relationship types (Figure 1.2B), resulting in straightforward queries directly describing the biological question (Figure 1.2C).

Compared to relational databases, they excel for queries where multi-step paths are retrieved from the graph. The equivalent query in a relational database would use multiple nested JOIN operations which show a significantly lower performance [114, 115]. Moreover, graph databases provide linear local path query performance independent of global graph size. Since many of the questions in todays biology are centered on the interactions between elements of a cell and thus on relationships in the graph model of a cell, graph database have a huge potential to improve storage for high-dimensional biological data.

Graph databases have been used in various fields of biology. Most applications are based on neo4j, the most widely used graph database. The Disease Ontology project is a knowledge base of human diseases [116]. Other ontology related projects have used neo4j to integrate clinical ontologies with medical patient records [117] and to perform semantic text-mining [118]. Structured models from modeling markup languages such as SMBL were also translated to graph databases [119, 120]. Data analysis tools for high-content microscopy [121] and visual data prioritization [122] use neo4j as principal data storage. Also, recent studies store data on protein interactions [123] and Arabidopsis signal transduction [124]. The alternative graph database OrientDB was used for the analysis of networks of non-coding RNAs [125].

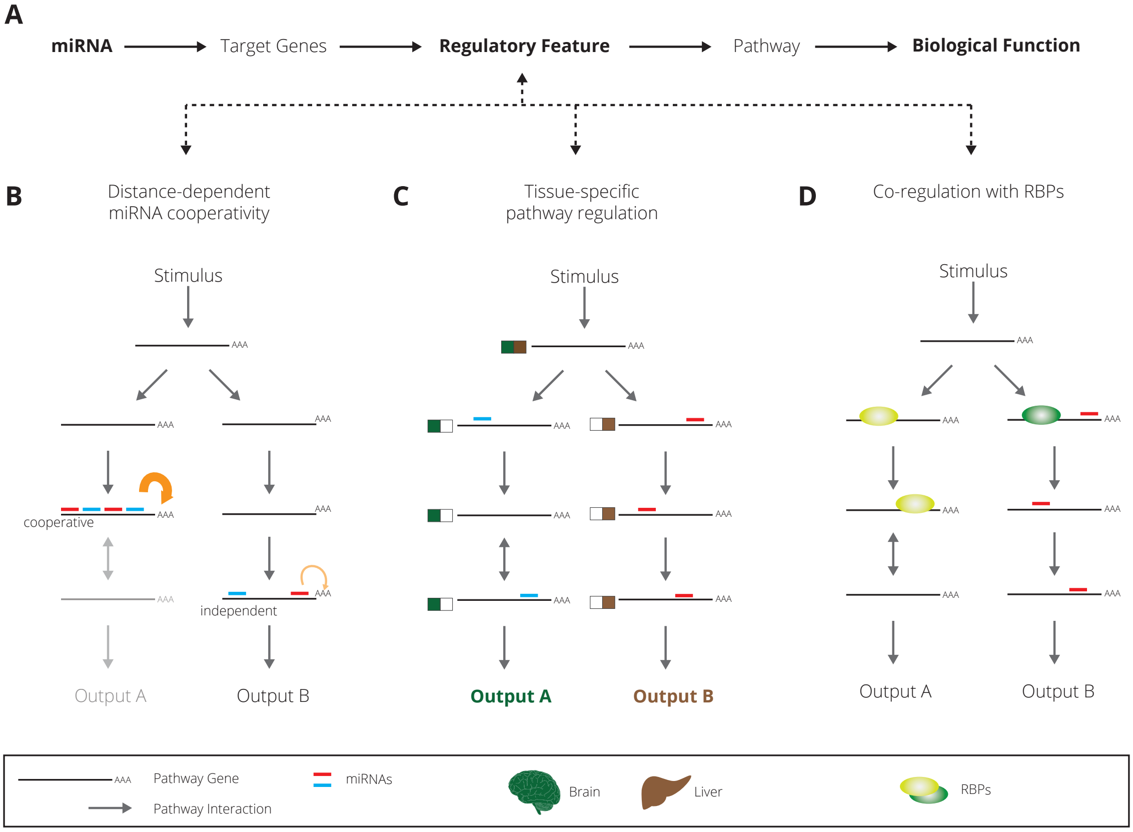


Figure 1.3: (A) Additional regulatory features are introduced to capture systemic effects of miRNA-mediated regulation in functional pathway analysis. (B) Neighboring miRNA binding sites lead to an increased down-regulation of target genes (red and blue miRNA). (C) Genes are not uniformly expressed among tissues. MiRNA regulation of pathways is thus tissue-specific (blue miRNA in brain and red miRNA in liver). (D) miRNAs and miRNA-independent RBPs have been reported to cooperate in regulation of gene expression.

1.5 Research Questions

Post-transcriptional down-regulation of target genes by miRNAs is a widespread phenomenon that influences most mammalian genes. In many cases, however, the actual function of miRNA-mediated regulation *in vivo* is not clear.

The primary question addressed in this thesis was how we can elucidate the biological function of miRNAs (Figure 1.3A). Current functional analyses do not account for the complexity of miRNA regulation due to limitations of targeting data and enrichment methods. We thus included three systemic features of miRNA mediated regulation into functional miRNA analysis in order to capture systemic effects (Figure 1.3B-D):

1. We first addressed distance-dependent cooperativity of miRNAs (Figure 1.3B). It is generally believed that mammalian mRNAs carry multiple miRNA binding sites and are in fact regulated by multiple miRNAs simultaneously. Experimental studies with reporter constructs suggested that binding sites in close proximity increase the down-regulation of target genes and produce cooperative effects, that is the repression of the target gene is higher than the additive effects of the individual binding sites [126, 127]. We asked if binding-site distance functions as a genome-wide predictor of miRNA cooperativity and if cooperative regulation has implications for miRNA function (**Section 2.1**).

2. Protein coding genes are not uniformly expressed among different cell types and tissues [128]. Consequently, miRNA mediated regulation of biological pathways could be tissue-specific and may contribute to cell-type specific modulation which has been reported for various signaling pathways [129] (Figure 1.3C). We addressed the question if tissue specific gene expression is relevant for functional miRNA analysis and how the biological significance of pathway enrichment methods can be improved by incorporating gene expression data to account for tissue specific miRNA regulation (**Section 2.2**).

3. RNAs are constantly bound by numerous RNA-binding proteins. They participate in regulation of all steps of the mRNA life-cycle from transcription to translation. It has been shown that miRNAs and other RBPs interact in regulation of gene expression (Figure 1.3D). We asked if miRNA target genes are also regulated by miRNA independent RBPs and if combined activity of miRNAs and RBPs predicts novel regulatory mechanisms (**Section 2.3**).

Secondly, we asked if we can use these regulatory features to develop tools for experimental miRNA research. Here, a central issue is to choose individual candidate miRNAs from larger sets of potentially relevant ones. We used the analyses presented in this thesis to develop three tools which allow to filter miRNAs for cooperative regulation, tissue specific effects and co-regulation with RBPs (**Section 3.1**).

Thirdly, we investigated novel data storage solutions and specifically addressed the use of graph databases for biological data sets. We devised a unified graph data model for functional miRNA analysis in order to cope with the challenges posed by integration of heterogenous public data sources (**Section 3.2**).

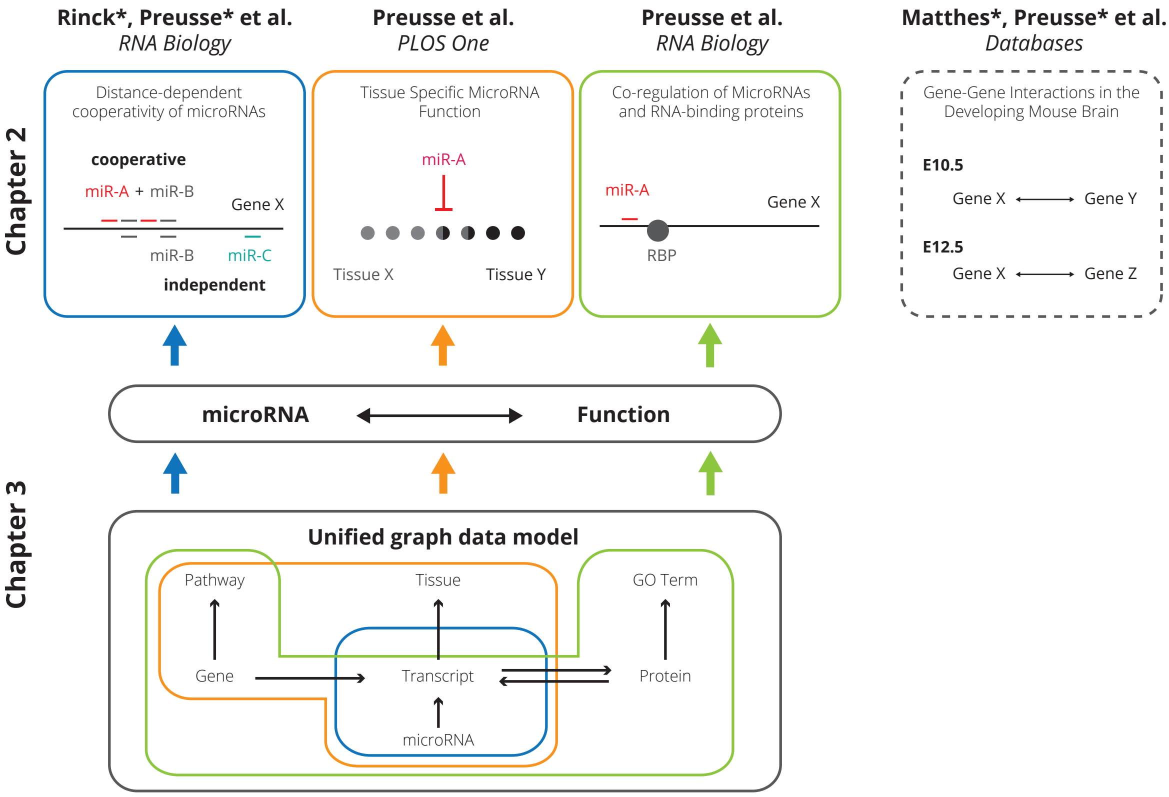


Figure 1.4: Graphical overview. Chapter 2 summarizes the publications composing this thesis. Publication 1, 2 and 3 extend functional miRNA analysis with additional features beyond individual miRNA-mRNA interactions to increase their biological relevance. Publication 4 describes a reference database for genetic interactions in the developing mouse brain. In Chapter 3 we discuss a unified graph data model for the presented analyses and the overall advances in functional analysis of miRNAs.

1.6 Overview of this Thesis

The following provides an overview of this thesis. A graphical overview is presented in Figure 1.4.

The first-author publications composing this thesis are summarized in **Chapter 2**. Publications 1 to 3 describe three novel regulatory features for the functional analysis of miRNAs and present accompanying web applications. For each publication, the comprehensive computational analysis of the regulatory feature is briefly summarized, followed by an overview of the respective web application and the connection to the unified graph data model. Publication 4 is similar in scope but does not involve miRNAs or the unified graph data model.

In **Chapter 3.1** we summarize the aggregated advances in functional miRNA analysis through publication 1, 2 and 3.

Chapter 3.2 provides a detailed description of the unified graph data model which was used to integrate datasources in publication 2 and 3.

Lastly, in **Chapter 3.3** we discuss possible extensions and future projects.

Chapter 2

Publications

In this chapter, the key findings of the first author publications are summarized. The contribution of the author of this thesis is highlighted. Shared first authorships are indicated by * symbols in bibliographic nominations.

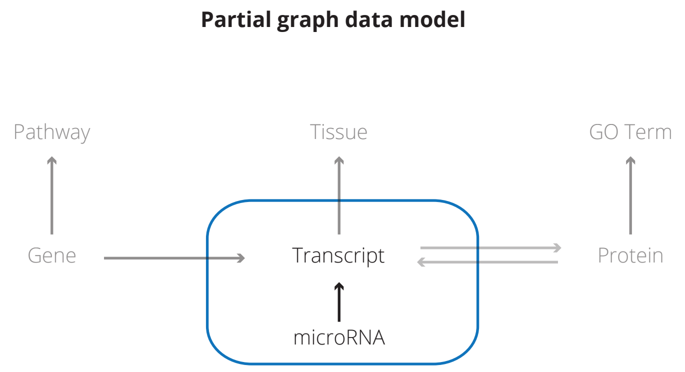


Figure 2.1: The partial data model for this analysis focuses on the relationships between miRNAs and their target transcripts (blue line).

2.1 miRco

Rinck A*, **Preusse M***, Laggerbauer B, Lickert H, Engelhardt S, Theis FJ. *The human transcriptome is enriched for miRNA-binding sites located in cooperativity-permitting distance.* RNA Biol. 2013;10(7):1125–35.

In this publication we investigated the cooperative regulation of miRNA target genes (Figure 1.3B). We addressed the question if binding-site distance functions as a predictor of miRNA cooperativity and further analyzed if cooperativity is a functional aspect of miRNA regulation.

We analyzed the distance distribution of miRNA binding sites genome-wide and found that computationally predicted sites with a distance of 26 nucleotides were enriched. The over-representation demonstrates the biological relevance of distance-dependent cooperativity. Next, we showed that the fraction of cooperative target genes increases if multiple different miRNAs are analyzed together. The increase was significant for miRNA target genes predicted by miRanda and TargetScan as well as target genes identified by HITS-CLIP and PAR-CLIP. Interestingly, the CLIP-seq data sets showed a higher increase. We further analyzed miRNAs which are either co-expressed in a tissue or co-regulated in a disease context. They had a higher number of cooperatively regulated target genes than unrelated miRNAs, indicating that functionally similar miRNAs regulate their target genes in a cooperative manner. In summary, our results demonstrated that distance-dependent miRNA cooperativity is a wide-spread phenomenon that is especially relevant for regulation by multiple different miRNAs.

In order to support experimental miRNA research, we developed *miRco*, a user friendly web tool that predicts cooperative miRNA regulation. It was designed to either predict genes which are cooperatively regulated by a given set of miRNAs or, reciprocally, miRNAs which cooperatively regulate a given gene. *miRco* thereby allows filtering of miRNA lists for candidates which are expected to down-regulate specific genes with a large effect-size.

The data model for this publication was initially implemented using the relational database mySQL. All relevant data, that is the binding sites of miRNAs from prediction tools and CLIP-seq studies, were later included in a graph data model which constituted the first building block of the unified graph data model used in the following publications (Figure 2.1).

The author of this thesis designed and performed the computational analyses presented in this publication in joint work with Andrea Rinck. The underlying data sources, database infrastructure and the web application *miRco* were developed by the author. The corresponding results, methods and discussion were written by the author.

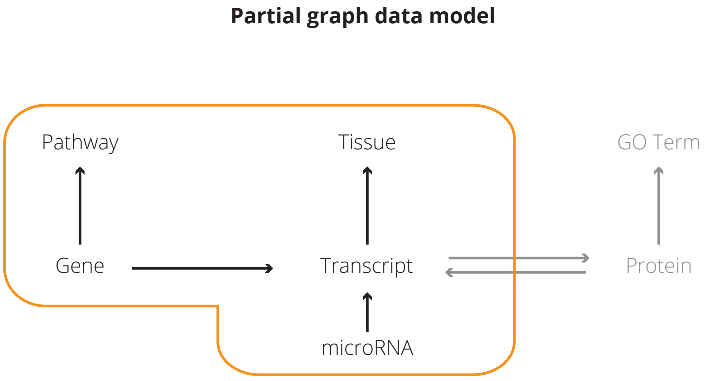


Figure 2.2: The partial data model for this analysis uses miRNAs and their target transcripts from the unified graph data model and augments the molecular data with annotations of tissue expression and pathways (orange line).

2.2 miTALOS

Preusse M, Theis FJ, Mueller N. *miTALOS v2: Analyzing Tissue Specific microRNA Function*. PLOS One. 2016. Accepted.

In this publication, we analyzed the tissue specific regulation of biological pathways by miRNAs (Figure 1.3C). We addressed whether tissue-specific gene expression is relevant for miRNA target genes and pathway enrichment methods. We further analyzed how functional enrichment methods can account for tissue-specific regulation in order to increase the biological relevance of the results.

Firstly, we provide evidence that both miRNA target genes and pathway genes are indeed expressed in a tissue specific manner. On average, only 75% of the miRNA target genes are expressed among 42 human tissues. Pathway genes show an even lower average expression rate with a much higher variance compared to miRNA target genes. Previous tools for functional miRNA analysis do not consider the tissue specificity but instead use the full set of miRNA targets independent of their expression. We thus developed a novel methodology for tissue specific pathway analysis of miRNAs. Here, we filter both miRNA target genes and pathway genes for expression in a given tissue and calculate a pathway enrichment as a proxy for the miRNA function. We highlight the power of the tissue specific enrichment with a comprehensive analysis of miR-199a-3p, miR-199b-3p, miR-571 and the miR-200 family which have been shown to be dysregulated in hepatocellular carcinoma and liver fibrosis. When the tissue filter for liver is applied, our methodology identifies pathways which are in accordance with recent findings regarding the role of the aforementioned miRNAs in pathogenesis. Moreover, we suggest cross-talk of MAPK and Wnt signaling in cancer formation for the miR-200 family and a role of miR-571 in up-regulation of Notch signaling during liver fibrosis. The pathway analysis uses data from various public resources. Pathways from KEGG, Reactome and WikiPathways were included. MiRNA target genes from TargetScan and Miranda as well as from CLIP-seq studies collected by StarBase were used. The tissue filter was based on expression data from the EBI expression atlas.

To support experimental miRNA research, the enrichment methodology was included in the updated *miTALOS v2*, a web application that predicts tissue specific pathway regulation by single or multiple miRNAs. The user can select miRNAs and a tissue of interest and *miTALOS v2* calculate a tissue specific pathway enrichment. *miTALOS v2* thereby helps researchers to identify miRNAs which are likely to influence biological processes in a tissue of interest.

For this publication, we extended the graph data model which was derived from the initial relational data model used in *miRco* (Section 2.1). MiRNA targeting data curated for the analysis of miRNA cooperativity was augmented with the aforementioned pathway and gene expression date sets (Figure 2.2). The graph data model implemented in neo4j has several advantages compared to relational databases: Data queries are more flexible, data updates are easier and, most importantly, changes to the data structure can be performed more easily.

The author of this thesis designed the study, performed all computational analyses, developed the novel database backend and pathway enrichment methodology, implemented the *miTALOS v2* web application, and wrote the paper.

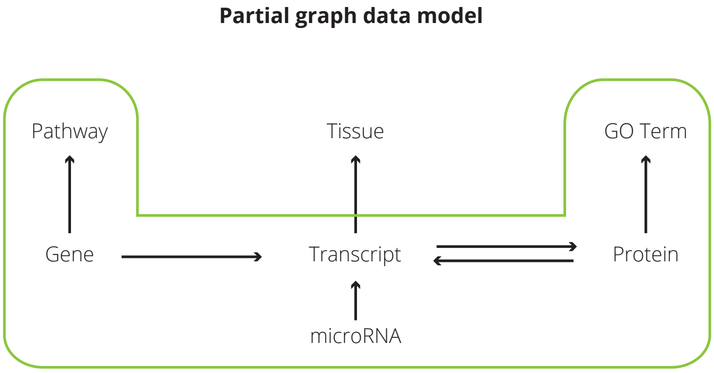


Figure 2.3: In this study the graph data model was further extended with GO annotations and links between RBPs and their target transcripts (green line).

2.3 SimiRa

Preusse M, Marr C, Saunders S, Maticzka D, Lickert H, Backofen R and Theis FJ. *SimiRa: A tool to identify coregulation between microRNAs and RNA-binding proteins.* RNA Biol. 2015;12(9):998–1009.

In this publication, we analyze co-regulation between miRNAs and miRNA-independent RBPs (Figure 1.3D). We addressed the question whether genes are regulated by both miRNAs and RBPs on a genome wide scale. We further asked if we can combine RBPs and miRNAs in pathway and GO term enrichment methods to predict functional similarity and regulatory interactions.

We localized a set of 19 RBPs on different levels of mRNA processing based on a GO term analysis. We further showed that RBPs have on average more target genes than miRNAs but indeed have distinct target sets, which indicates specific functions. We analyzed two features of RBPs to highlight their importance in gene regulation: Many genes are targeted by more RBPs and miRNAs than statistically expected and RBPs, but not miRNAs, preferably target cellular interaction network hubs. This suggested that some genes are under tight control by both RBPs and miRNAs and that RBPs regulate genes with important roles in signaling. For this study, we used CLIP-seq based targeting data for both miRNAs and RBPs. While CLIP-seq methods provide more reliable target genes than computational prediction, they still suffer from large error rates. In order to overcome these errors and to identify functionally similar miRNAs and RBPs, we developed a method which compares enriched functional categories such as pathways and GO terms for both classes of regulators. We compared pairs of miRNAs and RBPs and showed that similarity in target genes and similarity in enriched categories do not correlate. In a case study with human Pumilio proteins, we identified a known interaction with miR-221 and miR-222 in regulation of the tumor suppressor gene p27. The similarity of enriched pathways and GO terms is in the top 10% of all miRNA-RBP pairs. The similarity of target gene sets, however, is ranked at 78%. Next, we suggested a cooperation of the nuclear RBP TAF15 with miR-590-3p and miR-495 with a potential role in regulation of cell cycle and differentiation.

Our method to compare post-transcriptional regulators was included in *simiRa* in order to support experimental miRNA research. *SimiRa* is a web application that allows to identify similar miRNAs and RBPs by exploring their functional neighborhood. Here, the user can select a miRNA or RBP of interest and display the network of similar miRNAs and RBPs defined by both overlap in target genes and associated pathways and GO terms. *SimiRa* thereby identifies miRNA candidates which participate in combined regulation with RBPs.

For this study, the graph data model from the *miTALOS v2* study (Section 2.2) was extended with data on RBPs by including a link from proteins to transcripts (Figure 2.3). Moreover, GO annotations for proteins for proteins were included further expanding the functional annotations. The flexibility of the neo4j implementation supported the extension of the graph data model without refactoring of existing data.

The author of this thesis designed this study, performed all computational analyses, curated all miRNA data, developed the database backend, implemented the *simiRa* web application and wrote the paper.

2.4 IDGenes

Matthes M*, **Preusse M***, Zhang J*, Schechter J, Mayer D, Lentes B, Theis FJ, Prakash N, Wurst W, Trümbach D. Mouse IDGenes: a reference database for genetic interactions in the developing mouse brain. *Database (Oxford). 2014;2014(0): bau083 – bau083.*

The development of the brain in the mouse and other vertebrates is a complex process of patterning along all axes in the embryo. Knowledge of this process is necessary for the understanding of complex neurodegenerative diseases such as Parkinson's disease and Alzheimer's disease as well as neuropsychiatric disorders such as schizophrenia and autism.

Public databases provide gene expression data of the developing mouse brain with high spatio-temporal resolution. However, they neglect the genetic interactions that control neural development. In this study, we developed *Mouse IDGenes*, a reference database of genetic interactions in the mouse brain. The database is manually curated and contains detailed information on gene expression and gene-gene interactions. A novel spatio-temporal model of the developing mouse brain at stages E8.5, E10.5 and E12.5 was developed to allocate interactions with high resolution. To highlight the utility of *Mouse IDGenes*, we used a support vector machine to infer new target genes of Wnt/β-catenin signaling. Dkk3 was predicted as a direct Wnt1 target and validated experimentally using luciferase reporter assays.

Mouse IDGenes was made publicly available as a web application. The interface allows to search for specific genes or brain regions and developmental stages. The gene expression and genetic interaction data are displayed in a single result table. Gene-gene interaction are further classified as direct or indirect and activation or repression. The published source of the interaction is directly linked to PubMed. A key concept of *Mouse IDGenes* is its extensibility. New gene expression data and genetic interactions can be added by the research community. *Mouse IDGenes* can thus keep up with new findings in brain development and serves as an important resource for future research into neuronal diseases.

The author of this thesis primarily implemented the data input and update methodology, developed the web interface and wrote the respective result chapters.

Chapter 3

Discussion

In this chapter, we will discuss the overarching topics of the publications presented in this thesis. Firstly, the novel approaches for functional miRNA analyses are discussed with respect to recent developments in this fields. Secondly, we derive a unified graph data model from the database infrastructure which was developed to support the miRNA analyses. The model is described in detail and applications beyond miRNA research are discussed. Lastly, the outlook is presented.

3.1 Functional MicroRNA Analysis

There is no doubt that silencing of gene expression is a central function of miRNAs. The recent advances in system-level analyses of miRNAs with new *omics* technologies provided valuable insight into the concepts behind miRNA mediated regulation. However, they also left many questions unanswered and raised concerns about paradigms which have been commonly accepted before.

Ambiguities in miRNA-mRNA interactions hinder system-level analyses of miRNA regulation. Indirect regulatory effects in the miRNA-mRNA network have been shown but the relevance of the ceRNA concept has to be strengthened. Further, the impact of miRNAs and their mis-regulation on biological processes and complex phenotypes is difficult to decipher and the apparent tissue and cell type specificity of miRNA regulation is not yet explained.

Functional enrichment analyses are a key tool in miRNA research in order to circumvent incomplete miRNA targeting data and improve the system-level understanding of miRNA-mediated regulation of biological processes. They map the presumed target genes of miRNAs onto known functional categories and look for distinctive features such as over-representation. The basic assumption is that these features do not occur by chance and hold a biological meaning. Consequently, if the target genes of a miRNA are over-represented in a MAPK signaling pathway, one would assume that the miRNA actually influence this signaling cascade and its corresponding biological processes. Functional enrichment tools were developed to bridge the gap between individual miRNA-mRNA interactions and biological functions [130]. However, they rely on miRNA target genes and are therefore subject to the uncertainties in miRNA targeting data. Moreover, enrichment methods are difficult to validate and falsify. The predicted functions can only be tested for individual cases. As long as there is no commonly accepted, comprehensive, accurate and stable catalogue of defined miRNA functions, *in silico* predictions cannot be compared and evaluated systematically. Even though some methods use self-compiled sets of miRNA-function associations as benchmark [105], the limited knowledge of miRNA function is likely biased and incomplete.

In the publications presented in this thesis, we sought to address the targeting uncertainties and problems with functional enrichment methods by incorporating additional features beyond individual miRNA-mRNA interactions. We used distance-dependent miRNA cooperativity, tissue specifc gene expression and interaction of miRNAs and RBPs. We only used features which have been demonstrated experimentally and did not infer them from miRNA targeting data. This was done in order to deal with the problem of limited assessability of functional predictions by providing a clear line of argument from biological motivation to our genome-wide analyses and extensive case studies.

First, we analyzed the concept of distance-dependent miRNA cooperativity. It was shown that miRNAs which bind in close proximity show a drastically increased effect on target regulation [127]. Presumably, this effect is a result of interactions between adjacent miRISCs which stabilize the miRISC-mRNA complex. It has long been known that miRNAs target multiple mRNAs and that mRNAs carry binding sites for multiple miRNAs [28, 35]. Consequently, it has been assumed that multiple miRNAs act in concert to achieve target regulation [131] and miRNAs and mRNAs form a complex regulatory network. The concept of cooperativity sheds light on this network and moves the target analyses towards are more quantitative perspective. If multiple miRNAs target a gene cooperatively, i.e. they bind in very close proximity, one can assume a strong effect on target regulation. This assumption can then be used to filter the large number of potential miRNA target genes for those which are more strongly repressed and therefore more relevant for the function of the analyzed miRNA. A similar study recently published another distance-dependent cooperativity analysis, supporting our approach [132]. Schmitz et al. incorporate minimum free energy predictions, molecular dynamics simulation of the cooperative targeting and resulting binding affinity into their workflow. However, their analysis focuses on triplets of two miRNAs and one mRNA while our approach can predict cooperative regulation for groups of miRNAs. Indeed, our analysis indicated that cooperativity is more prevalent if multiple miRNAs are considered, that is clusters of miRNA binding sites in cooperative distance are likely composed of alternating sites for several miRNAs. More importantly, the concept of distance-dependent cooperativity has recently been shown to be relevant for miRNA-mediated regulation in myotonic dystrophy type 1 [133], neural stem cells [134] and Hepatitis C virus replication [135]. These findings clearly support the notion of miRNA cooperativity, its impact on biological processes and our *in silico* method to predict cooperative regulation.

We then moved from the level of miRNA-mRNA interactions to the functional aspects of miRNAs. More specifically, we focused on the influence of miRNA regulation of biological pathways. We considered pathways as a proxy for the effect of miRNAs on biological processes and used them to narrow done the miRNA function. However, only a subset of genes is expressed in each cell at any given point in time. Consequently, only a subset of miRNA target genes and pathway genes are available for regulation. Indeed, many miRNAs show tissue or cell type specific effects and the limited reproducibility of *in vitro* effects in *in vivo* studies has been partly attributed to the absence of suggested target genes [13]. Other miRNA enrichment tools, such as DIANA-miRPath [136] or miRGator [137], do not consider this effect. In order to improve the biological relevance of the predicted pathway associations, we incorporated tissue specific effects into a pathway enrichment. Most importantly, we used the latest version of the EBI Expression Atlas which introduces the concept of baseline expression derived from RNA-seq experiments [138]. Compared to microarray based gene expression experiments, this allows to define whether a gene is expressed or not over several tissues and cell types. We were able to show that filtering miRNA target genes and pathway genes for tissue expression drastically changes the result of the pathway enrichment and therefore the biological interpretation of the results. In conjunction with expression of the miRNA itself, this also provides a step towards understanding of the ceRNA hypothesis. If only few miRNA target genes are expressed, repression of these targets might be elevated because competition for the miRNA from other binding sites decreases.

We then extended the functional analysis of miRNAs and also considered other miRNA-independent RNA-binding proteins. In a cell under physiological conditions, RNAs are constantly bound by many different proteins and packaged into ribonucleoprotein particles. MiRNA-AGO complexes are only a part of the mRNA interactome. Many examples for regulatory interactions between miRNAs and miRNA-independent RBPs are known and data from CLIP-seq methods was used to generate mRNA binding maps of both regulators. Similar to the idea of miRNA cooperativity, it has been shown that RBPs bind in close proximity to effectual miRNA target sites [139, 140]. Hence, the hypothesis that miRNAs and RBPs cooperate in order to carry out their biological function ensues. We used 19 RBPs and their CLIP-seq derived target genes from the doRiNA database. We first found that all RBPs have distinct sets of targets ranging from hundreds to few thousands of genes. While this observation may seem trivial, it does prove an important point: All RBPs, even those which have been described as generic splicing factors, interact only with specific, differing genes. This implies that they convey regulatory functions for specific biological processes. Therefore, we extended the functional analysis of individual miRNAs and moved on to comparing the functional environment as a whole in order to find possible interactions between miRNAs and RBPs. We performed a pathway and GO term enrichment for all miRNAs and RBPs. By comparison of the functional neighborhood we identify miRNAs and RBPs which are functionally similar and likely interact. We used this approach to predict miRNAs interacting with the RBP TAF15. TAF15 has been shown to indirectly interact with miR-17-5p and miR-20a-5p in regulation of cell-cycle and proliferation. We proposed a combined role of TAF15 and miR-590-3p, a miRNA which has been poorly characterized at that time. Intriguingly, several very recent experimental studies support an involvement of miR-590-3p in cell-cycle, proliferation and migration in the context of different cancer types [141, 142, 143]. In general, the combined analysis of miRNAs and RBPs came more into focus. Other *in silico* tools studied the cooperation of both in post-transcriptional regulation [144]. However, HafezQorani et al. focus on the combined regulation of individual target genes while we demonstrated that known interactions between miRNAs and RBPs are evident in their functional similarity. Next to *in silico* efforts, several new miRNA-RBP interactions have been demonstrated experimentally [145, 146]. This further supports the idea of combined, functional analysis of miRNAs and RBPs.

In summary, our computational approaches use additional features of miRNA regulation to identify novel regulatory effects. They stepwise extended the scope of the analysis, from cooperative miRNA-mRNA interactions through functional analysis of individual miRNAs to comparing the complete functional environment of miRNAs and RBPs. All the used features have been demonstrated experimentally and are not only deduced from theoretical analyses. The basic concepts of our computational methods have been picked up by related and complimentary tools. The most important evidence for the impact of our methods, however, is the growing experimental evidence for the proposed regulatory mechanisms.

We used the presented concepts of functional miRNA analysis to develop three web applications: *miRco*, *miTALOS v2* and *simiRa*. They all share a common purpose: Identify testable candidates from a set of potentially interesting miRNAs. In experimental miRNA research, large-scale experiments such as differential expression analysis in the system of interest often result in lists of dozens of potentially interesting miRNAs. Also, due to the extensive and confusing amount of literature on miRNAs, a simple PubMed search yields many potentially relevant miRNAs. However, in many cases functional testing is not feasible for so many candidates. High-throughput screenings are not possible in all biological systems and individual testing is labour intensive and costly. Therefore, the long list of interesting miRNAs has to be narrowed down to a smaller set that can be handled experimentally. Our tools were developed to achieve just that. The workflow generally starts by selecting a miRNA of interest, such as the top regulated one from a differential expression study. Our analysis tools then highlight functional features of this miRNA and allow the experimental researcher to select the miRNA that is most likely relevant for his biological question.

3.2 A Unified Graph Data Model

In the three presented publications dealing with functional miRNA analysis, we used many data sets from various public resources. They include data on different molecular levels, that is genes, transcript, proteins and miRNAs, as well as annotation data such as pathways and GO terms. The extended approaches for functional miRNA analysis based on these datasets required a way to integrate them and perform queries spanning different data sources and molecular levels. During development of the functional analysis tools, we developed a novel data management system to store all used data in a central place. Such a system has to fulfill specific requirements in order to handle biological data sets:

- The underlying data sources are updated frequently and, therefore, data updates must be simple and fast.
- Scientific development requires a data model that is easy to refactor.

- The data model must be extendable and allow to store additional entities and relations from future data sources.

- Experimental resources are often burdened by missing data. The data management system must be able to flexibly store and query incomplete data.

- Queries are generally centered around biological entities (i.e a miRNA) and their interaction or annotation (i.e. the target genes of a miRNA within a specific pathway). Thus, efficient queries over multi-step paths must be possible.

- In order to increase accessibility of the database, query mechanisms for complex questions have to be simple and straight-forward.

Requirements for data management became more demanding because the data generated in biology has changed during the past decade. This change was driven by two major developments: The new *omics* technologies allow the study of biological processes on the systems level and multiple *omics* data sets are combined to answer biological questions. Arguably the most important experimental technologies are next-generation sequencing (NGS) in the field of genomics, epigenomics and transcriptomics and mass spectroscopy based methods for proteomics and metabolomics. They were refined, improved and, most importantly, became affordable for researchers in all fields of biology. As a consequence, the volume of data grew exponentially. A single NGS experiment analyzing, for example, binding of a transcription factor to DNA, produces sequence data in the range of 10 to 20 gigabyte. Projects like the Roadmap Epigenomics Mapping Consortium [147] have to deal not only with a single NGS data set but with hundreds of those experiments. Another example are population scale whole-genome sequencing projects which aim at elucidating human genome variation. The "1000 genomes" [148] project and the 100,000 genomes project of the UK Department of Health have to handle massive amounts of sequencing data.

However, the sheer size of the data is not the most pressing problem. In comparison to the data volume that is produced by social networks like Facebook or Twitter, who have been driving the development of new big data technologies, even the large genomics and epigenomics projects are manageable. Big-data solutions to store large datasets are well-suited to handle them and parallel computing approaches are used for processing and analysis. The central issue in modern biology lies within the complexity of the data under study. Firstly, the data is very heterogenous and, secondly, multiple data sets have to be integrated in order to answer biological questions.

To cope with these challenges, we developed, over the course of the presented publications, a graph model of molecular biological entities, their molecular interactions and annotations. Parts of the graph model were implemented for the respective analyses. From this we derive a unified graph data model, laying the groundwork for a comprehensive model of the cell which can be extended to incorporate all known entities and their interactions.

3.2.1 Implementation of the Combined Graph Model

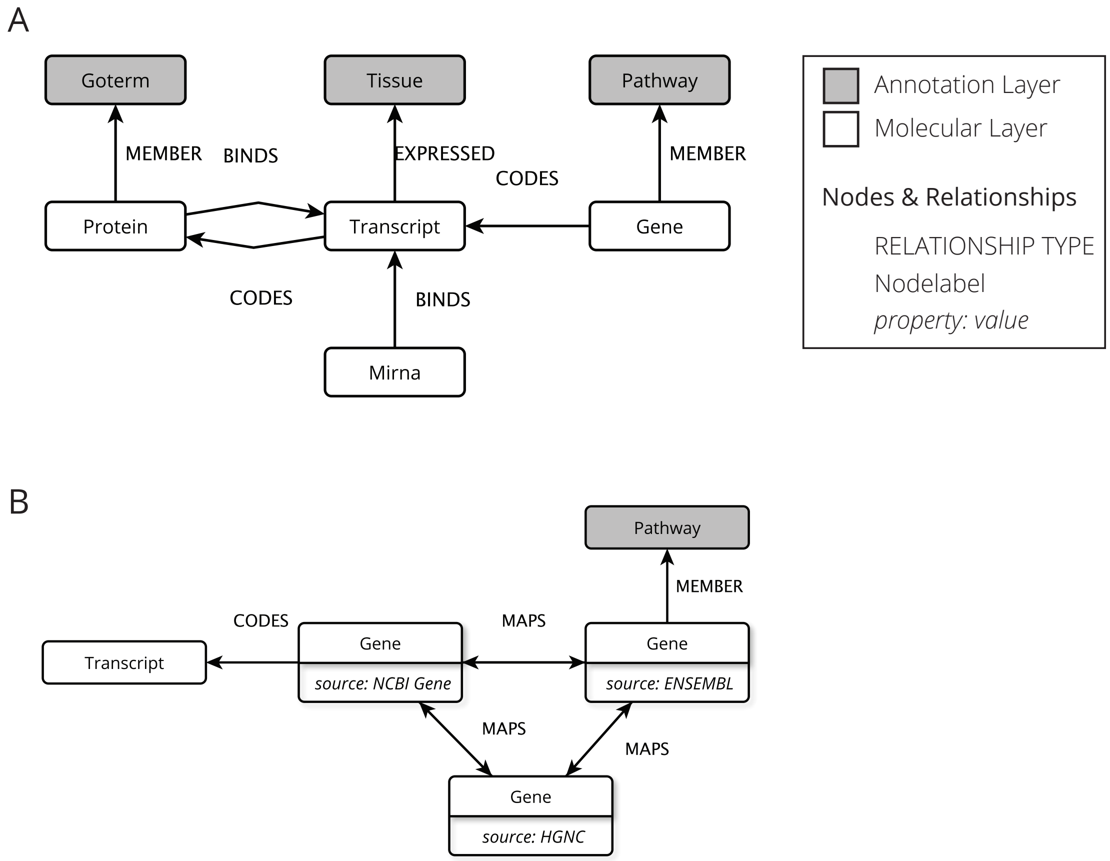


Figure 3.1: Combined graph model used in this thesis. Following general neo4j conventions, node labels are capitalized and relationship types are all capitals. A) Molecular entities are modeled as nodes, their interactions as relationships. The annotation layer models annotation elements and links them to the molecular layer. B) Multiple databases for the same molecular entity are modeled with nodes carrying the data source as property. The mappings are modeled by relationships.

The combined graph model of the analyses presented in this thesis contains a molecular layer with genes, transcripts, proteins and miRNAs and an annotation layer with pathways, GO terms and tissues (Figure 3.1A). Relationship types were chosen for expressive and easy-to-understand queries. Data was extracted and integrated from various public resources, including the large genome consortia (ENSEMBL, NCBI Gene) and UniProt, the principal protein knowledge base (Table 3.1).

The model was implemented with the graph database neo4j which allows to natively store data as a labeled property graph (Section 1.4.1). In a relational database, each node type would be modeled with a separate table. The properties are predefined by the table structure. Similarly, all relationship types would be modeled in tables with primary keys linking to the respective entities, resulting in a total of 13 tables. Each relationship between previously unconnected entities would need an additional table, leading to a growing complexity in database structure. Moreover, adding properties to single entities or relationships requires expensive refactoring of the complete table structure. In neo4j, nodes and relationships are schema-free and properties can be added or removed for individual nodes and relationships.

For a single organism, the combined graph data model contains several hundred thousand nodes for molecular entities and several ten thousand nodes for the annotation layer. The number of relationships is two orders of magnitudes larger. The core data on miRNA-transcript interactions curated from TargetScan, miRanda and StarBase adds up to more than 14 million relationships. The data model is implemented for both mouse and human independently, that is all molecular entities are specific for the respective organism.

Table 3.1: Datasources used for the combined graph data model

Node label	Datasource	Relationship type	Datasource
Gene	ENSEMBL, NCBI Gene	Gene-CODES-Transcript	NCBI Gene
Transcript	NCBI RefSeq	Transcript-CODES-Protein	UniProt
Proteins	UniProt	Mirna-BINDS-Transcript	TargetScan, Miranda, StarBase
miRNAs	miRBase	Gene-MEMBER-Pathway	KEGG, Reactome, WikiPathways
Pathways	KEGG, Reactome, WikiPathways	Transcript-EXPRESSED-Tissue	EBI Gene Atlas
GO Terms	Gene Ontology	Protein-MEMBER-Goterm	Gene Ontology
Tissues	EBI Expression Atlas		

3.2.2 Advantages of the Graph Model

The query language of neo4j, Cypher, provides a simple way to retrieve data by describing path patterns. A query that finds all target genes of the human miRNA hsa-miR-21 first matches the miRNA as start node, moves along the path to transcripts which are bound by the miRNA and subsequently gets the genes coding for the transcript (Listing 3.1).

Listing 3.1: Query to find miRNA target genes

```
MATCH (m:Mirna {name: "hsa-miR-21"})-[:BINDS]->(t:Transcript)<-[:CODES]
        -(g:Gene)
RETURN distinct(g.name)
```

A key advantage of our novel data model is that more complex queries retain readability and are still easy to understand. The query in Listing 3.1 can be extended to return all KEGG pathways which contain target genes of hsa-miR-21 along with the respective target genes (Listing 3.2). If data was stored in a relational database with entities modeled in normalized tables and relationships implemented with primary keys, the equivalent SQL query would be more complex due to three sequential JOIN operations. Neo4j outperforms relational databases for path queries typical in biology [149].

Listing 3.2: Query to find miRNA target genes in KEGG pathways

```
MATCH (m:Mirna {name: "hsa-miR-21"})-[:BINDS]->(t:Transcript)<-[:CODES]
        -(g:Gene)-[:MEMBER]->(p:Pathway {source: 'kegg'})
RETURN p.name, collect(g.name)
```

When multiple datasources are combined and integrated, ID mapping is a common problem. For all molecular entities, there are several competing consortia that collect and maintain information, e.g. ENSEMBL and NCBI Gene for genes, ENSEMBL and NCBI RefSeq for transcripts and ENSEMBL and UniProt for proteins. They usually have distinct basic assumptions on the definition of the respective molecular entity and thus incompatible naming schemes. ID mapping is, in fact, not a trivial problem and has been discussed for a long time [150, 151]. It is most evident on the level of genes since the large reference genome annotation projects do not agree on a common definition of a gene. The consensus coding sequence (CCDS) project has been established to generate a unified source of protein coding genes for mouse and human with stable mappings to other gene identifiers [152].

Our novel graph data model allows to transparently include mapping data. We included genes with their primary ID from ENSEMBL, NCBI Gene and HGNC, and added mapping relationships between them (Figure 3.1B). This approach allows to include the ID mapping in path queries without relying on external ID conversion tools. The central advantage is that the mapping step can be included in the query without altering path elements left and right of the mapping.

Public resources from the same field often use different ways to identify data. The pathway database KEGG, for example, contains genes identified by NCBI gene IDs. WikiPathways, on the other hand, does not specify the gene identifier and uses NCBI gene IDs, ENSEMBL

IDs and gene names concurrently. With our graph data model, all gene identifications in WikiPathways can be mapped to nodes representing the specific gene ID source. Then, flexible queries allow to include all possible gene nodes in the aforementioned example of identifying pathways targeted by a miRNA (Listing 3.3). The mappings were used for both *miTALOS v2* and *simiRa* and highlight the extensibility of our data model. If, for example, a new database with gene identifiers is developed, it can easily be included into the existing model.

Listing 3.3: Query with gene ID mapping

```
MATCH (m:Mirna {name: ”hsa−miR−21”})−[:BINDS]−>(t:Transcript)<−[:CODES]
      −(g:Gene {source: ’ensembl’})−[:MAPS]−(g2:Gene {source: ’ncbi’})
      −[:MEMBER]−>(p:Pathway {source: ’kegg’})
RETURN p.name, collect(g.name)
```

ID mapping becomes drastically more difficult when multiple levels of molecular entities are considered. Here, biological conditions increase the complexity. Not only different concepts of a gene have to be integrated but also the complex relationships between the central axis of genes, their transcripts and encoded proteins. Generally, genes can have multiple transcripts and protein-coding transcripts can give rise to multiple proteins. Several public resources provide ID mapping service spanning molecular levels, however, as assessment of their performance demonstrated large discrepancies between the results and lack of data updates [153]. While our data model does not solve the ID mapping issue per se, it provides valuable advantages by, firstly, integrating ID mapping steps into the biology-centered queries and, secondly, allowing to transparently use several data sources for each mapping step.

3.2.3 Concept of a Unified Graph Data Model

As described for functional miRNA analysis, our combined graph model is able to capture biological questions and integrate multiple molecular levels and associated annotations. In general, novel methods for data integration in biology became more relevant with the new *omics* technologies [154]. We thus developed the concept of a unified graph data model of the cell based on the model presented above. Graph databases are scalable, highly flexible and allow to directly map biological systems in the way we usually depict them. This provides the ideal basis for a system that stores all molecular entities, their interactions and annotations. The result is a native and easy-to-interpret graph model of the cell which can be used to integrate complex, heterogenous biological datasets.

Sydney Brenner described this approach as a "Cell Map" [155]. He envisioned the cell map as "at once a map of the molecules within cells and a map of the cells in the organism", based on the assumption that everything in biology can be represented as a graph. He further argued that a complete cell map takes the middle-ground between a gene-centric bottom-up and a more organismic top-down approach and therefore allows to integrate data from the genome through all *omics* levels up to physiological processes which can be mapped back to cells and their interactions. However, such a "Cell Map" would need to be complete in order to allow for precise answers of biological questions.

With our combined graph model and its implementation in a graph database we took a first step towards a unified graph data model of the cell, or a "Cell Map". The concept of mapping molecular entities and enriching them with an annotation layer enabled our functional miRNA analyses. During development of our methods, we gradually extended the model and demonstrated that it can adapt to incorporate new data sources without altering existing structures. For example, it is straight forward to integrate more classes of molecular entities, such as metabolites, and their interactions with existing data. Relational databases are, of course, theoretically able to store the same set of data that is included in a graph database. However, the complexity of the database structure would increase exponentially and subsequent changes of the data model would be almost impossible to handle.

3.3 Outlook

While the functional miRNA analysis provided useful insight into miRNA-mediated regulation, we considered additional features for miRNA regulation separately. In future, a combination of the described approaches will allow to decipher further regulatory effects and thus increase the biological relevance.

By extending distance-dependent cooperativity to RBPs, we can generate a clearer picture of the complex network of post-transcriptional regulation. The proposed mechanism behind the cooperative effects is stabilization of protein complexes. By considering all proteins that bind a specific mRNA, we can deduce the strength of the effect of combined regulation through miRNAs and RBPs. The cooperative regulation could be introduced into the functional analysis as a quantitative property, that is by assigning weights to the relationships between miRNA/RBPs and mRNAs. The concept of tissue specificity can be extended in two ways. Firstly, it should not only consider expression of target genes and pathway genes, but also expression of miRNAs and RBPs themselves. Secondly, tissue-specific expression should be included in the cooperativity analysis. By integrating all feature into a single analysis pipeline, we will close the loop from regulators to regulated genes and allow explain tissue specific effects on a much more detailed level. It would allow a tissue specific, combined functional analysis of miRNAs and RBPs with inclusion of cooperativity as a way to predict the strength of regulation.

The combined data model can easily be extended to include more public data sources. Several other pathway databases exist which would extend the scope of pathway analyses. Next to the gene ontology, there are a multitude of ontology and annotation projects which would allow association not only to functional categories but also to diseases. Data sources on genetic variation can be included to link positional data from genome annotations to their molecular manifestations and analyze e.g. the role of SNPs in miRNA function.

In this thesis, we focused on miRNAs and RBPs. However, in future other classes of molecules might be involved in regulation of gene expression. Long non-coding RNAs are a growing field of research and interesting candidates to be included into the extended, combined pipeline for tissue-specific functional analysis.

Bibliography

[1] James B. Sumner. The isolation and crystallization of the enzyme urease preliminary paper. *Journal of Biological Chemistry*, 69:435–41, 1926.

[2] G W Beadle and E L Tatum. Genetic Control of Biochemical Reactions in Neurospora. *Proceedings of the National Academy of Sciences of the United States of America*, 27(11):499–506, nov 1941.

[3] F Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163, jan 1958.

[4] F H CRICK, L BARNETT, S BRENNER, and R J WATTS-TOBIN. General nature of the genetic code for proteins. *Nature*, 192:1227–32, dec 1961.

[5] F Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, aug 1970.

[6] Tobias Maier, Marc Güell, and Luis Serrano. Correlation of mRNA and protein in complex biological samples. *FEBS letters*, 583(24):3966–73, dec 2009.

[7] Christine Vogel and Edward M Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews. Genetics*, 13(4):227–32, apr 2012.

[8] R W HOLLEY, J APGAR, G A EVERETT, J T MADISON, M MARQUISEE, S H MERRILL, J R PENSWICK, and A ZAMIR. STRUCTURE OF A RIBONUCLEIC ACID. *Science (New York, N.Y.)*, 147(3664):1462–5, mar 1965.

[9] C Guerrier-Takada, K Gardiner, T Marsh, N Pace, and S Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3 Pt 2):849–57, dec 1983.

[10] K Kruger, P J Grabowski, A J Zaug, J Sands, D E Gottschling, and T R Cech. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, 31(1):147–57, nov 1982.

[11] G J Hannon, F V Rivas, E P Murchison, and J A Steitz. The expanding universe of noncoding RNAs. *Cold Spring Harbor symposia on quantitative biology*, 71:551–64, jan 2006.

[12] Donny D Licatalosi and Robert B Darnell. RNA processing and its regulation: global insights into biological networks. *Nature reviews. Genetics*, 11(1):75–87, jan 2010.

[13] Joana a. Vidigal and Andrea Ventura. The biological functions of miRNAs: lessons from in vivo studies. *Trends in Cell Biology*, 25(3):137–147, 2014.

[14] Wigard P. Kloosterman and Ronald H.A. Plasterk. The Diverse Functions of MicroRNAs in Animal Development and Disease. *Developmental Cell*, 11(4):441–450, oct 2006.

[15] Witold Filipowicz, Suvendra N Bhattacharyya, and Nahum Sonenberg. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature reviews. Genetics*, 9(2):102–14, feb 2008.

[16] B Wightman, I Ha, and G Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5):855–62, dec 1993.

[17] R C Lee, R L Feinbaum, and V Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–54, dec 1993.

[18] B J Reinhart, F J Slack, M Basson, A E Pasquinelli, J C Bettinger, A E Rougvie, H R Horvitz, and G Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772):901–6, feb 2000.

[19] Eric Huntzinger and Elisa Izaurralde. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nature reviews. Genetics*, 12(2):99–110, feb 2011.

[20] Gyorgy Hutvagner and Martin J Simard. Argonaute proteins: key players in RNA silencing. *Nature reviews. Molecular cell biology*, 9(1):22–32, jan 2008.

[21] Ramesh S Pillai, Caroline G Artus, and Witold Filipowicz. Tethering of human Ago proteins to mRNA mimics the miRNA-mediated repression of protein synthesis. *RNA (New York, N.Y.)*, 10(10):1518–25, oct 2004.

[22] Marc R Fabian and Nahum Sonenberg. The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nature structural & molecular biology*, 19(6):586–93, jun 2012.

[23] Jacek Krol, Inga Loedige, and Witold Filipowicz. The widespread regulation of microRNA biogenesis, function and decay. *Nature reviews. Genetics*, 11(9):597–610, jul 2010.

[24] Jidong Liu, Michelle a Carmell, Fabiola V Rivas, Carolyn G Marsden, J Michael Thomson, Ji-Joon Song, Scott M Hammond, Leemor Joshua-Tor, and Gregory J Hannon. Argonaute2 is the catalytic engine of mammalian RNAi. *Science (New York, N.Y.)*, 305(5689):1437–41, sep 2004.

[25] Ji-Joon Song, Stephanie K Smith, Gregory J Hannon, and Leemor Joshua-Tor. Crystal structure of Argonaute and its implications for RISC slicer activity. *Science (New York, N.Y.)*, 305(5689):1434–7, sep 2004.

[26] Benjamin P Lewis, I-hung Shih, Matthew W Jones-Rhoades, David P Bartel, and Christopher B Burge. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–98, dec 2003.

[27] Alexander Stark, Julius Brennecke, Robert B Russell, and Stephen M Cohen. Identification of Drosophila MicroRNA targets. *PLoS biology*, 1(3):E60, dec 2003.

[28] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, jan 2005.

[29] Yanli Wang, Stefan Juranek, Haitao Li, Gang Sheng, Greg S Wardle, Thomas Tuschl, and Dinshaw J Patel. Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. *Nature*, 461(7265):754–61, oct 2009.

[30] Yanli Wang, Gang Sheng, Stefan Juranek, Thomas Tuschl, and Dinshaw J Patel. Structure of the guide-strand-containing argonaute silencing complex. *Nature*, 456(7219):209–13, nov 2008.

[31] Yanli Wang, Stefan Juranek, Haitao Li, Gang Sheng, Thomas Tuschl, and Dinshaw J Patel. Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature*, 456(7224):921–6, dec 2008.

[32] Andrew Grimson, Kyle Kai-How Farh, Wendy K Johnston, Philip Garrett-Engele, Lee P Lim, and David P Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell*, 27(1):91–105, jul 2007.

[33] William H Majoros and Uwe Ohler. Spatial preferences of microRNA targets in 3' untranslated regions. *BMC Genomics*, 8(1):152, jan 2007.

[34] Sarah M Peterson, Jeffrey A Thompson, Melanie L Ufkin, Pradeep Sathyanarayana, Lucy Liaw, and Clare Bates Congdon. Common features of microRNA target prediction tools. *Frontiers in genetics*, 5:23, jan 2014.

[35] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome research*, 19(1):92–105, jan 2009.

[36] Vikram Agarwal, George W Bell, Jin-wu Nam, and David P Bartel. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4:1–38, aug 2015.

[37] Hyeyoung Min and Sungroh Yoon. Got target? Computational methods for microRNA target prediction and their extension. *Experimental & molecular medicine*, 42(4):233–44, apr 2010.

[38] William Ritchie and John E J Rasko. Refining microRNA target predictions: sorting the wheat from the chaff. *Biochemical and biophysical research communications*, 445(4):780–4, mar 2014.

[39] Marshall Thomas, Judy Lieberman, and Ashish Lal. Desperately seeking microRNA targets. *Nature structural & molecular biology*, 17(10):1169–74, oct 2010.

[40] T M Witkos, E Koscianska, and W J Krzyzosiak. Practical Aspects of microRNA Target Prediction. *Current molecular medicine*, 11(2):93–109, mar 2011.

[41] Dong Yue, Hui Liu, and Yufei Huang. Survey of Computational Algorithms for MicroRNA Target Prediction. *Current genomics*, 10(7):478–92, nov 2009.

[42] Donny D Licatalosi, Aldo Mele, John J Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A Clark, Anthony C Schweitzer, John E Blume, Xuning Wang, Jennifer C Darnell, and Robert B Darnell. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–9, nov 2008.

[43] Sung Wook Chi, Julie B Zang, Aldo Mele, and Robert B Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–86, jul 2009.

[44] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–41, apr 2010.

[45] Julian König, Kathi Zarnack, Gregor Rot, Tomaz Curk, Melis Kayikci, Blaz Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*, 17(7):909–15, jul 2010.

[46] Aleksandra Helwak, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, 153(3):654–65, apr 2013.

[47] Michael J Moore, Troels K H Scheel, Joseph M Luna, Christopher Y Park, John J Fak, Eiko Nishiuchi, Charles M Rice, and Robert B Darnell. miRNA-target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity. *Nature communications*, 6:8864, jan 2015.

[48] Florian Erhard, Lars Dölken, Lukasz Jaskiewicz, and Ralf Zimmer. PARma: identification of microRNA target sites in AGO-PAR-CLIP data. *Genome biology*, 14(7):R79, jan 2013.

[49] William H Majoros, Parawee Lekprasert, Neelanjan Mukherjee, Rebecca L Skalsky, David L Corcoran, Bryan R Cullen, and Uwe Ohler. MicroRNA target site identification by integrating sequence and binding information. *Nature methods*, 10(7):630–3, jul 2013.

[50] Mohsen Khorshid, Jean Hausser, Mihaela Zavolan, and Erik van Nimwegen. A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nature methods*, 10(3):253–5, mar 2013.

[51] Shobha Vasudevan, Yingchun Tong, and Joan a Steitz. Switching from repression to activation: microRNAs can up-regulate translation. *Science (New York, N.Y.)*, 318(5858):1931–4, dec 2007.

[52] Shobha Vasudevan and Joan A Steitz. AU-rich-element-mediated upregulation of translation by FXR1 and Argonaute 2. *Cell*, 128(6):1105–18, mar 2007.

[53] Ulf Andersson Ørom, Finn Cilius Nielsen, and Anders H Lund. MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. *Molecular cell*, 30(4):460–71, may 2008.

[54] Yvonne Tay, John Rinn, and Pier Paolo Pandolfi. The multilayered complexity of ceRNA crosstalk and competition. *Nature*, 505(7483):344–52, jan 2014.

[55] Lei Wang, Zhang-Yan Guo, Rui Zhang, Bo Xin, Rui Chen, Jing Zhao, Tao Wang, Wei-Hong Wen, Lin-Tao Jia, Li-Bo Yao, and An-Gang Yang. Pseudogene OCT4-pg4 functions as a natural microRNA sponge to regulate OCT4 expression by competing for miR-145 in hepatocellular carcinoma. *Carcinogenesis*, apr 2013.

[56] Jue Yang, Tong Li, Chao Gao, Xiaobo Lv, Kunmei Liu, Hui Song, Yingying Xing, and Tao Xi. FOXO1 3'UTR functions as a ceRNA in repressing the metastases of breast cancer cells via regulating miRNA activity. *FEBS letters*, 588(17):3218–24, aug 2014.

[57] M-h Lü, B Tang, S Zeng, C-j Hu, R Xie, Y-y Wu, S-m Wang, F-t He, and S-m Yang. Long noncoding RNA BC032469, a novel competing endogenous RNA, upregulates hTERT expression by sponging miR-1207-5p and promotes proliferation in gastric cancer. *Oncogene*, (August):1–11, nov 2015.

[58] Tian Xia, Shengcan Chen, Zhen Jiang, Yongfu Shao, Xiaoming Jiang, Peifei Li, Bingxiu Xiao, and Junming Guo. Long noncoding RNA FER1L4 suppresses cancer cell growth by acting as a competing endogenous RNA and regulating PTEN expression. *Scientific reports*, 5(4):13445, 2015.

[59] Ning Bao, Khar-Wai Lye, and M Kathryn Barton. MicroRNA binding sites in Arabidopsis class III HD-ZIP mRNAs are required for methylation of the template chromosome. *Developmental cell*, 7(5):653–62, dec 2004.

[60] Anna Carina Jungkamp, Marlon Stoeckius, Desirea Mecenas, Dominic Grün, Guido Mastrobuoni, Stefan Kempa, and Nikolaus Rajewsky. In vivo and transcriptome-wide identification of RNA binding protein target sites. *Molecular Cell*, 44:828–840, 2011.

[61] Tina Glisovic, Jennifer L Bachorik, Jeongsik Yong, and Gideon Dreyfuss. RNA-binding proteins and post-transcriptional gene regulation. *FEBS letters*, 582(14):1977–86, jun 2008.

[62] Michaela Müller-McNicoll and Karla M Neugebauer. How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nature reviews. Genetics*, 14(4):275–87, apr 2013.

[63] Stefanie Gerstberger, Markus Hafner, and Thomas Tuschl. A census of human RNA-binding proteins. *Nature reviews. Genetics*, 15(12):829–45, dec 2014.

[64] S J Klug and M Famulok. All you wanted to know about SELEX. *Molecular biology reports*, 20(2):97–107, jan 1994.

[65] C G Burd and G Dreyfuss. Conserved structures and diversity of functions of RNA-binding proteins. *Science (New York, N.Y.)*, 265(5172):615–21, jul 1994.

[66] Somashe Niranjanakumari, Erika Lasda, Robert Brazas, and Mariano A Garcia-Blanco. Reversible cross-linking combined with immunoprecipitation to study RNA-protein interactions in vivo. *Methods (San Diego, Calif.)*, 26(2):182–90, feb 2002.

[67] Alfredo Castello, Bernd Fischer, Katrin Eichelbaum, Rastislav Horos, Benedikt M. Beckmann, Claudia Strein, Norman E. Davey, David T. Humphreys, Thomas Preiss, Lars M. Steinmetz, Jeroen Krijgsveld, and Matthias W. Hentze. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, 149(6):1393–406, jun 2012.

[68] Alexander G. Baltz, Mathias Munschauer, Björn Schwanhäusser, Alexandra Vasile, Yasuhiro Murakawa, Markus Schueler, Noah Youngs, Duncan Penfold-Brown, Kevin Drew, Miha Milek, Emanuel Wyler, Richard Bonneau, Matthias Selbach, Christoph Dieterich, and Markus Landthaler. The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Molecular cell*, 46(5):674–90, jun 2012.

[69] Sarah F Mitchell, Saumya Jain, Meipei She, and Roy Parker. Global analysis of yeast mRNPs. *Nature structural & molecular biology*, 20(1):127–33, jan 2013.

[70] Manuel Ascano, Markus Hafner, Pavol Cekan, Stefanie Gerstberger, and Thomas Tuschl. Identification of RNA-protein interaction networks using PAR-CLIP. *Wiley interdisciplinary reviews. RNA*, 3(2):159–77, mar 2012.

[71] Charles Danan, Sudhir Manickavel, and Markus Hafner. PAR-CLIP: A Method for Transcriptome-Wide Identification of RNA Binding Protein Interaction Sites. *Methods in molecular biology (Clifton, N.J.)*, 1358(3):153–73, feb 2016.

[72] Debashish Ray, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, Hong Na, Manuel Irimia, Leah H Matzat, Ryan K Dale, Sarah A Smith, Christopher A Yarosh, Seth M Kelly, Behnam Nabet, Desirea Mecenas, Weimin Li, Rakesh S Laishram, Mei Qiao, Howard D Lipshitz, Fabio Piano, Anita H Corbett, Russ P Carstens, Brendan J Frey, Richard A Anderson, Kristen W Lynch, Luiz O F Penalva, Elissa P Lei, Andrew G Fraser, Benjamin J Blencowe, Quaid D Morris, and Timothy R Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–7, jul 2013.

[73] Sarah F Mitchell and Roy Parker. Principles and properties of eukaryotic mRNPs. *Molecular cell*, 54(4):547–58, may 2014.

[74] Min-Sik Kim, Sneha M Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S Manda, Raghothama Chaerkady, Anil K Madugundu, Dhanashree S Kelkar, Ruth Isserlin, Shobhit Jain, Joji K Thomas, Babylakshmi Muthusamy, Pamela Leal-Rojas, Praveen Kumar, Nandini A Sahasrabuddhe, Lavanya Balakrishnan, Jayshree Advani, Bijesh George, Santosh Renuse, Lakshmi Dhevi N Selvan, Arun H Patil, Vishalakshi Nanjappa, Aneesha Radhakrishnan, Samarjeet Prasad, Tejaswini Subbannayya, Rajesh Raju, Manish Kumar, Sreelakshmi K Sreenivasamurthy, Arivusudar Marimuthu, Gajanan J Sathe, Sandip Chavan, Keshava K Datta, Yashwanth Subbannayya, Apeksha Sahu, Soujanya D Yelamanchi, Savita Jayaram, Pavithra Rajagopalan, Jyoti Sharma, Krishna R Murthy, Nazia Syed, Renu Goel, Aafaque A Khan, Sartaj Ahmad, Gourav Dey, Keshav Mudgal, Aditi Chatterjee, Tai-Chung Huang, Jun Zhong, Xinyan Wu, Patrick G Shaw, Donald Freed, Muhammad S Zahari, Kanchan K Mukherjee, Subramanian Shankar, Anita Mahadevan, Henry Lam, Christopher J Mitchell, Susarla Krishna Shankar, Parthasarathy Satishchandra, John T Schroeder, Ravi Sirdeshmukh, Anirban Maitra, Steven D Leach, Charles G Drake, Marc K Halushka, T S Keshava Prasad, Ralph H Hruban, Candace L Kerr, Gary D Bader, Christine A Iacobuzio-Donahue, Harsha Gowda, and Akhilesh Pandey. A draft map of the human proteome. *Nature*, 509(7502):575–81, may 2014.

[75] Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, Toby Mathieson, Simone Lemeer, Karsten Schnatbaum, Ulf Reimer, Holger Wenschuh, Martin Mollenhauer, Julia Slotta-Huspenina, Joos-Hendrik Boese, Marcus Bantscheff, Anja Gerstmair, Franz Faerber, and Bernhard Kuster. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–7, may 2014.

[76] Bryan T MacDonald, Keiko Tamai, and Xi He. Wnt/beta-catenin signaling: components, mechanisms, and diseases. *Developmental cell*, 17(1):9–26, jul 2009.

[77] Julia M Gohlke, Reuben Thomas, Yonqing Zhang, Michael C Rosenstein, Allan P Davis, Cynthia Murphy, Kevin G Becker, Carolyn J Mattingly, and Christopher J Portier. Genetic and environmental pathways to complex diseases. *BMC systems biology*, 3:46, jan 2009.

[78] Tomas Klingström and Dariusz Plewczynski. Protein-protein interaction and pathway databases, a graphical review. *Briefings in bioinformatics*, sep 2010.

[79] M. Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, jan 2000.

[80] David Croft, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, Steven Jupe, Irina Kalatskaya, Shahana Mahajan, Bruce May, Nelson Ndegwa, Esther Schmidt, Veronica Shamovsky, Christina Yung, Ewan Birney, Henning Hermjakob, Peter D'Eustachio, and Lincoln Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(Database issue):D691–7, jan 2011.

[81] Thomas Kelder, Martijn P. Van Iersel, Kristina Hanspers, Martina Kutmon, Bruce R. Conklin, Chris T. Evelo, and Alexander R. Pico. WikiPathways: Building research communities on biological pathways. *Nucleic Acids Research*, 40(D1):1301–1307, 2012.

[82] Saikat Chowdhury and Ram Rup Sarkar. Comparison of human cell signaling pathway databases–evolution, drawbacks and challenges. *Database : the journal of biological databases and curation*, 2015, jan 2015.

[83] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–9, may 2000.

[84] Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic acids research*, 43(Database issue):D1049–56, jan 2015.

[85] Bogumil M. Konopka. Biomedical ontologies—A review. *Biocybernetics and Biomedical Engineering*, 35(2):75–86, 2015.

[86] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–5, nov 2007.

[87] Richard Côté, Florian Reisinger, Lennart Martens, Harald Barsnes, Juan Antonio Vizcaino, and Henning Hermjakob. The Ontology Lookup Service: bigger and better. *Nucleic acids research*, 38(Web Server issue):W155–60, jul 2010.

[88] Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, Christopher G Chute, and Mark A Musen. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(Web Server issue):W170–3, jul 2009.

[89] John D Osborne, Jared Flatow, Michelle Holko, Simon M Lin, Warren a Kibbe, Lihua Julie Zhu, Maria I Danila, Gang Feng, and Rex L Chisholm. Annotating the human genome with Disease Ontology. *BMC genomics*, 10 Suppl 1:S6, jan 2009.

[90] Marion Gremse, Antje Chang, Ida Schomburg, Andreas Grote, Maurice Scheer, Christian Ebeling, and Dietmar Schomburg. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic acids research*, 39(Database issue):D507–13, jan 2011.

[91] Ryan R Brinkman, Mélanie Courtot, Dirk Derom, Jennifer M Fostel, Yongqun He, Phillip Lord, James Malone, Helen Parkinson, Bjoern Peters, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Larisa N Soldatova, Christian J Stoeckert, Jessica A Turner, and Jie Zheng. Modeling biomedical experimental processes with OBI. *Journal of biomedical semantics*, 1 Suppl 1:S7, jan 2010.

[92] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, jan 2009.

[93] Isabelle Rivals, Léon Personnaz, Lieng Taing, and Marie-Claude Potier. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics (Oxford, England)*, 23(4):401–7, feb 2007.

[94] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael a Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50, oct 2005.

[95] Sebastian Bauer, Julien Gagneur, and Peter N. Robinson. Going Bayesian: Model-based gene set analysis of genome-scale data. *Nucleic Acids Research*, 38(11):3523–3532, 2010.

[96] Qinghua Cui, Zhenbao Yu, Enrico O Purisima, and Edwin Wang. Principles of microRNA regulation of a human cellular signaling network. *Molecular systems biology*, 2:46, jan 2006.

[97] John Tsang, Jun Zhu, and Alexander van Oudenaarden. MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Molecular cell*, 26(5):753–67, jun 2007.

[98] Masafumi Inui, Graziano Martello, and Stefano Piccolo. MicroRNA control of signal transduction. *Nature reviews. Molecular cell biology*, 11(4):252–63, apr 2010.

[99] Henriett Butz, Károly Rácz, László Hunyady, and Attila Patócs. Crosstalk between TGF-β signaling and the microRNA machinery. *Trends in pharmacological sciences*, 33(7):382–93, jul 2012.

[100] Nastaran Mohammadi Ghahhari and Sadegh Babashah. Interplay between microRNAs and WNT/β-catenin signalling pathway regulates epithelial-mesenchymal transition in cancer. *European journal of cancer (Oxford, England : 1990)*, 51(12):1638–49, aug 2015.

[101] Margaret S Ebert and Phillip A Sharp. Roles for microRNAs in conferring robustness to biological processes. *Cell*, 149(3):515–24, may 2012.

[102] Daehyun Baek, Judit Villén, Chanseok Shin, Fernando D Camargo, Steven P Gygi, and David P Bartel. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, sep 2008.

[103] Hossein Zare, Arkady Khodursky, and Vittorio Sartorelli. An evolutionarily biased distribution of miRNA sites toward regulatory genes with high promoter-driven intrinsic transcriptional noise. *BMC evolutionary biology*, 14(1):74, jan 2014.

[104] Thomas Bleazard, Janine A Lamb, and Sam Griffiths-Jones. Bias in microRNA functional enrichment analysis. *Bioinformatics (Oxford, England)*, 31(10):1592–8, may 2015.

[105] Igor Ulitsky, Louise C Laurent, and Ron Shamir. Towards computational prediction of microRNA function and activity. *Nucleic acids research*, 38(15):e160, aug 2010.

[106] Zhengqing Ouyang, Qing Zhou, and Wing Hung Wong. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21521–21526, 2009.

[107] Claudia Angelini and Valerio Costa. Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data : statistical solutions to biological problems Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data : statistical solutions to biological proble. 2(May):1–8, 2014.

[108] Mark B Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, Renqiang Min, Pedro Alves, Alexej Abyzov, Nick Addleman, Nitin Bhardwaj, Alan P Boyle, Philip Cayting, Alexandra Charos, David Z Chen, Yong Cheng, Declan Clarke, Catharine Eastman, Ghia Euskirchen, Seth Frietze, Yao Fu, Jason Gertz, Fabian Grubert, Arif Harmanci, Preti Jain, Maya Kasowski, Phil Lacroute, Jing Leng, Jin Lian, Hannah Monahan, Henriette O'Geen, Zhengqing Ouyang, E Christopher Partridge, Dorrelyn Patacsil, Florencia Pauli, Debasish Raha, Lucia Ramirez, Timothy E Reddy, Brian Reed, Minyi Shi, Teri Slifer, Jing Wang, Linfeng Wu, Xinqiong Yang, Kevin Y Yip, Gili Zilberman-Schapira, Serafim Batzoglou, Arend Sidow, Peggy J Farnham, Richard M Myers, Sherman M Weissman, and Michael Snyder. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, sep 2012.

[109] Daogang Guan, Jiaofang Shao, Youping Deng, Panwen Wang, Zhongying Zhao, Yan Liang, Junwen Wang, and Bin Yan. CMGRN: A web server for constructing multilevel gene regulatory networks using ChIP-seq and gene expression data. *Bioinformatics*, 30(8):1190–1192, 2014.

[110] Christa Buecker, Rajini Srinivasan, Zhixiang Wu, Eliezer Calo, Dario Acampora, Tiago Faial, Antonio Simeone, Minjia Tan, Tomasz Swigut, and Joanna Wysocka. Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell stem cell*, 14(6):838–53, jun 2014.

[111] Daniel C. Factor, Olivia Corradin, Gabriel E. Zentner, Alina Saiakhova, Lingyun Song, Josh G. Chenoweth, Ronald D. McKay, Gregory E. Crawford, Peter C. Scacheri, and Paul J. Tesar. Epigenomic Comparison Reveals Activation of "Seed" Enhancers during Transition from Naive to Primed Pluripotency. *Cell stem cell*, 14(6):854–63, jun 2014.

[112] Alexandre G de Brevern, Jean-Philippe Meyniel, Cécile Fairhead, Cécile Neuvéglise, and Alain Malpertuy. Trends in IT Innovation to Build a Next Generation Bioinformatics Solution to Manage and Analyse Biological Big Data Produced by NGS Technologies. *BioMed research international*, 2015:904541, 2015.

[113] Shicai Wang, Ioannis Pandis, Chao Wu, Sijin He, David Johnson, Ibrahim Emam, Florian Guitton, and Yike Guo. High dimensional biological data retrieval optimization with NoSQL technology. *BMC genomics*, 15 Suppl 8(Suppl 8):S3, 2014.

[114] Chad Vicknair, Michael Macias, Zhendong Zhao, Xiaofei Nan, Yixin Chen, and Dawn Wilkins. A comparison of a graph database and a relational database. In *Proceedings of the 48th Annual Southeast Regional Conference on - ACM SE '10*, page 1, New York, New York, USA, 2010. ACM Press.

[115] Salim Jouili and Valentin Vansteenberghe. An Empirical Comparison of Graph Databases. In *2013 International Conference on Social Computing*, pages 708–715. IEEE, sep 2013.

[116] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(Database issue):D940–6, jan 2012.

[117] W Scott Campbell, Jay Pedersen, James C. McClay, Praveen Rao, Dhundy Bastola, and James R. Campbell. An alternative database approach for management of SNOMED CT and improved patient data queries. *Journal of biomedical informatics*, 57:350–357, aug 2015.

[118] Dimitar Hristovski, Andrej Kastrin, Dejan Dinevski, and Thomas C Rindflesch. Constructing a Graph Database for Semantic Literature-Based Discovery. *Studies in health technology and informatics*, 216:1094, jan 2015.

[119] David Johnson, Anthony J Connor, Steve McKeever, Zhihui Wang, Thomas S Deisboeck, Tom Quaiser, and Eliezer Shochat. Semantically linking in silico cancer models. *Cancer informatics*, 13(Suppl 1):133–43, 2014.

[120] Ron Henkel, Olaf Wolkenhauer, and Dagmar Waltemath. Combining computational models, semantic annotations and simulation experiments in a graph database. *Database : the journal of biological databases and curation*, 2015(0):bau130–, jan 2015.

[121] Bálint Antal, Anatole Chessel, and Rafael E Carazo Salas. Mineotaur: a tool for high-content microscopy screen sharing and visual analytics. *Genome biology*, 16(1):283, 2015.

[122] Daniel Bottomly, Shannon K McWeeney, and Beth Wilmot. HitWalker2: visual analytics for precision medicine and beyond. *Bioinformatics (Oxford, England)*, dec 2015.

[123] Pablo Pareja-Tobes, Raquel Tobes, Marina Manrique, Eduardo Pareja, and Eduardo Pareja-Tobes. Bio4j: a high-performance cloud-enabled graph-based data platform. Technical report, mar 2015.

[124] Xinbin Dai, Jun Li, Tingsong Liu, and Patrick Xuechun Zhao. HRGRN: A Graph Search-Empowered Integrative Database of Arabidopsis Signaling Transduction, Metabolism and Gene Regulation Networks. *Plant & cell physiology*, 57(1):e12, dec 2015.

[125] Vincenzo Bonnici, Francesco Russo, Nicola Bombieri, Alfredo Pulvirenti, and Rosalba Giugno. Comprehensive reconstruction and visualization of non-coding regulatory networks in human. *Frontiers in bioengineering and biotechnology*, 2:69, jan 2014.

[126] Pål Saetrom, Bret S E Heale, Ola Snøve, Lars Aagaard, Jessica Alluin, and John J Rossi. Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic acids research*, 35(7):2333–42, jan 2007.

[127] Jennifer a. Broderick, William E. Salomon, Sean P. Ryder, Neil Aronin, and Phillip D. Zamore. Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing. *RNA*, pages 1858–1869, aug 2011.

[128] Andrew I Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith a Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, Michael P Cooke, John R Walker, and John B Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–7, apr 2004.

[129] Anette H H van Boxel-Dezaire, M R Sandhya Rani, and George R Stark. Complex modulation of cell type-specific signaling in response to type I interferons. *Immunity*, 25(3):361–72, sep 2006.

[130] Steffen Sass, Adriana Pitea, Kristian Unger, Julia Hess, Nikola S Mueller, and Fabian J Theis. MicroRNA-Target Network Inference and Local Network Enrichment Analysis Identify Two microRNA Clusters with Distinct Functions in Head and Neck Squamous Cell Carcinoma. *International journal of molecular sciences*, 16(12):30204–22, jan 2015.

[131] John S Tsang, Margaret S Ebert, and Alexander van Oudenaarden. Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures. *Molecular cell*, 38(1):140–53, apr 2010.

[132] Ulf Schmitz, Xin Lai, Felix Winter, Olaf Wolkenhauer, Julio Vera, and Shailendra K Gupta. Cooperative gene regulation by microRNA pairs and their identification using a computational workflow. *Nucleic acids research*, 42(12):7539–52, jul 2014.

[133] Edyta Koscianska, Tomasz M Witkos, Emilia Kozlowska, Marzena Wojciechowska, and Wlodzimierz J Krzyzosiak. Cooperation meets competition in microRNA-mediated DMPK transcript regulation. *Nucleic acids research*, 43(19):9500–18, oct 2015.

[134] Marijn Schouten, Silvina A Fratantoni, Chantal J Hubens, Sander R Piersma, Thang V Pham, Pascal Bielefeld, Rob A Voskuyl, Paul J Lucassen, Connie R Jimenez, and Carlos P Fitzsimons. MicroRNA-124 and -137 cooperativity controls caspase-3 activity through BCL2L13 in hippocampal neural stem cells. *Scientific reports*, 5:12448, jan 2015.

[135] Patricia A Thibault, Adam Huys, Yalena Amador-Cañizares, Julie E Gailius, Dayna E Pinel, and Joyce A Wilson. Regulation of Hepatitis C Virus Genome Replication by Xrn1 and MicroRNA-122 Binding to Individual Sites in the 5' Untranslated Region. *Journal of virology*, 89(12):6294–311, jun 2015.

[136] Ioannis S Vlachos, Konstantinos Zagganas, Maria D Paraskevopoulou, Georgios Georgakilas, Dimitra Karagkouni, Thanasis Vergoulis, Theodore Dalamagas, and Artemis G Hatzigeorgiou. DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic acids research*, 43(W1):W460–6, jul 2015.

[137] Sooyoung Cho, Insu Jang, Yukyung Jun, Suhyeon Yoon, Minjeong Ko, Yeajee Kwon, Ikjung Choi, Hyeshik Chang, Daeun Ryu, Byungwook Lee, V Narry Kim, Wankyu Kim, and Sanghyuk Lee. MiRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting. *Nucleic acids research*, 41(Database issue):D252–7, jan 2013.

[138] Robert Petryszak, Tony Burdett, Benedetto Fiorelli, Nuno a Fonseca, Mar Gonzalez-Porta, Emma Hastings, Wolfgang Huber, Simon Jupp, Maria Keays, Nataliya Kryvych, Julie McMurry, John C Marioni, James Malone, Karine Megy, Gabriella Rustici, Amy Y Tang, Jan Taubert, Eleanor Williams, Oliver Mannion, Helen E Parkinson, and Alvis Brazma. Expression Atlas update–a database of gene and

transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic acids research*, 42(1):D926–32, jan 2014.

[139] Anders Jacobsen, Jiayu Wen, Debora S Marks, and Anders Krogh. Signatures of RNA binding proteins globally coupled to effective microRNA target sites. *Genome research*, 20(8):1010–9, aug 2010.

[140] Peng Jiang, Mona Singh, and Hilary a Coller. Computational assessment of the cooperativity between RNA binding proteins and MicroRNAs in Transcript Decay. *PLoS computational biology*, 9(5):e1003075, may 2013.

[141] Hengyuan Pang, Yongri Zheng, Yan Zhao, Xiaoqing Xiu, and Jianjiao Wang. miR-590-3p suppresses cancer cell migration, invasion and epithelial-mesenchymal transition in glioblastoma multiforme by targeting ZEB1 and ZEB2. *Biochemical and biophysical research communications*, 468(4):739–45, dec 2015.

[142] Yanxia Chu, Yunwei Ouyang, Fei Wang, Ai Zheng, Liping Bai, Ling Han, Yali Chen, and Hui Wang. MicroRNA-590 promotes cervical cancer cell growth and invasion by targeting CHL1. *Journal of cellular biochemistry*, 115(5):847–53, may 2014.

[143] Tianming Liu, Fang Nie, Xianggui Yang, Xiaoyan Wang, Yue Yuan, Zhongshi Lv, Li Zhou, Rui Peng, Dongsheng Ni, Yuping Gu, Qin Zhou, and Yaguang Weng. MicroRNA-590 is an EMT-suppressive microRNA involved in the TGFβ signaling pathway. *Molecular medicine reports*, 12(5):7403–11, nov 2015.

[144] Saber HafezQorani, Atefeh Lafzi, Ruben G de Bruin, Anton Jan van Zonneveld, Eric P van der Veer, Yeşim Aydın Son, and Hilal Kazan. Modeling the combined effect of RNA-binding proteins and microRNAs in post-transcriptional regulation. *Nucleic acids research*, feb 2016.

[145] Yoonseo Kim, Nicole Noren Hooten, Douglas F Dluzen, Jennifer L Martindale, Myriam Gorospe, and Michele K Evans. Posttranscriptional Regulation of the Inflammatory Marker C-Reactive Protein by the RNA-Binding Protein HuR and MicroRNA 637. *Molecular and cellular biology*, 35(24):4212–21, dec 2015.

[146] Ania Wilczynska, Anna Git, Joanna Argasinska, Eulàlia Belloc, and Nancy Standart. CPEB and miR-15/16 Co-Regulate Translation of Cyclin E1 mRNA during Xenopus Oocyte Maturation. *PloS one*, 11(2):e0146792, jan 2016.

[147] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra,

Arthur L Beaudet, Joseph R Ecker, Peggy J Farnham, Martin Hirst, Eric S Lander, Tarjei S Mikkelsen, and James A Thomson. The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*, 28(10):1045–8, oct 2010.

[148] Gonçalo R Abecasis, David Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard A Gibbs, Matt E Hurles, and Gil A McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, oct 2010.

[149] Christian Theil Have and Lars Juhl Jensen. Are graph databases ready for bioinformatics? *Bioinformatics (Oxford, England)*, 29(24):3107–3108, oct 2013.

[150] Lincoln D Stein. Integrating biological databases. *Nature reviews. Genetics*, 4(5):337–45, may 2003.

[151] Carole Goble and Robert Stevens. State of the nation in data integration for bioinformatics. *Journal of biomedical informatics*, 41(5):687–93, oct 2008.

[152] Kim D Pruitt, Jennifer Harrow, Rachel a Harte, Craig Wallin, Mark Diekhans, Donna R Maglott, Steve Searle, Catherine M Farrell, Jane E Loveland, Barbara J Ruef, Elizabeth Hart, Marie-Marthe Suner, Melissa J Landrum, Bronwen Aken, Sarah Ayling, Robert Baertsch, Julio Fernandez-Banet, Joshua L Cherry, Val Curwen, Michael Dicuccio, Manolis Kellis, Jennifer Lee, Michael F Lin, Michael Schuster, Andrew Shkeda, Clara Amid, Garth Brown, Oksana Dukhanina, Adam Frankish, Jennifer Hart, Bonnie L Maidak, Jonathan Mudge, Michael R Murphy, Terence Murphy, Jeena Rajan, Bhanu Rajput, Lillian D Riddick, Catherine Snow, Charles Steward, David Webb, Janet a Weber, Laurens Wilming, Wenyu Wu, Ewan Birney, David Haussler, Tim Hubbard, James Ostell, Richard Durbin, and David Lipman. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome research*, 19(7):1316–23, jul 2009.

[153] Roger S Day, Kevin K McDade, Uma R Chandran, Alex Lisovich, Thomas P Conrads, Brian L Hood, V S Kumar Kolli, David Kirchner, Traci Litzi, and G Larry Maxwell. Identifier mapping performance for integrating transcriptomics and proteomics experimental results. *BMC bioinformatics*, 12(1):213, jan 2011.

[154] David Gomez-Cabrero, Imad Abugessaisa, Dieter Maier, Andrew Teschendorff, Matthias Merkenschlager, Andreas Gisel, Esteban Ballestar, Erik Bongcam-Rudloff, Ana Conesa, and Jesper Tegnér. Data integration in the era of omics: current and future challenges. *BMC systems biology*, 8 Suppl 2(Suppl 2):I1, jan 2014.

[155] Sydney Brenner. Sequences and consequences. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1537):207–212, 2010.

Appendix A

Accepted Publications

The publications described in Chapter 2 are attached in the same order.

A.1 Publication 1

Rinck A*, **Preusse M***, Laggerbauer B, Lickert H, Engelhardt S, Theis FJ. *The human transcriptome is enriched for miRNA-binding sites located in cooperativity-permitting distance.* RNA Biol. 2013;10(7):1125–35.

RNA Biology 10:7, 1125–1135; July 2013; © 2013 Landes Bioscience

The human transcriptome is enriched for miRNA-binding sites located in cooperativity-permitting distance

Andrea Rinck,[1,2,†] Martin Preusse,[1,3,†] Bernhard Laggerbauer,[2] Heiko Lickert,[3,4] Stefan Engelhardt,[2,5,*] and Fabian J. Theis[1,6,*]

[1]Institute of Computational Biology; Helmholtz Zentrum München; Neuherberg, Germany; [2]Institute of Pharmacology and Toxicology; Technische Universität München; Munich, Germany; [3]Institute of Diabetes and Regeneration Research; Helmholtz Zentrum München; Neuherberg, Germany, Germany; [4]Institute of Stem Cell Research; Helmholtz Zentrum München; Neuherberg, Germany; [5]DZHK (German Center for Cardiovascular Research); partner site Munich Heart Alliance; Munich, Germany; [6]Technische Universität München; Garching, Germany

[†]These authors contributed equally to this work.

Keywords: microRNA, target regulation, target prediction, cooperativity

Abbreviations: miRNA, microRNA; miRISC, miRNA-induced silencing complex; AGO, Argonaute protein; 3'-UTR, 3' untranslated region; siRNAs, short interfering RNAs; iPSCs, induced pluripotent stem cells; RBP, RNA-binding protein

MiRNAs are short, non-coding RNAs that regulate gene expression post-transcriptionally through specific binding to mRNA. Deregulation of miRNAs is associated with various diseases and interference with miRNA function has proven therapeutic potential. Most mRNAs are thought to be regulated by multiple miRNAs and there is some evidence that such joint activity is enhanced if a short distance between sites allows for cooperative binding. Until now, however, the concept of cooperativity among miRNAs has not been addressed in a transcriptome-wide approach. Here, we computationally screened human mRNAs for distances between miRNA binding sites that are expected to promote cooperativity. We find that sites with a maximal spacing of 26 nucleotides are enriched for naturally occurring miRNAs compared with control sequences. Furthermore, miRNAs with similar characteristics as indicated by either co-expression within a specific tissue or co-regulation in a disease context are predicted to target a higher number of mRNAs cooperatively than unrelated miRNAs. These bioinformatic data were compared with genome-wide sets of biochemically validated miRNA targets derived by Argonaute crosslinking and immunoprecipitation (HITS-CLIP and PAR-CLIP). To ease further research into combined and cooperative miRNA function, we developed *miRco*, a database connecting miRNAs and respective targets involved in distance-defined cooperative regulation (mips.helmholtz-muenchen.de/mirco). In conclusion, our findings suggest that cooperativity of miRNA-target interaction is a widespread phenomenon that may play an important role in miRNA-mediated gene regulation.

Introduction

MicroRNAs (miRNAs) are small single-stranded non-coding RNAs, which are endogenously expressed and predominantly downregulate the expression of mRNA targets. They achieve post-transcriptional regulation of gene expression as part of the miRNA-induced silencing complex (miRISC), which consists of a miRNA and several proteins, including a member of the Argonaute (AGO) protein family. Binding of miRISC to its target sequence is guided by the miRNA and most commonly occurs within the 3'-untranslated region (3'-UTR) of the mRNA, thereby inducing translational repression or degradation of the mRNA (reviewed in refs. 1–3).

It becomes increasingly apparent that deregulated expression of miRNAs is causally related to the development of various complex disorders. This includes cardiac disease,[4,5] lung cancer,[6] leukemia,[7] neurological disorders such as Alzheimer disease,[8] metabolic abnormalities like diabetes mellitus,[9] and rheumatoid arthritis.[10]

Unbiased approaches to miRNA function, for example, by application of synthetic miRNA libraries to cells,[11] indicated that cellular pathways are regulated by multiple miRNAs[12,13] or are subject to regulation by a single miRNA acting on different levels.[14] On the other hand, almost every miRNA investigated has been assigned several, often contradictory, physiological roles.[15] Obviously, identifying the target mRNAs is crucial to understand the function of a disease-related miRNA and, consequently, to develop therapeutic approaches. To achieve this goal, we need to know the criteria according to which miRNAs (in the context of miRISCs) are guided to their respective targets and the principles leading to effective target regulation.

*Correspondence to: Stefan Engelhardt; Email: stefan.engelhardt@tum.de; Fabian J. Theis; Email: fabian.theis@helmholtz-muenchen.de
Submitted: 12/23/12; Revised: 05/03/13; Accepted: 05/06/13
http://dx.doi.org/10.4161/rna.24955

However, the mechanisms of miRNA-mRNA interactions are still about to be elucidated, and versatile, often contradictory modes of action have been reported.[16-19] In metazoans, the suppressive effect of an individual miRNA on a target is often small,[20] potentially due to the fact that miRNAs form only imperfect and thermodynamically unfavorable RNA-RNA hybrids with their targets over a short sequence (called the miRNA seed region nucleating the interaction).

A set of interaction rules has been formulated[1] based on biochemical and bioinformatic analyses, but functional miRNA sites often show aberrant characteristics. In spite of these difficulties, there is good evidence that contiguous and perfect base pairing of nucleotide positions 2–8 of the miRNA (seed region) with the cognate mRNA sequence is predictive of true interactions between them.[1,21]

Therefore, one comparably successful approach to bioinformatically predict miRNA targets is to focus on the seed region in miRNA targets. The online tool TargetScan searches for conserved seed regions of 7 and 8 nucleotides in length as well as for 3' compensated sites in 3'-UTRs. It ranks its predicted results based on further miRNA-mRNA binding properties summarized in a so called context+ score, including seed-pairing stability and target-site abundance.[22-24] A similar tool to predict miRNA target sites, miRanda, scores and ranks its results based on a machine learning algorithm called mirSVR.[25-27] The authors use support vector regression (SVR) to train on target site information as well as context features and calibrate their scores to correlate with observed downregulation of a published experimental data set.

More recently, computational methods were successfully combined with experimental miRISC-RNA crosslinking approaches to identify target mRNAs and characterize their miRNA binding sites: High-throughput sequencing of RNA isolated after UV crosslinking and immunoprecipitation (HITS-CLIP),[28] photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP)[29] and individual nucleotide resolution CLIP (iCLIP).[30] These approaches are helpful to reduce the search space for miRNA targets since they select for RNA fragments that are bound to active miRISC complexes. By UV irradiation of living cells, native protein-RNA contacts will be covalently crosslinked and, thereby, the information about the binding region preserved for later miRNA binding site predictions. Next to CLIP methods, there is a range of approaches used for target identification that do not rely on crosslinking, such as pull down of biotinylated miRNAs.[31]

However, miRNA-target interactions are not only bidirectional but rather form complex networks.[32,33] For the formation of a RISC on mRNA, seed pairing with as little as 6 or 7 nucleotides between miRNA and mRNA target seems sufficient (albeit thermodynamically unfavored and most likely dependent on further interaction between RISC components and the mRNA). Therefore, almost every miRNA known to date is computationally predicted to target more than one mRNA, and experimental evidence confirms this notion.[17,22-24,34-36] Further, one mRNA is often controlled by multiple miRNAs. It has been shown that mRNAs with strong miRNA-mediated effects on their expression level typically contain multiple miRNA binding sites for the same or different miRNAs instead of a single one, which is therefore useful as a predictor of miRNA target regulation.[35]

The possibility that miRNAs could regulate their targets in a concerted—potentially cooperative—fashion has already been considered short after their identification, when the 3'-UTR of *C. elegans* lin-41 mRNA was shown to contain multiple targets sites (seven) for miRNA lin-4.[37,38] Later, the integrity of more than one site for miRNA let-7 on the same target has been shown to be essential for efficient translational repression.[39] Additional support comes from assays which showed that luciferase reporter mRNAs with two or four binding sites for an exogenous small RNA (CXCR4 siRNA) in their 3'-UTR were more efficiently repressed than single-site constructs.[40]

In principle, the combined regulation of a mRNA by several miRNAs could be achieved by (1) independent or (2) cooperative target interaction (**Fig. 1A**). Independent binding of several miRNAs (in the context of a miRISC) to the same mRNA may be presumed to confer additive regulatory effects, whereas cooperative binding enhances the individual regulatory potency of miRNAs. Depending on their experimental design, assays for translational repression of reporter constructs verified both independent[23,41,42] and cooperative[42,43] activities of small RNAs. According to these studies, additive effects on the same mRNA are, at best, moderate, whereas regulation by two or more sites within a certain distance amplified miRNA-mediated repression to an extent greater than expected from independent sites, supporting a concept of cooperative activity.

However, only one of these studies[42] characterized cooperative RISC binding in a quantitative way. The authors used siRNAs (i.e., small interfering RNAs of 18–21 nucleotides, which completely hybridize to their targets) instead of miRNAs and multiple binding sites on one reporter mRNA molecule. The Hill coefficient, a measure of cooperativity,[44] was determined by fitting reporter repression as a function of the siRNA concentration to the Hill equation. Next to the identity of the involved Argonaute protein, Broderick et al.[42] showed strong dependency of cooperative silencing on the distance between two adjacent binding sites. They could show that at least for AGO 1, 3 and 4, miRNA cooperativity is limited to directly adjacent binding sites. This is in accordance with previous approaches studying the spacing pattern between neighboring binding sites leading to cooperative repression. Although the conclusions drawn in these studies were not fully unanimous, they concurred that cooperative effects are facilitated when miRNA binding sites are directly adjoining[40,42] (i.e., the distance from the 5'-end of one miRNA to the 5'-end of the next is 20–22 nucleotides or the length of exactly one miRNA). Enhanced repression of mRNAs with two (vs. one) miRNA binding sites was also observed when sites partially overlapped (5'-end of downstream site moved four nucleotides into the accessory but not the seed region of the upstream site) or when they were separated by few additional nucleotides (5'-to-5' distance of 25 nucleotides).[43] It seems, however, unclear what mode of combined miRNA activity (independent or cooperative) underlies translational repression of these constructs. Broderick et al.[42] found that cooperative effects are lost when miRNA sites (except bulged AGO 2 sites) were separated by 19 nucleotides

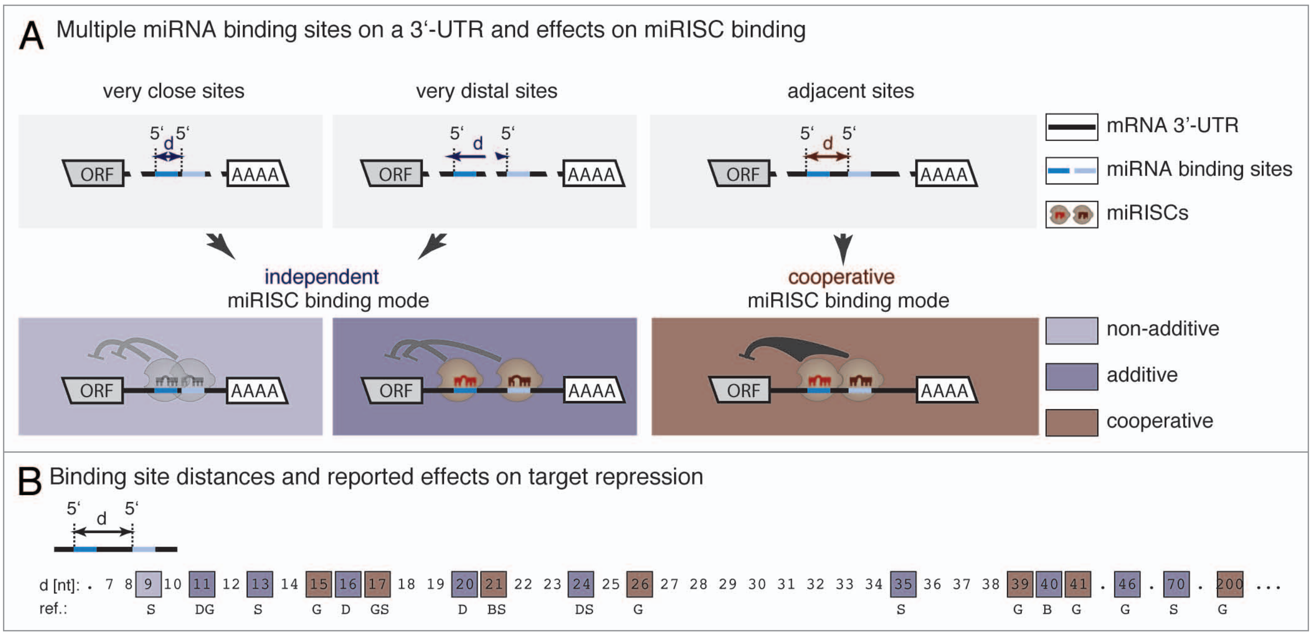

A Multiple miRNA binding sites on a 3'-UTR and effects on miRISC binding
very close sites
very distal sites
adjacent sites
5' d 5'
ORF
AAAA
mRNA 3'-UTR
miRNA binding sites
miRISCs
independent
miRISC binding mode
cooperative
miRISC binding mode
non-additive
additive
cooperative
B Binding site distances and reported effects on target repression
d [nt]: . 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 . 46 . 70 . 200 ...
ref.: S DG S G D GS D BS DS G S G B G G S G

Figure 1. Distance between multiple miRNA binding sites as predictor of cooperative target regulation. (**A**) Concerted miRNA target regulation (in the context of miRISCs) may be described by independent or cooperative activities. An independent mode of repression has been described for very close and for very distant sites. Non-additive effects would be expected if overly close sites exclude simultaneous binding of miRNAs. Additive effects may occur when miRNAs occupy sites autonomously without activity-enhancing interactions between their miRISCs. In contrast, cooperative activity has been shown for miRNAs, whose binding sites on a specific mRNA are within a certain distance referred to as cooperativity range. The term cooperativity refers to a synergistic effect amplifying miRNA-mediated repression to an extent greater than expected from independent sites. (**B**) Summary of previous experimental studies investigating distance-dependency of cooperative target repression by multiple small RNA binding sites. Inter-site distances (d) tested in the reports (squares) are shown as the distance between adjacent miRNA 5' ends on the respective mRNA target. Square colors, light and dark blue indicate that the repressive effect of multiple binding sites was weaker/similar or stronger than for a single target site. Brown squares emphasize 5'-to-5' distances for which repression has been reported significantly greater than expected from additive effects (cooperative miRISC binding mode). The accumulation of cooperative regulation for distances between 15–26 nucleotides indicates that directly adjacent miRISCs (with certain variations) have the highest potential to repress their target in a cooperative way. Cooperative regulation outside of the core cooperativity range might occur due to secondary structure formations of the target sequence.[23,41-43]

(5'-to-5' distance 40 nucleotides). These studies suggest that direct adjacency of binding sites promotes cooperative miRNA activities, whereas deviation from this rule may result in loss of combined effects or a shift toward independent activities. A schematic drawing of these correlations is shown in **Figure 1B**. Some outliers with larger 5'-to-5' distance of cooperative binding sites may be explained by spatial proximity due to suitable secondary structure of the mRNA sequence.

Mechanistically, distance constraints between miRNA binding sites have been suggested to result from interactions between adjacent RNA-induced silencing complexes which stabilize target mRNA binding and increase the probability of occupancy (binding cooperativity).[41,42] Another possibility would be a cooperative influence on the recruitment or effectiveness of further proteins leading to enhanced target degradation or repression (functional cooperativity[42]). Too close sites might be underrepresented due to steric hindrance of neighboring RISCs resulting in reduced effectiveness, possibly even lower than for a single site. On the other hand, if binding sites are too distant, the RNA-protein complexes might not be able to positively interact. However, it has to be elucidated if cooperative target regulation reflects a general concept of miRNA mediated mRNA regulation.

Here we present a systematic distance analysis of predicted miRNA target sites in human 3'-UTRs. Compared with randomized controls, distances shown by experimental studies to generate cooperative effects were enriched for naturally occurring miRNAs and miRNA binding sites. Further, functionally related miRNAs tend to bind more distance-defined cooperative targets, as the number increases for groups of miRNAs co-expressed in the same tissue or co-regulated in specific disease contexts. Our results, which are based on binding sites predicted by TargetScan are in good agreement with both a second computational target site predictor (miRanda/mirSVR) and experimentally verified miRNA interaction sites derived from HITS-CLIP or PAR-CLIP experiments.

Our findings support the importance of inter-site distance as a parameter defining miRNA-mediated repression. The comprehensive analysis of multiple miRNAs per target rather than miRNA-mRNA pairs appears essential to exploit disease-associated miRNAs and respective targets suitable for therapeutic purposes. To facilitate further research in miRNA cooperativity we developed *miRco*, a public web application that predicts cooperative miRNA-target interactions based on inter-site distance constrains: www.mips.helmholtz-muenchen.de/mirco.

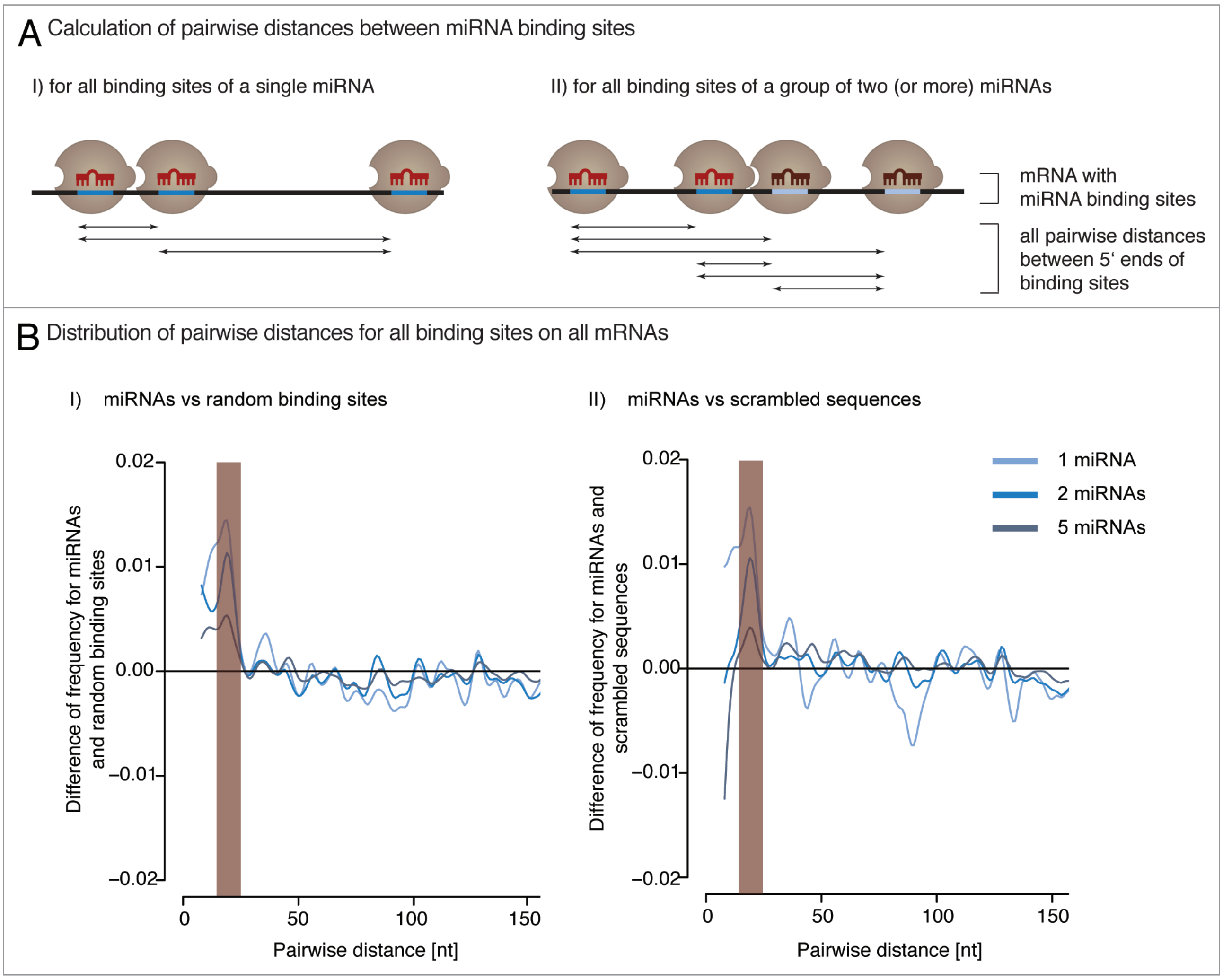


Figure 2. Naturally occurring miRNA binding sites are more frequently spaced within the cooperativity range (15–26 nucleotides) than expected by chance. (**A**) The pairwise distance between all binding sites of a single (1) or multiple (2) miRNAs is calculated for each mRNA 3'-UTR. (**B**) The distribution of pairwise distances shows an enrichment of the cooperativity permitting distance for miRNAs compared with randomly picked sites (1) and predicted binding sites of scrambled sequences (2) ($P < 2.2 \times 10^{-16}$, Wilcoxon Rank Sum test).

Results

MiRNA target sites are enriched within cooperativity-promoting distance. If a cooperative mode of action was functionally relevant, then cooperativity-promoting distance between target sites should be statistically overrepresented for intact binding sites of miRNAs. To test this, we computed the distribution of pairwise distances between predicted binding sites of evolutionary conserved human miRNAs. MiRNA targets were predicted using TargetScan, version 6.2.[45] The data set contained 1,537 conserved human miRNAs. We calculated the distribution of distances between all binding sites of each miRNA individually. In addition, we determined distances for all binding sites of groups of two and five miRNAs (**Fig. 2A**). These group sizes have been defined in order to analyze combinatory effects. They were sampled 1,000 times from the complete set of miRNAs.

To test for statistical significance, the results were compared with two different null models: (1) randomly chosen binding sites and (2) predicted target sites for scrambled miRNA-like sequences. For the first null model, we randomly selected target sites in a sequence-independent manner and, thus, generated artificial target sets with random binding positions. We picked random positions from the complete set of real human 3'-UTRs. The number of sites was normalized to predictions for human miRNAs. For the second, we designed arbitrary sequences of 22 nucleotides with the constraint that they are not similar to known miRNAs. We predicted targets with TargetScan and kept only those results that have a similar number of targets than native human miRNAs.

The distribution of all pairwise distances significantly differed for endogenous miRNAs and randomly selected binding sites in the range of 15–26 nucleotides (**Fig. 2B**, P value < 2.2×10^{-16}). This is in accordance with experimental findings (**Fig. 1B**) and is hence referred to as cooperativity range. This holds true for individual as well as for combinations of two and five different miRNAs. The enrichment of miRNA binding sites shows a peak for an inter-site distance of ~21 nucleotides (i.e., when two miRNAs bind in immediate vicinity). The distance distribution of predicted binding sites for scrambled sequences was also found to be different from miRNAs, again with significant underrepresentation within the cooperativity range (**Fig. 2B**).

In summary, when only small distances are considered (< 3 miRNA lengths), predictions for randomly picked sites and scrambled sequences produced similar results, while predicted target sites for human miRNAs displayed significant enrichment of inter-site spacing between 15–26 nucleotides. These findings correlate with previous studies (**Fig. 1B**) and, thus, we used this window of inter-site spacing in subsequent analyses to determine cooperatively regulated miRNA targets.

HITS-CLIP and PAR-CLIP data sets show cooperativity. We calculated the fraction of targets that are potentially regulated in a cooperative manner for four distinct sets of miRNA targets: (1) TargetScan predictions for human miRNAs,[44] (2) miRanda/mirSVR predictions for human miRNAs as a second target prediction tool,[26,27] (3) experimentally validated data from a HITS-CLIP[28] and (4) from a PAR-CLIP study.[29] All data sets were compared with random target sites and random sequences. As above, we analyzed single as well as groups of two and five miRNAs to take combined activity into account.

Looking at target prediction, both tools show significantly more cooperative targets than random binding sites and random sequences (p-values $< 2.2 \times 10^{-16}$). This holds true for single and groups of two and five miRNAs. Interestingly, in all cases, miRanda/mirSVR has a higher percentage of cooperative targets. We find a mean of 2%, 4%, and 8% for miRanda/mirSVR and a mean of 1%, 1.7%, and 2.2% for TargetScan.

The difference between miRNAs and controls increased with the number of miRNAs and we found the largest difference for groups of five naturally occurring miRNAs. This indicates that targets controlled by multiple different miRNAs are more frequently regulated in a cooperative fashion than mRNAs with multiple binding sites for an identical miRNA species.

Recently, biochemical methods to identify miRNA binding sites on a genome-wide scale have been developed. For an experimental validation of our in silico results, we analyzed the published HITS-CLIP and PAR-CLIP data sets. The former contains mRNA binding sites for the 20 most abundant miRNAs from mouse brain while the latter contains 47 human miRNAs. Both HITS-CLIP and PAR-CLIP identifie similar numbers of targets as TargetScan and miRanda/mirSVR and, thus, allow for comparison with our findings for computational prediction.

We retrieved all binding sites for both data sets and calculated the proportion of cooperative targets (**Fig. 3**, brown and blue boxes). For a single miRNA, only HITS-CLIP shows significantly more cooperative targets than controls with a mean of 2.5% compared with 0.5% and 0.2% for random sequences and random sites. PAR-CLIP data shows a mean of 0.4% cooperative targets and thus is not different from controls.

This picture changed when groups of two or five miRNAs were taken into account. While both data sets retrieved by experimental methods exhibited the same tendency as the target prediction tools, both HITS-CLIP and PAR-CLIP showed a stronger gain. The mean fraction of cooperative targets increased to 4.7% and 12.5% for HITS-CLIP and to 3.4% and 5.7% for PAR-CLIP. The mean percentage of coperative targets for the controls increased only to 0.4%/0.8% for random positions and 1%/1.3% for random sequences. In general, miRanda/mirSVR resembled HITS-CLIP and PAR-CLIP more closely while for TargetScan the number of cooperative targets was slightly lower.

These results show that cooperative regulation is likely to involve different miRNAs. Most importantly, the data for computational target prediction was confirmed with two independent sets of experimentally validated miRNA targets.

Functionally related miRNAs show an enrichment of target sites within cooperativity permitting distance. As shown above, endogenous miRNAs are more likely to posess target sites in a cooperativity range than randomly picked sites or scrambled sequences. If cooperativity is relevant in miRNA-mediated gene regulation, then functionally related miRNAs may share more cooperative targets than others. As most miRNAs are not comprehensively understood with respect to function, the field widely relies on two criteria that may be indicative of functional relation: (1) Co-expression of miRNAs within a particular tissue and (2) co-regulation in a common disease context. To put the first criterion to the test, we used the miRNA expression profiling database mimiRNA.[46] For the second, we employed PhenomiR,[47] a database of differentially regulated miRNA expression in diseases. For all miRNAs in both databases, targets were retrieved from TargetScan as described before.

The mimiRNA database employs normalized human miRNA expression profiles from four different sources: Sequencing data from the miRNA Atlas,[48] quantitative real-time PCR data from Gaur et al.[49] and Lee et al.[50] and microarray and deep sequencing data from the Gene Expression Omnibus (GEO).[51] The complete data set for 188 different tissues was used to calculate the proportion of cooperative targets among all targets for single and groups of two and five miRNAs. Co-expressed miRNAs were compared with all non-expressed miRNAs as a control. As shown exemplary for brain, liver, heart and lung (**Fig. 4**), miRNAs that are co-expressed in a tissue target more mRNAs in a potentially cooperative manner than miRNAs that are not co-expressed in a particular tissue (one-sided Wilcoxon Rank Sum test with $P < 2.2 \times 10^{-16}$). Moreover, the difference increases for groups of two and five co-expressed miRNAs, suggesting that these co-expressed miRNAs are in a functional relation with each other.

To test for the second presumed indicator of functional interrelation, i.e. co-regulation in disease, we applied the latest release of PhenomiR, a database which comprises 126 diseases and 615 associated miRNAs. Again, the fraction of cooperative targets was determined for single and sampled combinations of two and five miRNAs. The complete set of non-regulated miRNAs was used as control for each disease. Targeting with at least two binding sites within the cooperativity range was more often found for co-regulated miRNAs than for control groups not associated with

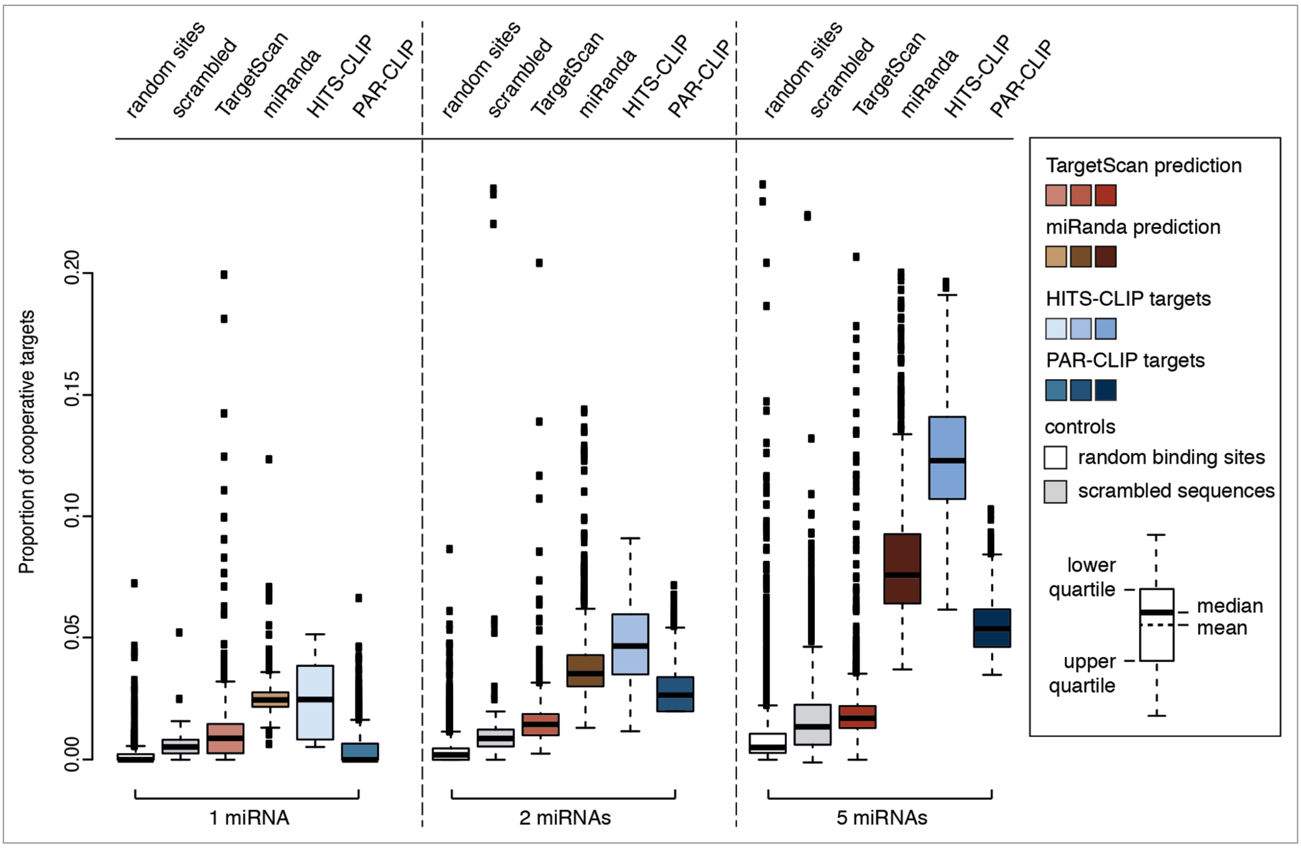


Figure 3. The fraction of cooperative targets per total targets grows for increasing numbers of miRNAs. Analysis of cooperative targets was performed with computationally predicted (TargetScan, red; miRanda, brown) and experimentally identified (HITS-CLIP, light blue; PAR-CLIP, dark blue) target sets. The proportion of cooperative targets is plotted for single miRNAs and sampled groups of two and five miRNAs. The mean is always higher for existing miRNAs than either of the controls ($P < 2.2 \times 10^{-16}$, tested with Wilcoxon Rank Sum test), except PAR-CLIP data for single miRNAs. This indicates that they are more often located in a potential cooperative distance than expected by chance.

a particular disease (one-sided Wilcoxon Rank Sum test with $P < 2.2 \times 10^{-16}$) (data not shown). Notably, this holds true for all diseases covered by PhenomiR. Similar to co-expressed miRNAs, we found an increase of the difference between disease-associated miRNAs and controls for groups of two and five miRNAs.

***miRco*: A tool to predict miRNA targets with binding sites in a cooperativity-permitting distance.** We have shown that miRNA binding sites are more often located in the cooperativity range than expected by chance. Additionally, functionally related miRNAs show an enrichment of cooperative binding sites. Still, the biological relevance of cooperativity in miRNA function has to be shown experimentally. To support further research in this topic, we developed the web application *miRco,* a tool to predict potentially cooperative miRNA interactions and their mRNA targets. Upon user input of miRNAs and distance allowance between miRNA binding sites, *miRco* identifies mRNAs that may be controlled by cooperative miRNA activities. Additionally, *miRco* can find all miRNAs that bind cooperatively to a given list of genes or mRNAs (**Fig. 5A and B**). To predict mRNAs that are cooperatively regulated, *miRco* searches by default for target sites within a distance of 15–26 nucleotides between two consecutive miRNA 5' ends. As described above, this setting was chosen based on our findings and reported experimental data.[23,41-43] Alternatively, the user may define a custom lower and upper limit of the distance. The tool includes miRNAs and mRNAs from human, mouse and rat. Target predictions are obtained from the current release of TargetScan (version 6.2).

First, the user is asked to choose the species for which the search is to be performed. Then, a list of miRNAs, or genes, or both may be submitted. If either miRNAs or genes are left blank, the complete data set is used for analysis. Our tool is connected to the PhenomiR database. The user can select a disease annotated in PhenomiR and input a set of disease-associated miRNAs (**Fig. 5B**). The output of *miRco* is presented as a list of target genes with corresponding binding sites in the aforementioned cooperativity range. Data are initially sorted based on the context+ score calculated by TargetScan and can subsequently be listed by target gene symbol and average distance between the binding sites. Furthermore, the result table can be filtered for the occurrence of one or multiple miRNAs within the list of candidate mRNAs.

The improved data set of the latest TargetScan release is a solid fundament for prediction of cooperative targets for three major model organisms used for medical research. *miRco* may serve as a hypothesis-generator to aid further research on the

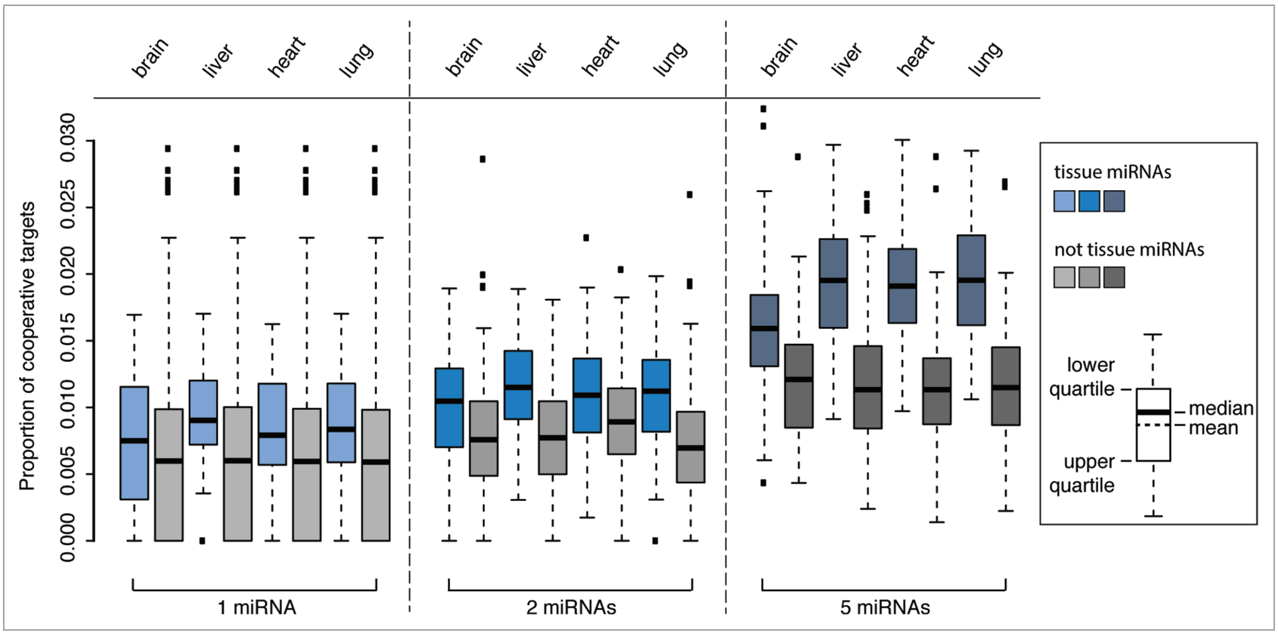


Figure 4. Functionally related miRNAs show an enrichment of target sites within the cooperativity range. Fraction of potentially cooperative targets for miRNAs in four exemplary tissues (blue) compared with a control set of miRNAs not expressed within the respective tissue (gray). Targeting within the cooperativity permitting distance is over-represented for co-expressed miRNAs (one-sided Wilcoxon Rank Sum test with $P < 2.2 \times 10^{-16}$).

mechanisms underlying concerted miRNA-mediated target regulation.

Discussion

The study presented here addresses a largely unresolved question in miRNA research: Do miRNAs confer physiological effects on their own, or do they function in a concerted, possibly cooperative manner?

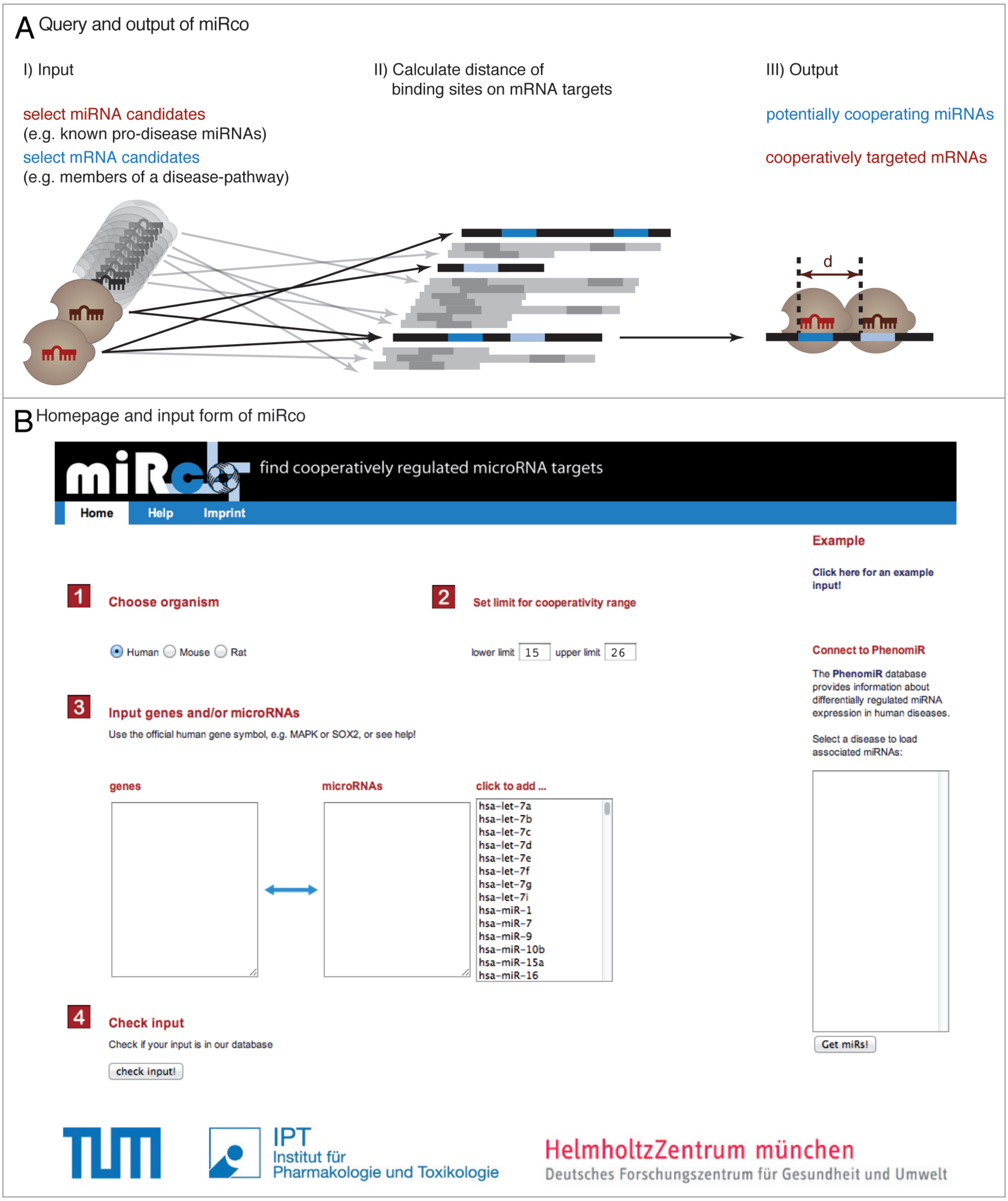
Literature provides certain evidence: Experiments in which a single small regulatory RNA binds to a single site within a mRNA often fail to show effects (e.g. refs. 23 and 40). On the other hand, studies indicated that miRNA-mediated target regulation is particularly effective if several miRNAs bind within a close distance.[23,41-43] However, these results rely on expression of artificial reporter constructs and do not provide comprehensive evidence that cooperativity is a general principle of miRNA-mediated target regulation.

In principle, one way to explain the basic concept of miRNA cooperativity is that proximity of binding sites on mRNAs stabilizes miRISCs' binding to their mRNA targets, leading to an increased silencing effect. This proximity concept has already been discussed in literature and several of our observations provide further support for it on a genome-wide scale: First, we showed that mRNAs with more than one miRNA site are more likely to have these sites placed in cooperativity-promoting distance (15–26 nucleotides, 5'-to-5') than randomized controls. Interestingly, the peak distance of ~21 nucleotides reflects binding of two miRNAs in direct neighborhood. Second, by in silico prediction (TargetScan, miRanda/mirSVR), as well as experimentally supported (HITS-CLIP, PAR-CLIP), we retrieved more mRNAs with miRNA sites in cooperativity range than from control setups. Third, the higher proportion of such mRNAs goes along with the co-regulation of miRNAs in tissue as well as similar disease context, underscoring the suspected functional interplay of these miRNAs on the respective mRNAs. The enrichment of miRNA binding sites in cooperativity-promoting distance speaks for a prevalently concerted, maybe cooperative way of miRNA target regulation.

However, the mechanisms of targeting are complex. Apparently, the type of Argonaute protein involved in a particular silencing complex has great influence on the nature of target regulation.[42] For example AGO 1 and AGO 2 show distinct characteristics with respect to the distance requirement between binding sites leading to cooperative targeting.[42] While bulged binding of miRNAs within AGO 1-complexes shows cooperativity only for adjacent binding sites, bulged sites of AGO 2-containing RICSs can act cooperatively in adjacent as well as in extended compositions. Consequently, the cell-specific proportion of the different AGO subtypes as well as the concentration of other potential effector proteins may be important. Furthermore, the sequence context around miRNA sites might affect cooperative actions of miRISCs, with other protein binding sites and secondary structure as the most likely determinants.

Therefore, the next step in studying miRNA cooperativity will be to more comprehensively analyze it in a biological context. Analysis of several instead of single miRNAs and their potential cooperativity could lead to a better understanding of the complex interplay of miRNAs and genetic networks in health and disease.

Cooperativity as a moderator of strongly increased effects would be interesting for the therapeutic use of miRNAs: If two miRNAs downregulate an mRNA target cooperatively, a lower


A Query and output of miRco
I) Input
select miRNA candidates
(e.g. known pro-disease miRNAs)
select mRNA candidates
(e.g. members of a disease-pathway)
II) Calculate distance of binding sites on mRNA targets
III) Output
potentially cooperating miRNAs
cooperatively targeted mRNAs
d
B Homepage and input form of miRco
miRco
Home
Help
Imprint
check input!
Get miRs!

find cooperatively regulated microRNA targets

Example

Click here for an example input!

1 Choose organism

● Human ○ Mouse ○ Rat

2 Set limit for cooperativity range

lower limit 15 upper limit 26

Connect to PhenomiR

The **PhenomiR** database provides information about differentially regulated miRNA expression in human diseases.

Select a disease to load associated miRNAs:

3 Input genes and/or microRNAs

Use the official human gene symbol, e.g. MAPK or SOX2, or see help!

genes

microRNAs

click to add ...

hsa-let-7a
hsa-let-7b
hsa-let-7c
hsa-let-7d
hsa-let-7e
hsa-let-7f
hsa-let-7g
hsa-let-7i
hsa-miR-1
hsa-miR-7
hsa-miR-9
hsa-miR-10b
hsa-miR-15a
hsa-miR-16

4 Check input

Check if your input is in our database

TUM

IPT Institut für Pharmakologie und Toxikologie

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Figure 5. Functionality of the *miRco* web application. (**A**) A search for cooperative miRNA-target interactions is performed by selecting miRNA candidates, relevant target genes or both. The user is able to specify parameters for the range in which the spacing between two adjacent miRISC binding sites (d) is assumed to lead to cooperative target repression. Default values are 15–26 nucleotides. Predictions for three species are available: human, mouse, and rat. (**B**) Screenshot of the user interface of the online tool.

level of expression might suffice to exert the designated effect. This would potentially decrease side effects of the miRNAs or miRNA mimics and, thereby, lead to a more tolerable treatment.

In addition, combinations of miRNAs could be employed to improve experimental protocols. Similar to the idea of decreased side effects of therapeutic miRNAs, the combination of different miRNAs might increase the specific effect on the targets of interest. Interestingly, the combined activity of multiple miRNAs has recently been reported to facilitate the reprogramming of fibroblasts to cardiomyocyte-like cells[52] as well as the induction of pluripotent stem cells (iPSCs).[53,54]

Recently, several studies highlighted the interaction of AGO/miRNAs with other RNA binding proteins (RBP).[55-57] In the future, the concept of cooperativity may extend to all RBPs in order to better predict mRNA regulation.

In the context of this work, we also developed *miRco* (mips.helmholtz-muenchen.de/mirco), a web application meant to aid experimental research into the cooperative action of miRNAs. It predicts potentially cooperatively targeted mRNAs based on binding site distances and, thus, might help to identify key regulatory miRNA-mRNA networks. *miRco* serves as a starting point for wet lab scientists: It allows one to input miRNAs and search for cooperative targets. In addition, the user can specify a set of genes and find all miRNAs that target these genes in a cooperative fashion. This dual approach helps to narrow down lists of candidate genes and miRNAs and makes it more feasible to test cooperativity in a complex biological context.

Taken together, our data indicate that cooperativity of miRNA-target interaction is a wide-spread phenomenon that may play an important role in miRNA-mediated gene regulation.

Materials and Methods

Criteria for the prediction of cooperativity. Cooperativity of two miRNAs is defined by the distance between the 5'-starts of their binding sites. We used 15 nucleotides as the lower and 26 nucleotides as the upper limit of the cooperative distance, following experimental studies of distance-dependent cooperative effects and our data showing an enrichment of binding sites for human miRNAs in this window.

To determine whether a mRNA may be cooperatively regulated, we take a single gene, acquire all binding sites of a given set of miRNAs on this mRNA and cluster them in groups where the distance between two adjacent sites lies within the cooperativity interval. If at least two binding sites fulfill this criterion, a mRNA is considered to be potentially regulated in a cooperative manner.

All data sets are stored in a MySQL database containing tables for genes, miRNAs and binding sites as well as their relations. Analyses are performed with Python programs combined with data plotting using R.

MiRNA target prediction. We used computational target prediction of human miRNAs from TargetScan[22,45] release 6.2 and miRanda/mirSVR release August 2010.[26,27] For TargetScan, we used the predictions for conserved miRNAs and targets. Scores of target sites are the context+ scores calculated by TargetScan. The release 6.2 contains 1,536 conserved human miRNAs and prediction is performed on a multiple sequence alignment of 18413 3'-UTRs from 23 species. For miRanda/mirSVR, we used the predictions for conserved miRNAs with a good mirSVR score. The release contains 249 human miRNAs.

Random distribution of target sites. Randomly distributed target positions were used as a null model for cooperativity. We picked random positions within the real set of human 3'-UTRs. UTR data for all human genes (assembly GRCh37.p10) was downloaded from ENSEMBL BioMart (http://www.ensembl.org/). The number of positions per UTR was normalized to lie within the range of TargetScan predictions. This approach is completely independent of miRNAs, their sequences and pairing determinants. Thus, this represents the most basic null model for binding site allocation and does not rely on any prior knowledge.

Random miRNA-like sequences. To augment the basic random position control, we generated 1000 completely random 22 nucleotides long sequences. We only used sequences which are not known human miRNAs and do not contain seeds (nucleotide 2–8) of known human miRNAs. We predicted targets for these seeds with the TargetScan 6.2 software and the UTR data provided by TargetScan. For the subsequent analyses, only random sequences that produce the same numbers of targets (i.e., between 10–2,719) as human miRNAs were taken into account.

Sampling of groups. For analyses using single miRNAs, the complete data set was considered. Groups of two and five miRNAs or controls were sampled randomly 1000 times from the complete set with no recurrence.

HITS-CLIP data set. The data set of Chi et al.[28] is available at ago.rockefeller.edu, including mapping of miRNA binding sites onto genomic positions. The authors of this study used neocortex of P13 mouse brain, crosslinked RNA binding proteins and RNA with UV irradiation and immunoprecipitated AGO-RNA complexes. Subsequently, RNA was purified and sequenced. Computational analysis produced a miRNA-mRNA interaction map. We used the mapping on mouse genome assembly mm9.

PAR-CLIP data set. The data set of Hafner et al.[29] is available through starBase (starbase.sysu.edu.cn), a database providing gene mappings for a wide range of CLIP experiments.[58] We used the "target site interaction" tool of starBase with settings for at least one microRNA read and "stringent miRNA targets" as described in the starBase publication.

Statistics. The distributions of pairwise distances (in a given distance window) as well as the percentage of cooperative targets were tested for a significant difference between miRNAs and controls with a one-sided Wilcoxon Rank Sum test.[59] We used the wilcox.test function in the "stats" package of the R statistical computing software with a confidence interval of 0.95 to calculate P values. P values $< 2.2 \times 10^{-16}$ occur due to the limits in floating point precision in R.

***miRco* web application.** The *miRco* web tool is implemented as a JAVA EE application running on a Tomcat 6 servlet engine, using the same MySQL database described above. It employs TargetScan release 6.2. For a given set of mRNAs and miRNAs, *miRco* produces groups of miRNA binding sites that fulfill the user set distance criteria. Whenever two binding sites are at the exact same position or overlap (i.e., their distance is smaller than the lower limit), the binding site with the best context+ score calculated by TargetScan is used.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We would like to thank Nikola Müller, Carsten Marr and Dominik Lutter for helpful comments and proofreading of the manuscript. This work was supported by the Bundesministerium für Bildung und Forschung (MedSys project "LungSys" and programmes m4, Personalized Medicine and ANR-BMBF-01KU0902A). This work was further supported in part by grants from the Fondation Leducq (to SE); the Bavarian Ministry of Sciences, Research and the Arts in the framework of the Bavarian Molecular Biosystems Research Network (to SE and FT), the Helmholtz Alliance on Systems Biology (project "CoReNe"), the German Science Foundation (DFG) within the SPP 1395 (InKoMBio) and the Studienstiftung des Deutschen Volkes (to AR).

References

1. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell 2009; 136:215-33; PMID:19167326; http://dx.doi.org/10.1016/j.cell.2009.01.002
2. Brodersen P, Voinnet O. Revisiting the principles of microRNA target recognition and mode of action. Nat Rev Mol Cell Biol 2009; 10:141-8; PMID:19145236; http://dx.doi.org/10.1038/nrm2619
3. Fabian MR, Sonenberg N, Filipowicz W. Regulation of mRNA translation and stability by microRNAs. Annu Rev Biochem 2010; 79:351-79; PMID:20533884; http://dx.doi.org/10.1146/annurev-biochem-060308-103103
4. van Rooij E, Sutherland LB, Liu N, Williams AH, McAnally J, Gerard RD, et al. A signature pattern of stress-responsive microRNAs that can evoke cardiac hypertrophy and heart failure. Proc Natl Acad Sci USA 2006; 103:18255-60; PMID:17108080; http://dx.doi.org/10.1073/pnas.0608791103
5. Thum T, Gross C, Fiedler J, Fischer T, Kissler S, Bussen M, et al. MicroRNA-21 contributes to myocardial disease by stimulating MAP kinase signalling in fibroblasts. Nature 2008; 456:980-4; PMID:19043405; http://dx.doi.org/10.1038/nature07511
6. Garofalo M, Romano G, Di Leva G, Nuovo G, Jeon YJ, Ngankeu A, et al. EGFR and MET receptor tyrosine kinase-altered microRNA expression induces tumorigenesis and gefitinib resistance in lung cancers. Nat Med 2012; 18:74-82; PMID:22157681; http://dx.doi.org/10.1038/nm.2577
7. Bousquet M, Harris MH, Zhou B, Lodish HF. MicroRNA miR-125b causes leukemia. Proc Natl Acad Sci USA 2010; 107:21558-63; PMID:21118985; http://dx.doi.org/10.1073/pnas.1016611107
8. Geekiyanage H, Chan C. MicroRNA-137/181c regulates serine palmitoyltransferase and in turn amyloid β, novel targets in sporadic Alzheimer's disease. J Neurosci 2011; 31:14820-30; PMID:21994399; http://dx.doi.org/10.1523/JNEUROSCI.3883-11.2011
9. Jordan SD, Krüger M, Willmes DM, Redemann N, Wunderlich FT, Brönneke HS, et al. Obesity-induced overexpression of miRNA-143 inhibits insulin-stimulated AKT activation and impairs glucose metabolism. Nat Cell Biol 2011; 13:434-46; PMID:21441927; http://dx.doi.org/10.1038/ncb2211
10. Baxter D, McInnes IB, Kurowska-Stolarska M. Novel regulatory mechanisms in inflammatory arthritis: a role for microRNA. Immunol Cell Biol 2012; 90:288-92; PMID:22249200; http://dx.doi.org/10.1038/icb.2011.114
11. Jentzsch C, Leierseder S, Loyer X, Flohrschütz I, Sassi Y, Hartmann D, et al. A phenotypic screen to identify hypertrophy-modulating microRNAs in primary cardiomyocytes. J Mol Cell Cardiol 2012; 52:13-20; PMID:21801730; http://dx.doi.org/10.1016/j.yjmcc.2011.07.010
12. Mestdagh P, Boström AK, Impens F, Fredlund E, Van Peer G, De Antonellis P, et al. The miR-17-92 microRNA cluster regulates multiple components of the TGF-β pathway in neuroblastoma. Mol Cell 2010; 40:762-73; PMID:21145484; http://dx.doi.org/10.1016/j.molcel.2010.11.038
13. Tsang JS, Ebert MS, van Oudenaarden A. Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures. Mol Cell 2010; 38:140-53; PMID:20385095; http://dx.doi.org/10.1016/j.molcel.2010.03.007
14. Ganesan J, Ramanujam D, Sassi Y, Ahles A, Jentzsch C, Werfel S, et al. MiR-378 Controls Cardiac Hypertrophy by Combined Repression of MAP Kinase Pathway Factors. Circulation 2013; 127:2097-106; PMID:23625957; http://dx.doi.org/10.1161/CIRCULATIONAHA.112.000882
15. Chekulaeva M, Filipowicz W. Mechanisms of miRNA-mediated post-transcriptional regulation in animal cells. Curr Opin Cell Biol 2009; 21:452-60; PMID:19450959; http://dx.doi.org/10.1016/j.ceb.2009.04.009
16. Place RF, Li LC, Pookot D, Noonan EJ, Dahiya R. MicroRNA-373 induces expression of genes with complementary promoter sequences. Proc Natl Acad Sci USA 2008; 105:1608-13; PMID:18227514; http://dx.doi.org/10.1073/pnas.0707594105
17. Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. Nature 2008; 455:64-71; PMID:18668037; http://dx.doi.org/10.1038/nature07242
18. Eulalio A, Huntzinger E, Nishihara T, Rehwinkel J, Fauser M, Izaurralde E. Deadenylation is a widespread effect of miRNA regulation. RNA 2009; 15:21-32; PMID:19029310; http://dx.doi.org/10.1261/rna.1399509
19. Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 2010; 466:835-40; PMID:20703300; http://dx.doi.org/10.1038/nature09267
20. Thomas M, Lieberman J, Lal A. Desperately seeking microRNA targets. Nat Struct Mol Biol 2010; 17:1169-74; PMID:20924405; http://dx.doi.org/10.1038/nsmb.1921
21. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet 2008; 9:102-14; PMID:18197166; http://dx.doi.org/10.1038/nrg2290
22. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005; 120:15-20; PMID:15652477; http://dx.doi.org/10.1016/j.cell.2004.12.035
23. Grimson A, Farh KKH, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 2007; 27:91-105; PMID:17612493; http://dx.doi.org/10.1016/j.molcel.2007.06.017
24. Friedman RC, Farh KKH, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 2009; 19:92-105; PMID:18955434; http://dx.doi.org/10.1101/gr.082701.108
25. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets. PLoS Biol 2004; 2:e363; PMID:15502875; http://dx.doi.org/10.1371/journal.pbio.0020363
26. Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNA.org resource: targets and expression. Nucleic Acids Res 2008; 36(Database issue):D149-53; PMID:18158296; http://dx.doi.org/10.1093/nar/gkm995
27. Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 2010; 11:R90; PMID:20799968; http://dx.doi.org/10.1186/gb-2010-11-8-r90
28. Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature 2009; 460:479-86; PMID:19536157; http://dx.doi.org/10.1038/nature08170
29. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 2010; 141:129-41; PMID:20371350; http://dx.doi.org/10.1016/j.cell.2010.03.009
30. König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol 2010; 17:909-15; PMID:20601959; http://dx.doi.org/10.1038/nsmb.1838
31. Lal A, Thomas MP, Altschuler G, Navarro F, O'Day E, Li XL, et al. Capture of microRNA-bound mRNAs identifies the tumor suppressor miR-34a as a regulator of growth factor signaling. PLoS Genet 2011; 7:e1002363; PMID:22102825; http://dx.doi.org/10.1371/journal.pgen.1002363
32. Kowarsch A, Marr C, Schmidl D, Ruepp A, Theis FJ. Tissue-Specific Target Analysis of Disease-Associated MicroRNAs in Human Signaling Pathways. Morris RJ, ed. *PLoS ONE* 2010; 5(6):e11154; PMID:20614023; http://dx.doi.org/10.1371/journal.pone.0011154
33. Kowarsch A, Preusse M, Marr C, Theis FJ. miTALOS: analyzing the tissue-specific regulation of signaling pathways by human and mouse microRNAs. RNA 2011; 17:809-19; PMID:21441347; http://dx.doi.org/10.1261/rna.2474511
34. Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. Nature 2008; 455:58-63; PMID:18668040; http://dx.doi.org/10.1038/nature07228
35. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. Cell 2003; 115:787-98; PMID:14697198; http://dx.doi.org/10.1016/S0092-8674(03)01018-3
36. Rajewsky N, Socci ND. Computational identification of microRNA targets. Dev Biol 2004; 267:529-35; PMID:15013811; http://dx.doi.org/10.1016/j.ydbio.2003.12.003
37. Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993; 75:843-54; PMID:8252621; http://dx.doi.org/10.1016/0092-8674(93)90529-Y

38. Wightman B, Ha I, Ruvkun G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*. Cell 1993; 75:855-62; PMID:8252622; http://dx.doi.org/10.1016/0092-8674(93)90530-4

39. Kloosterman WP, Wienholds E, Ketting RF, Plasterk RH. Substrate requirements for let-7 function in the developing zebrafish embryo. Nucleic Acids Res 2004; 32:6284-91; PMID:15585662; http://dx.doi.org/10.1093/nar/gkh968

40. Doench JG, Petersen CP, Sharp PA. siRNAs can function as miRNAs. Genes Dev 2003; 17:438-42; PMID:12600936; http://dx.doi.org/10.1101/gad.1064703

41. Saetrom P, Heale BSE, Snøve O Jr., Aagaard L, Alluin J, Rossi JJ. Distance constraints between microRNA target sites dictate efficacy and cooperativity. Nucleic Acids Res 2007; 35:2333-42; PMID:17389647; http://dx.doi.org/10.1093/nar/gkm133

42. Broderick JA, Salomon WE, Ryder SP, Aronin N, Zamore PD. Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing. RNA 2011; 17:1858-69; PMID:21878547; http://dx.doi.org/10.1261/rna.2778911

43. Doench JG, Sharp PA. Specificity of microRNA target selection in translational repression. Genes Dev 2004; 18:504-11; PMID:15014042; http://dx.doi.org/10.1101/gad.1184404

44. Hill AV. A new mathematical treatment of changes of ionic concentration in muscle and nerve under the action of electric currents, with a theory as to their mode of excitation. J Physiol 1910; 40:190-224; PMID:16993004

45. Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol 2011; 18:1139-46; PMID:21909094; http://dx.doi.org/10.1038/nsmb.2115

46. Ritchie W, Flamant S, Rasko JEJ. mimiRNA: a microRNA expression profiler and classification resource designed to identify functional correlations between microRNAs and their targets. Bioinformatics 2010; 26:223-7; PMID:19933167; http://dx.doi.org/10.1093/bioinformatics/btp649

47. Ruepp A, Kowarsch A, Schmidl D, Buggenthin F, Brauner B, Dunger I, et al. PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. Genome Biol 2010; 11:R6; PMID:20089154; http://dx.doi.org/10.1186/gb-2010-11-1-r6

48. Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, et al. A mammalian microRNA expression atlas based on small RNA library sequencing. Cell 2007; 129:1401-14; PMID:17604727; http://dx.doi.org/10.1016/j.cell.2007.04.040

49. Gaur A, Jewell DA, Liang Y, Ridzon D, Moore JH, Chen C, et al. Characterization of microRNA expression levels and their biological correlates in human cancer cell lines. Cancer Res 2007; 67:2456-68; PMID:17363563; http://dx.doi.org/10.1158/0008-5472.CAN-06-2698

50. Lee EJ, Baek M, Gusev Y, Brackett DJ, Nuovo GJ, Schmittgen TD. Systematic evaluation of microRNA processing patterns in tissues, cell lines, and tumors. RNA 2008; 14:35-42; PMID:18025253; http://dx.doi.org/10.1261/rna.804508

51. Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol 2006; 411:352-69; PMID:16939800; http://dx.doi.org/10.1016/S0076-6879(06)11019-8

52. Jayawardena TM, Egemnazarov B, Finch EA, Zhang L, Payne JA, Pandya K, et al. MicroRNA-mediated in vitro and in vivo direct reprogramming of cardiac fibroblasts to cardiomyocytes. Circ Res 2012; 110:1465-73; PMID:22539765; http://dx.doi.org/10.1161/CIRCRESAHA.112.269035

53. Anokye-Danso F, Trivedi CM, Juhr D, Gupta M, Cui Z, Tian Y, et al. Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. Cell Stem Cell 2011; 8:376-88; PMID:21474102; http://dx.doi.org/10.1016/j.stem.2011.03.001

54. Miyoshi N, Ishii H, Nagano H, Haraguchi N, Dewi DL, Kano Y, et al. Reprogramming of mouse and human cells to pluripotency using mature microRNAs. Cell Stem Cell 2011; 8:633-8; PMID:21620789; http://dx.doi.org/10.1016/j.stem.2011.05.001

55. Jiang P, Coller H. Functional Interactions Between microRNAs and RNA Binding Proteins. MicroRNA 2012; 1:70-9

56. Srikantan S, Tominaga K, Gorospe M. Functional interplay between RNA-binding protein HuR and microRNAs. Curr Protein Pept Sci 2012; 13:372-9; PMID:22708488; http://dx.doi.org/10.2174/138920312801619394

57. Jacobsen A, Wen J, Marks DS, Krogh A. Signatures of RNA binding proteins globally coupled to effective microRNA target sites. Genome Res 2010; 20:1010-9; PMID:20508147; http://dx.doi.org/10.1101/gr.103259.109

58. Yang JH, Li JH, Shao P, Zhou H, Chen YQ, Qu LH. starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. Nucleic Acids Res 2011; 39(Database issue):D202-9; PMID:21037263; http://dx.doi.org/10.1093/nar/gkq1056

59. Bauer DF. Constructing Confidence Sets Using Rank Statistics. J Am Stat Assoc 1972; 67:687-90; http://dx.doi.org/10.1080/01621459.1972.10481279

A.2 Publication 2

Preusse M, Theis FJ, Mueller NS. *miTALOS v2: Analyzing Tissue Specific microRNA Function.* PLoS One. 2016;11(3):e0151771.

RESEARCH ARTICLE

miTALOS v2: Analyzing Tissue Specific microRNA Function

Martin Preusse[1,2], **Fabian J. Theis**[1,3], **Nikola S. Mueller**[1]*

1 Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany, **2** Institute of Diabetes and Regeneration Research, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany, **3** Institute for Mathematical Sciences, Technische Universität München, Munich, Germany

* nikola.mueller@helmholtz-muenchen.de

OPEN ACCESS

Citation: Preusse M, Theis FJ, Mueller NS (2016) miTALOS v2: Analyzing Tissue Specific microRNA Function. PLoS ONE 11(3): e0151771. doi:10.1371/journal.pone.0151771

Editor: Bibekanand Mallick, National Institute of Technology, Rourkela, INDIA

Received: December 16, 2015

Accepted: March 3, 2016

Published: March 21, 2016

Data Availability Statement: All relevant data are in the paper and on the web application described in the article: http://mips.helmholtz-muenchen.de/mitalos.

Funding: This work was supported by Deutsche Forschungsgemeinschaft (InKoMBio: FKZ TH 900/3-2) to MP, an ERC starting grant award (LatentCauses) to FJT and Deutsche Forschungsgemeinschaft (PsyCourse: FI 981/9-1) to NSM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

MicroRNAs are involved in almost all biological processes and have emerged as regulators of signaling pathways. We show that miRNA target genes and pathway genes are not uniformly expressed across human tissues. To capture tissue specific effects, we developed a novel methodology for tissue specific pathway analysis of miRNAs. We incorporated the most recent and highest quality miRNA targeting data (TargetScan and StarBase), RNA-seq based gene expression data (EBI Expression Atlas) and multiple new pathway data sources to increase the biological relevance of the predicted miRNA-pathway associations. We identified new potential roles of miR-199a-3p, miR-199b-3p and the miR-200 family in hepatocellular carcinoma, involving the regulation of metastasis through MAPK and Wnt signaling. Also, an association of miR-571 and Notch signaling in liver fibrosis was proposed. To facilitate data update and future extensions of our tool, we developed a flexible database backend using the graph database neo4j. The new backend as well as the novel methodology were included in the updated miTALOS v2, a tool that provides insights into tissue specific miRNA regulation of biological pathways. miTALOS v2 is available at http://mips.helmholtz-muenchen.de/mitalos.

Introduction

MicroRNAs (miRNAs) are short, non-coding RNAs that regulate gene expression post transcriptionally through binding to a target mRNA. They are predicted to target hundreds of genes in mammals and most genes are thought to be regulated by miRNAs [1]. Consequently, most biological processes involve miRNAs and miRNA-mediated control of gene expression.

Functional analysis of miRNAs depends on accurate identification of gene targets in a given biological context [2]. Since there is no comprehensive catalogue of tissue and cell type specific miRNA-mRNA interactions, computational target prediction tools are still widely used. Although these prediction tools have improved in accuracy, they still suffer from large numbers of false-positive miRNA-mRNA interactions [2]. Recently, biochemical methods using sequencing of target RNA isolated after UV crosslinking and immunoprecipitation of Ago/miRNA complexes (CLIP-seq) were developed [3,4]. They produce a map of miRNA binding sites on their target mRNAs. CLIP-seq data is collected in the StarBase database [5], providing a constantly growing resource of experimentally supported interactions. While these experimental methods increase the specificity of miRNA target data, their explanatory power is limited due to differences in experimental procedures and lack of reproducibility [6]. Moreover, all human data sets in StarBase were measured in immortalized cell lines (HEK293, HeLa) and not in primary tissue.

Next to limitations of *in-silico* and experimental gene target identification, miRNA-mediated regulation suggested by *in-vitro* and cell culture experiments is often not supported by *in-vivo* validation studies [7]. This can be partly explained by the fact that most miRNAs show only limited effects on the level of individual target mRNAs under physiological conditions [8]. In addition, target prediction and CLIP-seq studies demonstrated that most mRNAs are regulated by multiple miRNAs [9–11]. Thus, the down-regulation of a target gene depends on the combined effect of multiple miRNAs. And analysis of individual miRNA-mRNA interactions is not sufficient to explain the regulatory role of miRNAs in biological process.

Computational approaches often perform a pathway analysis to increase the explanatory power of target gene sets and to circumvent the shortcomings in targeting data. They use the complete set of miRNA target genes and pathway genes to associate miRNAs to biological pathways as an indication of their biological function. In doing so they do not account for the characteristic tissue expression signature of mammalian genes [12] and thus disregard tissue specific effects of miRNAs. Indeed, miRNAs were shown to facilitate tissue specificity of gene regulation [13]. Moreover, other pathway analysis tools such as DIANA mirPath rely on target prediction only and do not use CLIP-seq based target data [14].

Tissue-specific gene expression data can be obtained using next-generation sequencing of RNA (RNA-seq). The EBI Expression Atlas [15] collects highly curated gene expression data sets and also includes baseline expression data for healthy tissue or untreated cell lines in various organisms. Baseline expression describes the abundance of a gene and is extracted from large-scale expression studies such as ENCODE cell lines.

We developed a novel pathway analysis methodology leveraging this high-quality tissue expression data in order to predict miRNA function. We used our new methodology to first analyze the role of miRNAs in hepatocellular carcinoma and identified the liver-specific effect of miR-199a/b-3p on pathways associated with proliferation and cell migration, a novel function that a recent study proposed. We next dissected the individual functions of the two genomic clusters of the miR-200 family and found hints to new signaling relationships, which were studied in other tissues and cell culture but not yet in liver cancer. We finally extended our analysis to liver fibrosis, which is in general less well studied than liver cancer. miR-571 is known to play a role here and we identified Notch signaling as a putative function. Interestingly, Notch signaling has already been proposed as a drug target for fibrosis in other tissues. With the three case studies we demonstrated the necessity to use tissue-specific target gene information for miRNA function prediction.

To make our novel pathway analysis methodology publicly available, we systematically integrated 1) high-quality miRNA targeting data from TargetScan and CLIP-seq studies from StarBase v2 [5], 2) tissue specific gene expression from the latest version of EBI Expression Atlas [15] with 3) three major pathway databases KEGG [16], WikiPathways [17] and Reactome [18]. A graph database was used to store the data in a flexible manner and increase the query performance compared to relational data stores. The data backend and the corresponding pathway analysis methodology were integrated into miTALOS version 2 (v2), a user-friendly web application to identify pathways regulated by miRNAs in a tissue specific manner. With miTALOS v2 users can analyze multiple miRNAs together to account for combinatorial effects.

MiTALOS v2 is complementary to other functional miRNA analysis tools such as miRGator [19] and ToppMir [20] and adds value with a tissue specific analysis of miRNA impact on signaling pathways. The integration of multiple new state-of-the art data sources increases the biological relevance of the results and a novel tissue filter allows every user to decipher complex miRNA functions.

Results

Tissue specific pathway enrichment

MiRNA target prediction tools and CLIP-seq based methods for target identification yield the full set of potential miRNA-mRNA interactions, i.e. all potential gene targets of a miRNA. However, different tissues and cell types have a characteristic gene expression signature and only a subset of genes are expressed in any cell under physiological conditions [12]. Thus, the function of miRNAs, which is exerted through repression of target genes, is tissue specific.

To learn about the tissue-specificities of miRNAs, we first analyzed the expression of all target genes of hsa-let-7a (TargetScan, see methods) in 42 human tissues from EBI Expression Atlas. The expression of target genes varied greatly between tissues (Fig 1A). To quantify the extent of tissue specificity of a miRNA, we calculated for each of the 42 tissues the fraction of target genes being expressed. The fraction is depicted in Fig 1B (color coded from green = 0 to red = 1). Fig 1C shows the respective distributions for ten representative miRNAs. Thereof, the median of target genes expressed in a tissue was 75%, with many tissues expressing only 60% target genes (Fig 1C). This is in line with studies showing tissue specific functions of miRNAs [13].

Next, we performed the same tissue-specificity analysis now only for genes of the same pathway. The pathway genes in well-described human MAPK signaling (KEGG) showed highly tissue specific expression (Fig 1D). Interestingly pathways showed a characteristic distribution of the fraction of expressed target genes when compared to miRNAs. Ten representative distributions across all 42 tissues are shown in Fig 1E. Some pathways (such as Cell adhesion molecules, Fig 1E) were more tissue specific than others, indicating highly tissue specific functions.

Having established that both miRNA and pathway associated genes have a characteristic gene expression signature across tissues, we next outlined the approach of standard miRNA pathway analysis methods. Typically the set of all miRNA targets are tested for over-representation in the set of all pathway genes (Fig 1F, left). This global analysis of all target and pathway genes will overlook miRNA-pathway associations with a small gene-overlap, while this gene-overlap may in turn be tissue-specific and, thus, functionally highly relevant. Pathway analysis tools that use all target genes to identify miRNA-pathway associations cannot capture tissue specific effects.

We thus propose a novel methodology for miRNA pathway analysis by using a tissue filter in order to increase the relevance of the association. If the target genes or pathway genes outside of the overlap are not expressed in a tissue, the relation of miRNA and pathway is much stronger (Fig 1F, right). Consequently, if the overlapping genes found in a miRNA-pathway association are not expressed in a tissue, the relation is discarded. The novel methodology calculates an enrichment of the target genes of a miRNA in all pathways of different pathway data sources. Significance of the associations is calculated with Fisher's exact test (see methods). Individually for each miRNA-pathway association test, we filtered for expression in a tissue by removing all miRNA target genes and pathway genes that are not expressed in this tissue. We thereby accounted for the highly tissue specific expression of many genes and seek to increase biological relevance of the pathway enrichment.

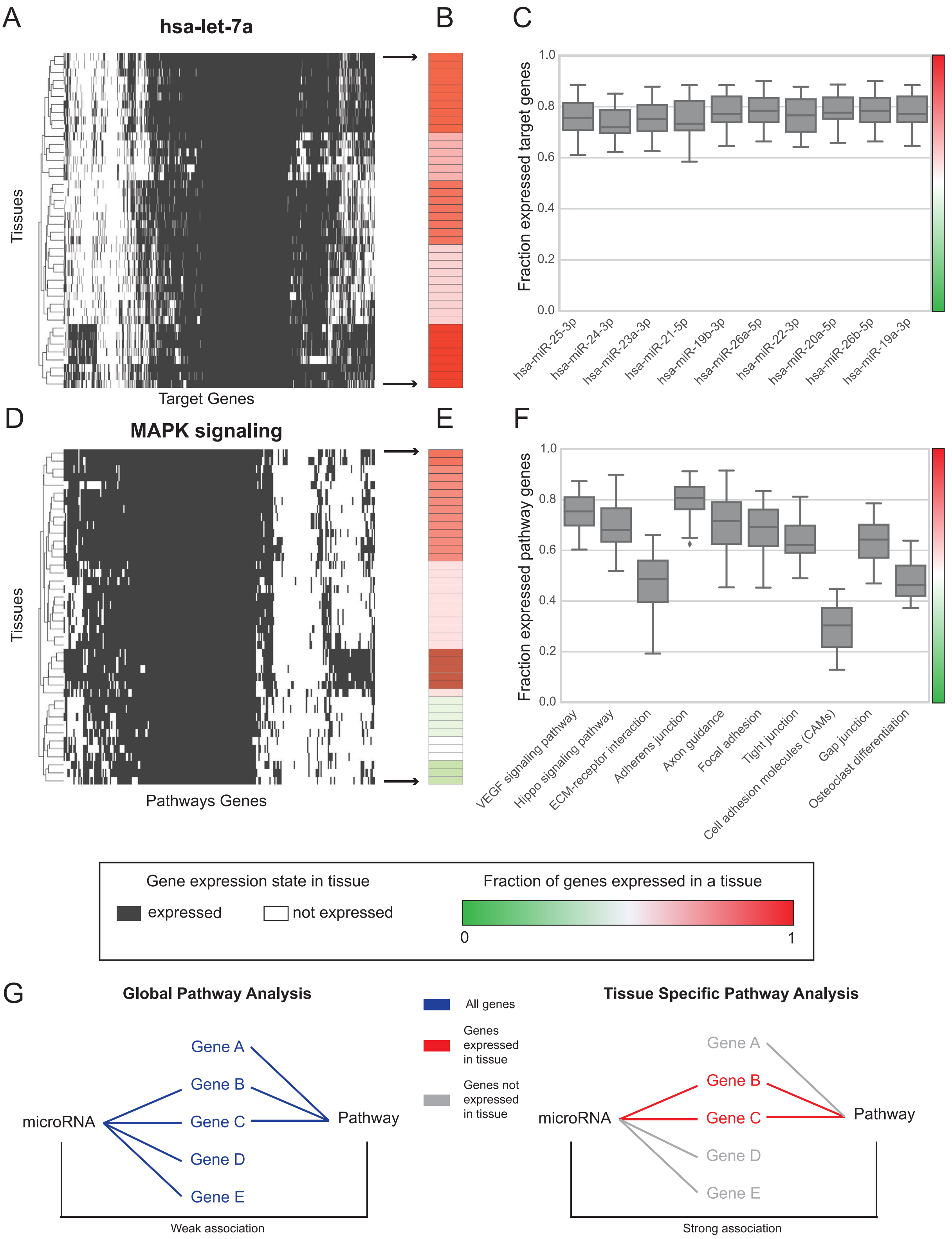


Fig 1. miRNA target genes and pathway genes are tissue specific. (A) Heatmap of all target genes of hsa-let-7a and their expression in 42 human tissues. Tissues are depicted in rows, genes in columns. (B) Fraction of target genes of hsa-let-7a expressed in each tissue, color coded in green (0) to red

(1). (C) Fraction of target genes expressed in all tissues for 10 representative miRNAs. (D)-(F) Corresponding analysis for pathway genes. (G) Pathway analysis with the global set of miRNA targets and pathway genes (left). The miRNA and pathway have only few common genes (gene B, gene C) compared to the other pathway genes (gene A) and miRNA targets (gene D, gene E). When applying a tissue filter (right), genes not in the set of miRNA targets and not in the pathway are discarded. The association derived from the overlap is much stronger, indicating a tissue specific regulation of the pathway by the miRNA.

doi:10.1371/journal.pone.0151771.g001

Case study: microRNAs in liver disease

We analyzed miRNAs known to be involved in liver disease with our novel methodology to evaluate the power of tissue specific pathway analysis. We focused on miRNAs in hepatocellular carcinoma (HCC) and liver fibrosis. Both diseases involve uncontrolled proliferation of liver cells.

First, we analyzed miR-199a-3p and miR-199b-3p. Both miRNAs are up-regulated in some tumor types, such as ovarian cancer and breast cancer [21]. In HCC, conversely, both miRNAs have been shown to be down-regulated [22,23]. While the function of miR-199a-3p and miR-199b-3p is not fully defined, they target members of Raf/MEK/ERK signaling [23]. In general, inhibition of Raf/MEK/ERK signaling will limit proliferation of cells. Thus, downregulation of miR-199a-3p and miR-199b-3p might be a part of the regulatory changes leading to increased proliferation of HCC cells. These miRNAs have consequently been considered as therapeutic targets for treatment of HCC [24].

When performing standard pathway analysis for miR-199a-3p and miR-199b-3p, no cancer-associated pathways were enriched (human, TargetScan). Using our methodology and the Illumina Body Map tissue filter for liver additionally identified two significantly associated pathways: Regulation of actin cytoskeleton (KEGG) and Regulation of Microtubule Cytoskeleton (WikiPathways) (Table 1). The miRNAs were previously not directly associated to regulation of the cytoskeleton, yet both pathways are fundamental for the processes of cell migration, EMT and metastasis. The regulation of actin cytoskeleton (KEGG) pathway overlaps with the MAPK signaling pathway from KEGG and includes several key components of Raf/MEK/ERK signaling (Fig 2A). The liver filter thus identified the known association of miR-199a-3p and miR-199b-3p with Raf/MEK/ERK signaling through associated regulatory pathways. Interestingly, the involvement of miR-199a/b-3p in cell migration and EMT has been described in other tissues [25,26].

Table 1. Pathway enrichment used in the case studies with liver filter.

source	Pathway	E	Corrected p-value	MP, Mn, Pn, U
	hsa-miR-199a, *hsa-miR-199b-3p*			
wp	Regulation of Microtubule Cytoskeleton	3,814	0,034	4, 232, 18, 3982
kegg	Regulation of actin cytoskeleton	2,010	0,049	11, 225, 95, 3905
	hsa-miR-200b, *hsa-miR-200c*, *hsa-miR-429*			
kegg	Focal adhesion	1,895	0,008	25, 593, 83, 3730
wp	Focal Adhesion	1,791	0,024	22, 596, 77, 3736
wp	Wnt Signaling Pathway	2,765	0,028	8, 610, 18, 3795
	hsa-miR-200a, *hsa-miR-141-3p*			
wp	MAPK Cascade	3,194	0,047	5, 408, 15, 3910
reactome	NR1D1 (REV-ERBA) represses gene expression	19,095	0,017	2, 411, 1, 3924
	hsa-miR-571-3p			
wp	Notch Signaling Pathway	9,844	0,000	3, 62, 20, 4069
kegg	Notch signaling pathway	8,945	0,001	3, 62, 22, 4067

doi:10.1371/journal.pone.0151771.t001

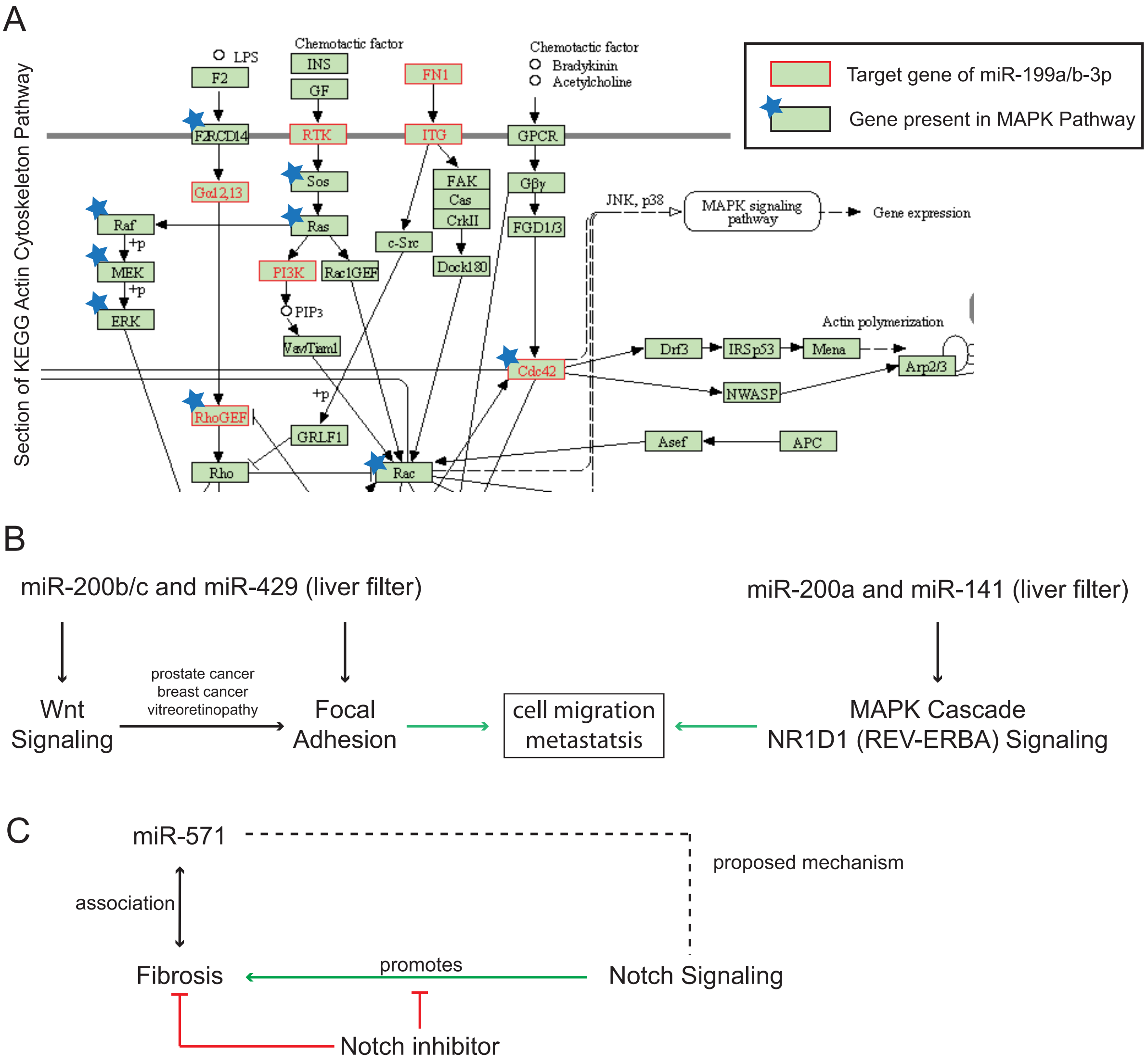


Fig 2. tissue specific enrichment of miRNAs in liver disease. (A) Targets of miR-199a/b-3p in the human KEGG Actin Cytoskeleton pathway (red). Only a section is shown, other parts are not targeted. Blue stars show genes also present in MAPK signaling. (B) Pathway analysis of miR-200 family using the liver filter. MiR-200b/c and miR-429 target Focal adhesion and Wnt signaling, pointing towards a regulatory interdependence in cancer formation. MiR-200a and miR-141 have different associated pathways but also target cancer related signaling. (C) MiR-571 is elevated in fibrosis and associated with notch signaling when using the liver filter. Notch inhibitors are in clinical studies for treatment of early stages of fibrosis.

doi:10.1371/journal.pone.0151771.g002

Our novel methodology with tissue filter suggested a role of miR-199a/b-3p in cell migration, EMT and ultimately metastasis through regulation of cytoskeleton. The decrease of miR-199a/b-3p in HCC might increase metastatic potential in HCC. Indeed, a recent study indicates a role for miR-199a/b-3p in HCC proliferation [27].

Second, we investigated the miR-200 family consisting of two genomic clusters (miR-200b/c/miR-429 and miR-200a/miR-141) that was shown to be involved in EMT and cell migration [28]. The family has been described as a potential cancer therapy target [29]. The miRNAs of the miR-200 family are often analyzed together. Here, we look at specific functions of the two clusters to show the power of combined pathway analysis of multiple miRNAs. When performing pathway analysis with liver filter for miR-200b/c/miR-429 (Illumina Body Map, TargetScan, human) we identified significant associations with focal adhesion pathways from both KEGG and WikiPathways (Table 1). This finding clearly points towards an involvement in cell migration and EMT (Fig 2B). Interestingly, we also identified Wnt pathway (KEGG) (Table 1). As of today, there was no direct evidence reported for involvement of Wnt signaling in regulation of cell migration, EMT and metastatis in HCC. There was, however, evidence for a connection in other diseases such as breast cancer [30], vitreorenopathy [31] and prostate cancer [32].

Our novel methodology suggested new roles for miR-200b/c/miR-429 in HCC and a functional connection of Wnt signaling with cell migration and EMT (Fig 2B). Pathway analysis with liver filter (Illumina Body Map, TargetScan, human) for the other genomic cluster (miR-200a/141) identifies MAPK signaling and MAPK associated NR1D1-(REV-ERBA) pathway (WikiPathways) (Table 1). MAPK signaling was indeed elevated in HCC [33,34] and has been suggested as target for HCC treatment with success in mouse model [35] (Fig 2B). In summary, our novel methodology found specific HCC related functions for both genomic clusters of the miR-200 family. Analyzing the entire miRNA-200 family did not identify focal adhesion, Wnt or MAPK as significant results.

MiRNAs also play a role in liver fibrosis [36] but are in general less well studied in this disease context. There is only few functional evidence or mechanistic insight into the role of miRNAs in fibrosis. This represents an interesting example for the primary use case of our novel methodology: To generate new hypotheses and filter candidate miRNAs to be tested in the wet lab. The serum levels of miR-571 were found increased in cirrhosis (the final stage of fibrosis) and miR-571 has been suggested as a biomarker [37]. With our pathway analysis, we identified Notch signaling (KEGG) as target of miR-571 with liver filter (Illumina Body Map, human, TargetScan) (Table 1). Interestingly, Notch signaling was shown to be over-active in fibrosis [38] and Notch inhibitors have been discussed as potential drugs for treatment of fibrosis [38,39]. As a result, our novel methodology suggests that miR-571 could potentially inhibit Notch signaling in liver tissue. Thus, miR-571 might be a potential therapeutic target in the context of fibrosis (Fig 2C). In summary, our updated novel methodology supported new functional hypotheses through tissue filtered pathway analysis.

Data sources

In order to make the novel methodology publicly available, we first integrated several data sources on miRNA targeting, biological pathways and gene expression for both mouse and human. We downloaded and integrated computational target prediction data from TargetScan 6.2 [40] and miRanda [41]. We also added miRNA-target interaction data of CLIP-seq studies from StarBase v2 [5]. TargetScan contained the majority of mammalian miRNAs while miRanda and StarBase only represented a small subset (Table 1). Due to the limited availability of CLIP-seq studies, we still rely on target prediction data for many miRNAs. Pathway data was extracted from KEGG [16], Reactome [18] and WikiPathways [17]. Pathways in the Reactome database were structured in top-level pathways with smaller sub pathways. This lead to larger numbers of pathways overall compared to KEGG and WikiPathways (Table 1). To allow for a tissue-specific pathway analysis, we used baseline gene expression data for a total of 68 human and mouse tissues and cell lines from the latest EBI Expression Atlas [15]. Baseline expression data was based on reliable RNA-seq experiments and represents abundance levels in healthy tissue or cell lines. We integrated tissue data sets from 6 different expression studies (Table 2).

Table 2. Overview of data sets.

	Human	Mouse
# miRNAs		
TargetScan v6	1529	1322
Miranda	249	238
StarBase v2	383	296
# Pathways		
KEGG	295	291
Reactome	2224	1882
WikiPathways	293	160
# Tissues		
Mammalian Tissues	8	6
Illumina Body Map	16	-
ENCODE Cell Lines	18	-
Vertebrate Tissues	-	5
6 Mouse Tissues	-	6
Nine Mouse Tissues	-	9

doi:10.1371/journal.pone.0151771.t002

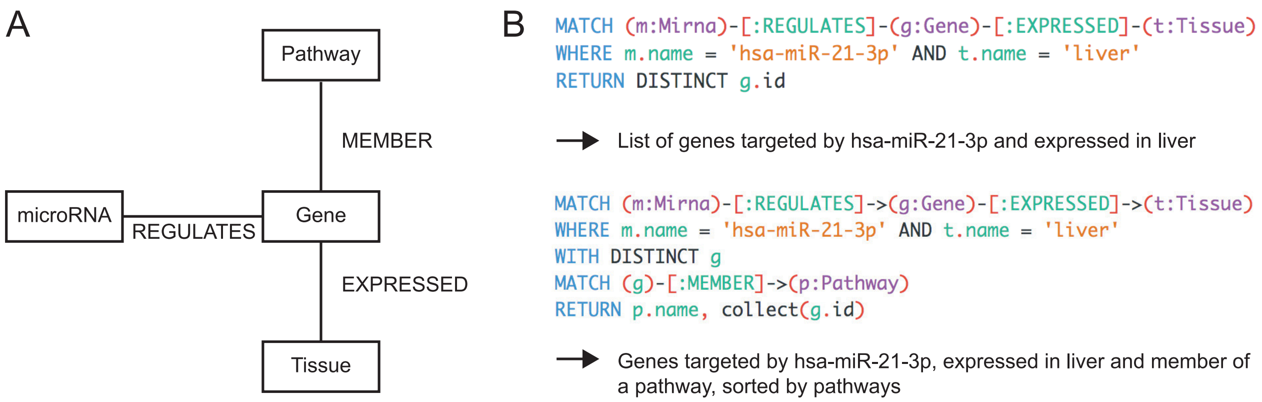
Database backend

Any system that integrates heterogeneous research data has to deal with two major challenges: I) Data has to be stored in a way that it can be queried efficiently and II) the data model must allow for easy updates for new releases of the underlying data sources.

Traditionally, SQL based relational database systems such as MySQL or PostgreSQL were the go-to solution for all data storage needs. In recent years however, new database technologies collectively termed noSQL (short for not-only SQL) were developed to cope with problems arising from big data. Such noSQL technologies have been used successfully in solutions for computational biology, especially in the field of NGS [42]. Among the diverse landscape of new database technologies, graph databases are particularly promising for biological data sets. They enable storing data natively as a property graph, i.e. nodes connected by edges with properties stored on both. Thus, they allow us to directly model biological systems as nodes representing molecular entities connected by edges representing their interaction. This leads to simple queries over multi-step paths through the interaction network and increased performance compared to JOIN operations in relational databases [43,44]. Since queries on biological data are usually centered on relationships between molecular entities (such as genes and miRNAs), graph database have a huge potential to improve data storage solutions. The key advantages are query performance and simple query syntax.

For our study, we used the graph database neo4j and developed a novel graph data model to integrate the data sources described above (Fig 3A). MiRNAs, genes, pathways and tissues were represented as nodes. MiRNAs were connected to genes with 'REGULATES' relationships, genes to tissues with 'EXPRESSED`relationships and genes to pathways with 'MEMBER' relationships. This data structure allowed us to e.g. query the target genes of a miRNA expressed in a tissue (Fig 3B, top) or the pathways in which the target genes are involved (Fig 3B, bottom).

Another challenge in studies based on integration of third party data sources is to keep up with data updates and new releases. Small, specialized data sources publish new versions on their own schedule and changes in one data source are not synchronized with others. Since neo4j is schema-less, changes of parts of the underlying data (e.g. miRNA targeting data for a single data source) and refactoring of the data structure (e.g. renaming of miRNAs) are easier



```
MATCH (m:Mirna)-[:REGULATES]-(g:Gene)-[:EXPRESSED]-(t:Tissue)
WHERE m.name = 'hsa-miR-21-3p' AND t.name = 'liver'
RETURN DISTINCT g.id
```

→ List of genes targeted by hsa-miR-21-3p and expressed in liver

```
MATCH (m:Mirna)-[:REGULATES]->(g:Gene)-[:EXPRESSED]->(t:Tissue)
WHERE m.name = 'hsa-miR-21-3p' AND t.name = 'liver'
WITH DISTINCT g
MATCH (g)-[:MEMBER]->(p:Pathway)
RETURN p.name, collect(g.id)
```

→ Genes targeted by hsa-miR-21-3p, expressed in liver and member of a pathway, sorted by pathways

Fig 3. Database structure of miTALOS v2. (A) The miTALOS v2 dataset is stored in a graph database. The network structure allows for easy extension of the dataset. (B) The Cypher query language allows for simple queries on the network. With one query, the targets of a miRNA in a pathway can be accessed and filtered for tissue expression.

doi:10.1371/journal.pone.0151771.g003

to implement. We thus seek to regularly update our pathway analysis with new data sets especially focusing on NGS based data for miRNA targets and gene expression.

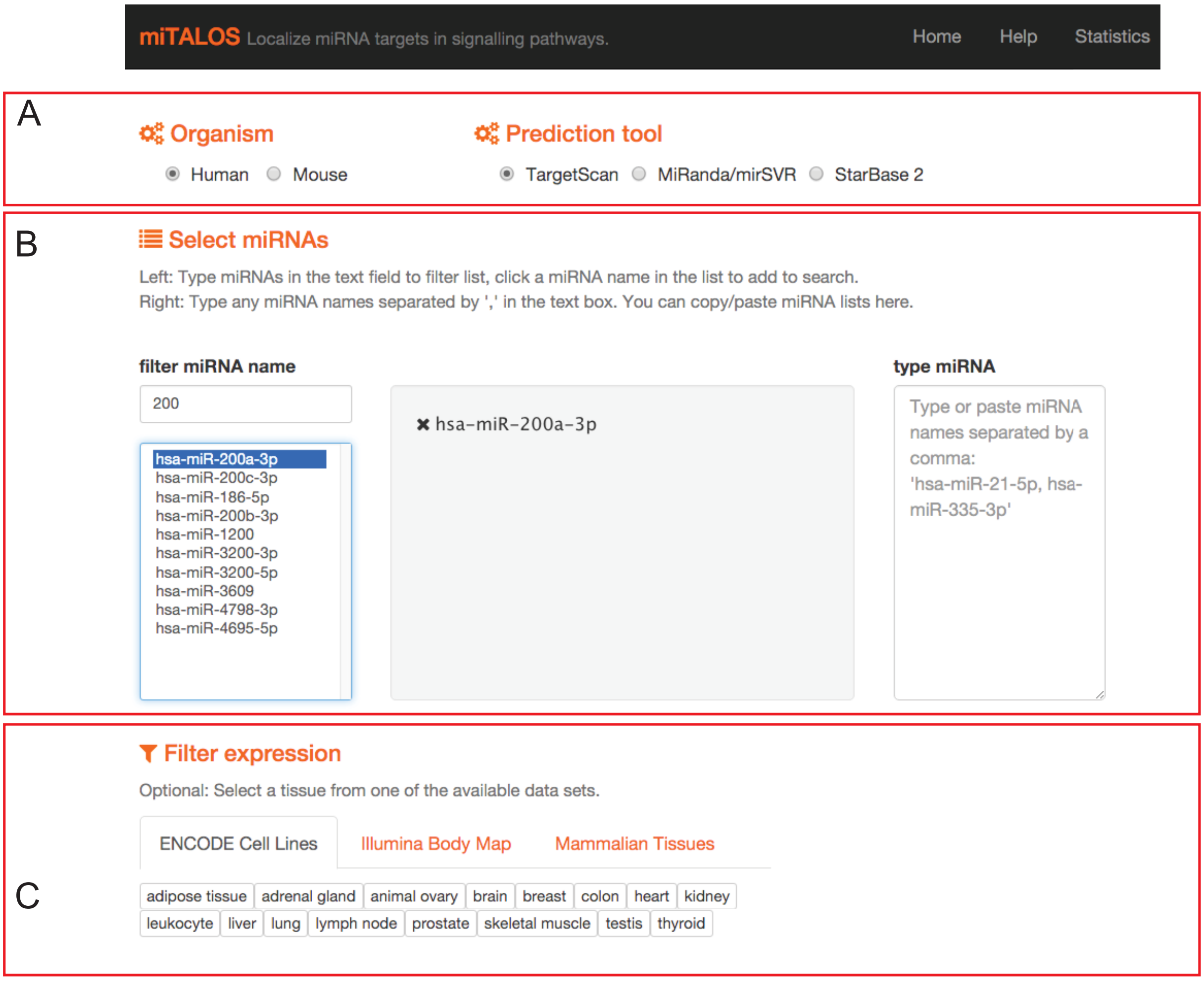
miTALOS v2

In order to make our integrated, tissue specific pathway analysis available to the research community, we included the new analysis methodology and data backend in an update to our miTALOS web application.

MiTALOS v2 is a user-friendly tool to perform tissue specific pathway analysis for a set of miRNAs and tissues of interest (Fig 4). It is available at http://mips.helmholtz-muenchen.de/mitalos. The user can analyze miRNAs from mouse and human. The user begins by selecting the organism and miRNA prediction method (Fig 4A) and then selects one or multiple miRNAs (Fig 4B). The pathway analysis is carried out dynamically by calculating the pathway enrichment (see Methods) on all pathway data sources. If more than one miRNA is selected, the union of target genes will be used for the analysis. All target genes are counted once and no additional ranking is applied. MiTALOS v2 thereby captures the biological impact of co-targeting by multiple miRNAs. If the user selects a tissue filter, all gene sets (miRNA target genes and genes in pathways) are filtered for this tissue (Fig 4C). All results with a corrected p-value > 0.05 and $E > 1$ are presented in a sortable table and can be accessed with a user specific URL for one week. For KEGG pathways, the user can access a graphical representation of the pathway with highlighted miRNA targets by clicking on a pathway name.

If a tissue filter is used, miTALOS v2 displays the expression score of the selected miRNAs inaddition to the tissue specific pathway enrichment. The user can thereby assess the impact of the selected miRNAs under physiological conditions. The absolute expression score is extended by a rank of the selected miRNA among all miRNAs expressed in this tissue and the miRNA with the overall highest expression value. This allows estimating the relative importance of the selected miRNA in the analyzed tissue.

MiTALOS v2 is geared towards wet-lab researchers working with miRNAs. MiTALOS v2 was designed for scenarios where a set of miRNAs (e.g. from expression studies or literature research) has to be filtered to identify the most promising miRNAs for testing in wet-lab experiments. With the tissue filter, the user can analyze the supposed biological effect of miRNAs in the particular tissue or cell line the user is working on.



Organism

● Human ○ Mouse

Prediction tool

● TargetScan ○ MiRanda/mirSVR ○ StarBase 2

Select miRNAs

Left: Type miRNAs in the text field to filter list, click a miRNA name in the list to add to search.
Right: Type any miRNA names separated by ',' in the text box. You can copy/paste miRNA lists here.

filter miRNA name

200

hsa-miR-200a-3p
hsa-miR-200c-3p
hsa-miR-186-5p
hsa-miR-200b-3p
hsa-miR-1200
hsa-miR-3200-3p
hsa-miR-3200-5p
hsa-miR-3609
hsa-miR-4798-3p
hsa-miR-4695-5p

✖ hsa-miR-200a-3p

type miRNA

Type or paste miRNA names separated by a comma: 'hsa-miR-21-5p, hsa-miR-335-3p'

Filter expression

Optional: Select a tissue from one of the available data sets.

ENCODE Cell Lines | Illumina Body Map | Mammalian Tissues

adipose tissue | adrenal gland | animal ovary | brain | breast | colon | heart | kidney
leukocyte | liver | lung | lymph node | prostate | skeletal muscle | testis | thyroid

Fig 4. User interface of miTALOS v2. (A) The user starts by selecting the organism and miRNA prediction tool. Next, multiple miRNAs can be selected by filtering the list of available miRNAs (B). Lastly, a tissue filter can be applied by selecting an expression experiment and tissue or cell line (C).

doi:10.1371/journal.pone.0151771.g004

Discussion

It has been established that miRNAs participate in almost all cellular processes but the functional impact of individual miRNAs and the precise mode of target gene regulation remains controversial. Consequently, the dynamic regulatory network of miRNAs and mRNAs under physiological conditions is not fully understood. One of the key issues in miRNA resarch is the identification and quantification of miRNA-mRNA interactions. While computational prediction methods and CLIP-seq approaches yield global sets of gene targets for individual miRNAs, they still suffer from lack of accuracy and fail to predict the regulatory landscape in-vivo.

One way to circumvent shortcomings in miRNA targeting data is to analyze the biological pathways which are incluenced by miRNAs. They can be considered a proxy for the miRNAs effect on biological processes and thus allow to classify miRNAs and generate new hypotheses. While pathway analyses have proven useful, they do not consider that most genes which are targeted by a miRNA or part of a pathway are not uniformly expressed across all cell types. The tissue specifity of miRNAs, which has been demonstrated extensively, is thus not taken into account.

By integrating tissue specific gene expression into our pathway analysis methodology, we seek to close this gap and improve the biological relevance of our miRNA-pathway associations. With our case studies, we recapitulated a common approach to generate new miRNA hypotheses for wet lab research: Based on prior knowledge, i.e. disregulation of several miRNAs in a disease context, the best candidates for experimental testing have to be identified. Our methodology aims at creating functional insight which is as specific as possible for the system studied by the user.

The distinctive feature of miTALOS v2 is the tissue specific pathway enrichment. Other pathway analysis tools, such as DIANA mirPath [45], do not account for this effect. MiTALOS v2 complements other methods for functional miRNA analysis. Tools analyzing the expression of miRNAs, such as MiRGator [19], aid in selecting the best miRNA candidates for a specific biological system. Ranking approaches, such as ToppMir [20], are used to limit the number of miRNAs based on preference for user-defined gene sets. MiTALOS v2 can be used in conjunction with these methods and adds a tissue specific perspective.

MiTALOS v2 includes CLIP-seq based miRNA targeting data from the StarBase database. CLIP-seq experiments generate the full set of target genes based on biochemically identified miRNA-mRNA interactions and likely produce more reliable targeting data than computational prediction. Several public resources, such as miRTarBase [46] and miRecords [47], collect miRNA targets validated in individual experiments. However, since these target sets contain only a potentially small subset of miRNA-mRNA interactions they would introduce a bias to the analysis and are thus not suitable for global pathway enrichment.

Next to TargetScan and miRanda, which were used in this study, there are several other miRNA target prediction tools. However, it is difficult to compare their performance due to the lack of a gold standard of known miRNA targets and systematic comparisons of target prediction tools generated inconsistent results [48–51]. TargetScan and miRanda were chosen based on their widespread use in the miRNA research community. If novel miRNA target data sources arise, the miTALOS v2 data can easily be integrated in miTALOS v2.

In general, the effect of a miRNA on its target genes cannot be quantified cell wide. The complexity of the miRNA-mRNA network was further increased when regulatory effects came into focus [52]. It was demonstrated that the total number of potential binding sites for a miRNA regulates its effect size. If the number of binding sites exceeds the number of miRNA molecules, mRNAs compete for binding to the miRNA and the regulatory impact decreases [53]. This has been subsumed under the concept of competing endogenous RNAs (ceRNAs). Recently, combined computational and experimental studies quantified these effects on a systems level [54]. Including these indirect effects into a pathway analysis presents a future direction for miTALOS v2. Here, using the relative expression levels of miRNAs and their target genes would allow to capture binding competition. However, more data on specific, quantitative effects will be necessary to devise a computational approach that properly describes the biological impact of competing RNAs.

When developing tools for the research community, the underlying data infrastructure is of pivotal importance. The state of the art, especially in research of post-transcriptional regulation, changes quickly and new methods for miRNA target identification might arise. We therefore developed a new database backend using neo4j, the leading graph database. It helps to integrate the numerous datasets used in miTALOS v2 and to keep up with new developments. The flexible backend also allows to integrate new aspects like lncRNAs as regulators of gene expression or disease specific expression profiles to extend tissue specific gene expression. New database technology is therefore instrumental in building tools which can adapt to the rapid generation of new research results.

In summary, our pathway analysis methodology and miTALOS v2 have been developed to generate testable hypotheses and to increase efficiency in experimental miRNA research.

Table 3. 2x2 cross table.

	Pathway *P*	
miRNA *M*	MP	Mn
	Pn	U

doi:10.1371/journal.pone.0151771.t003

Methods

Datasets

We integrated several data sources on miRNA targeting, biological pathways and gene expression in order to analyse tissue specific miRNA functions. For mouse and human, we offer computational target prediction data from the latest releases of TargetScan 6.2 [40] and miRanda [41]. We added miRNA-target interaction data of CLIP-seq studies from StarBase v2 [5] to the miTALOS v2 pathway analysis. Pathway data was extracted from KEGG, Reactome and WikiPathways. In order to analyze tissue specific pathway regulation, miTALOS v2 uses baseline gene expression data for 68 tissues and cell lines from the latest EBI Expression Atlas [15] for both mouse and human.

Pathway analysis

We calculate an enrichment of miRNA target genes in pathways. For a miRNA *M* and Pathway *P* miTALOS v2 calculates a 2x2 cross table, where *MP* is the number of targets of *M* in *P*, *Pn* is the number of not targeted genes in *P*, *Mn* is the number of targets of *M* not in *P* and *U* is the union of all pathway genes and miRNA targets without *MP*, *Pn* and *Mn* (Table 3):

An enrichment score *E* is calculated as the odds ratio of *M* and *P*:

$$E(M, P) = (MP/Pn)/(Mn/U)$$

E describes the dependence of variables *M* and *P*. $E > 1$ indicates an over-representation of targets of miRNA *M* in the pathway *P*. A p-value is calculated using Fisher's exact test and results for multiple pathways are corrected using the Benjamini-Hochberg procedure [55].

To perform a tissue specific pathway enrichment, we remove all genes from *MP*, *Mn*, *Pn* and *U* that are not expressed in the analyzed tissue. We then calculate *E* as described above. A gene is considered expressed if its baseline expression value is > 0.5 (as defined in the EBI Expression Atlas).

When multiple miRNAs are selected, the union of target genes is used for the analysis.

Database and webinterface

The integrated database backend is uses a neo4j graph database (v2.3.1). The miTALOS v2 frontend was developed with AngularJS 1.4.

Author Contributions

Conceived and designed the experiments: MP NSM FJT. Performed the experiments: MP. Analyzed the data: MP. Wrote the paper: MP NSM.

References

1. Friedman RC, Farh KK-H, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 2009; 19: 92–105. doi: 10.1101/gr.082701.108 PMID: 18955434

2. Thomas M, Lieberman J, Lal A. Desperately seeking microRNA targets. Nat Struct Mol Biol. Nature Publishing Group; 2010; 17: 1169–74. doi: 10.1038/nsmb.1921

3. Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature. Nature Publishing Group; 2009; 460: 479–86. doi: 10.1038/nature08170

4. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell. Elsevier Ltd; 2010; 141: 129–41. doi: 10.1016/j.cell.2010.03.009

5. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res. 2013; 1–6.

6. Jungkamp AC, Stoeckius M, Mecenas D, Grün D, Mastrobuoni G, Kempa S, et al. In vivo and transcriptome-wide identification of RNA binding protein target sites. Mol Cell. 2011; 44: 828–840. doi: 10.1016/j.molcel.2011.11.009 PMID: 22152485

7. Vidigal J a., Ventura A. The biological functions of miRNAs: lessons from in vivo studies. Trends Cell Biol. Elsevier Ltd; 2014; 25: 137–147. doi: 10.1016/j.tcb.2014.11.004

8. Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. Nature. 2008; 455: 64–71. doi: 10.1038/nature07242 PMID: 18668037

9. Cui Q, Yu Z, Purisima EO, Wang E. Principles of microRNA regulation of a human cellular signaling network. Mol Syst Biol. 2006; 2: 46. doi: 10.1038/msb4100089 PMID: 16969338

10. Inui M, Martello G, Piccolo S. MicroRNA control of signal transduction. Nat Rev Mol Cell Biol. 2010; 11: 252–63. doi: 10.1038/nrm2868 PMID: 20216554

11. Rinck A, Preusse M, Laggerbauer B, Lickert H, Engelhardt S, Theis FJ. The human transcriptome is enriched for miRNA-binding sites located in cooperativity-permitting distance. RNA Biol. 2013; 10: 1125–35. doi: 10.4161/rna.24955 PMID: 23696004

12. Su AI, Wiltshire T, Batalov S, Lapp H, Ching K a, Block D, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 2004; 101: 6062–7. doi: 10.1073/pnas.0400782101 PMID: 15075390

13. Farh KK-H, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, et al. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science. 2005; 310: 1817–21. doi: 10.1126/science.1121158 PMID: 16308420

14. Vlachos IS, Kostoulas N, Vergoulis T, Georgakilas G, Reczko M, Maragkakis M, et al. DIANA miRPath v.2.0: Investigating the combinatorial effect of microRNAs in pathways. Nucleic Acids Res. 2012; 40: 498–504. doi: 10.1093/nar/gks494

15. Petryszak R, Burdett T, Fiorelli B, Fonseca N a, Gonzalez-Porta M, Hastings E, et al. Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. Nucleic Acids Res. 2014; 42: D926–32. doi: 10.1093/nar/gkt1270 PMID: 24304889

16. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28: 27–30. doi: 10.1093/nar/28.1.27 PMID: 10592173

17. Kelder T, Van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, et al. WikiPathways: Building research communities on biological pathways. Nucleic Acids Res. 2012; 40: 1301–1307. doi: 10.1093/nar/gkr1074

18. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011; 39: D691–7. doi: 10.1093/nar/gkq1018 PMID: 21067998

19. Cho S, Jang I, Jun Y, Yoon S, Ko M, Kwon Y, et al. MiRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting. Nucleic Acids Res. 2013; 41: D252–7. doi: 10.1093/nar/gks1168 PMID: 23193297

20. Wu C, Bardes EE, Jegga AG, Aronow BJ. ToppMiR: ranking microRNAs and their mRNA targets based on biological functions and context. Nucleic Acids Res. 2014; 42: W107–13. doi: 10.1093/nar/gku409 PMID: 24829448

21. Chen R, Alvero a B, Silasi D a, Kelly MG, Fest S, Visintin I, et al. Regulation of IKKbeta by miR-199a affects NF-kappaB activity in ovarian cancer cells. Oncogene. 2008; 27: 4712–23. doi: 10.1038/onc.2008.112 PMID: 18408758

22. Murakami Y, Yasuda T, Saigo K, Urashima T, Toyoda H, Okanoue T, et al. Comprehensive analysis of microRNA expression patterns in hepatocellular carcinoma and non-tumorous tissues. Oncogene. 2006; 25: 2537–45. doi: 10.1038/sj.onc.1209283 PMID: 16331254

23. Hou J, Lin L, Zhou W, Wang Z, Ding G, Dong Q, et al. Identification of miRNomes in human liver and hepatocellular carcinoma reveals miR-199a/b-3p as therapeutic target for hepatocellular carcinoma. Cancer Cell. Elsevier Inc.; 2011; 19: 232–43. doi: 10.1016/j.ccr.2011.01.001

24. Callegari E, Elamin BK, D'Abundo L, Falzoni S, Donvito G, Moshiri F, et al. Anti-tumor activity of a miR-199-dependent oncolytic adenovirus. PLoS One. 2013; 8: e73964. doi: 10.1371/journal.pone.0073964 PMID: 24069256

25. Duan Z, Choy E, Harmon D, Liu X, Susa M, Mankin H, et al. MicroRNA-199a-3p Is Downregulated in Human Osteosarcoma and Regulates Cell Proliferation and Migration. Mol Cancer Ther. 2011; 10: 1337–1345. doi: 10.1158/1535-7163.MCT-11-0096 PMID: 21666078

26. Bonet F, Dueñas Á, López-Sánchez C, García-Martínez V, Aránega AE, Franco D. MiR-23b and miR-199a impair epithelial-to-mesenchymal transition during atrioventricular endocardial cushion formation. Dev Dyn. 2015; 244: 1259–1275. doi: 10.1002/dvdy.24309 PMID: 26198058

27. Song J, Gao L, Yang G, Tang S, Xie H, Wang Y, et al. MiR-199a regulates cell proliferation and survival by targeting FZD7. PLoS One. 2014; 9: e110074. doi: 10.1371/journal.pone.0110074 PMID: 25313882

28. Gregory P a, Bert AG, Paterson EL, Barry SC, Tsykin A, Farshid G, et al. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. Nat Cell Biol. 2008; 10: 593–601. doi: 10.1038/ncb1722 PMID: 18376396

29. Humphries B, Yang C. The microRNA-200 family: small molecules with novel roles in cancer development, progression and therapy. Oncotarget. 2015; 6: 6472–98. PMID: 25762624

30. Wu Z-Q, Li X-Y, Hu CY, Ford M, Kleer CG, Weiss SJ. Canonical Wnt signaling regulates Slug activity and links epithelial-mesenchymal transition with epigenetic Breast Cancer 1, Early Onset (BRCA1) repression. Proc Natl Acad Sci U S A. 2012; 109: 16654–9. doi: 10.1073/pnas.1205822109 PMID: 23011797

31. Chen H-C, Zhu Y-T, Chen S-Y, Tseng SCG. Wnt signaling induces epithelial-mesenchymal transition with proliferation in ARPE-19 cells upon loss of contact inhibition. Lab Invest. Nature Publishing Group; 2012; 92: 676–87. doi: 10.1038/labinvest.2011.201

32. Jiang Y-G, Luo Y, He D, Li X, Zhang L, Peng T, et al. Role of Wnt/beta-catenin signaling pathway in epithelial-mesenchymal transition of human prostate cancer induced by hypoxia-inducible factor-1alpha. Int J Urol. 2007; 14: 1034–9. doi: 10.1111/j.1442-2042.2007.01866.x PMID: 17956532

33. Schmidt CM, McKillop IH, Cahill P a, Sitzmann J V. Increased MAPK expression and activity in primary human hepatocellular carcinoma. Biochem Biophys Res Commun. 1997; 236: 54–8. doi: 10.1006/bbrc.1997.6840 PMID: 9223425

34. Huynh H, Nguyen TTT, Chow K-HP, Tan PH, Soo KC, Tran E. Over-expression of the mitogen-activated protein kinase (MAPK) kinase (MEK)-MAPK in hepatocellular carcinoma: its role in tumor progression and apoptosis. BMC Gastroenterol. 2003; 3: 19. doi: 10.1186/1471-230X-3-19 PMID: 12906713

35. Liu L, Cao Y, Chen C, Zhang X, McNabola A, Wilkie D, et al. Sorafenib blocks the RAF/MEK/ERK pathway, inhibits tumor angiogenesis and induces tumor cell apoptosis in hepatocellular carcinoma model PLC/PRF/5. Cancer Res. 2006; 66: 11851–11858. doi: 10.1158/0008-5472.CAN-06-1377 PMID: 17178882

36. Noetel A, Kwiecinski M, Elfimova N, Huang J, Odenthal M. microRNA are Central Players in Anti- and Profibrotic Gene Regulation during Liver Fibrosis. Front Physiol. 2012; 3: 49. doi: 10.3389/fphys.2012.00049 PMID: 22457651

37. Roderburg C, Mollnow T, Bongaerts B, Elfimova N, Vargas Cardenas D, Berger K, et al. Micro-RNA profiling in human serum reveals compartment-specific roles of miR-571 and miR-652 in liver cirrhosis. Lafrenie R, editor. PLoS One. 2012; 7: e32999. doi: 10.1371/journal.pone.0032999 PMID: 22412969

38. Sweetwyne MT, Tao J, Susztak K. Kick it up a notch: Notch signaling and kidney fibrosis. Kidney Int Suppl. 2014; 4: 91–96. doi: 10.1038/kisup.2014.17

39. Morell CM, Strazzabosco M. Notch signaling and new therapeutic options in liver disease. J Hepatol. 2014; 60: 885–90. doi: 10.1016/j.jhep.2013.11.028 PMID: 24308992

40. Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol. Nature Publishing Group; 2011; 18: 1139–1146. doi: 10.1038/nsmb.2115

41. Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol. 2010; 11: R90. doi: 10.1186/gb-2010-11-8-r90 PMID: 20799968

42. de Brevern AG, Meyniel J-P, Fairhead C, Neuvéglise C, Malpertuy A. Trends in IT Innovation to Build a Next Generation Bioinformatics Solution to Manage and Analyse Biological Big Data Produced by NGS Technologies. Biomed Res Int. 2015; 2015: 904541. doi: 10.1155/2015/904541 PMID: 26125026

43. Vicknair C, Macias M, Zhao Z, Nan X, Chen Y, Wilkins D. A comparison of a graph database and a relational database. Proceedings of the 48th Annual Southeast Regional Conference on—ACM SE '10. New York, New York, USA: ACM Press; 2010. p. 1.

44. Jouili S, Vansteenberghe V. An Empirical Comparison of Graph Databases. 2013 International Conference on Social Computing. IEEE; 2013. pp. 708–715.

45. Vlachos IS, Zagganas K, Paraskevopoulou MD, Georgakilas G, Karagkouni D, Vergoulis T, et al. DIANA-miRPath v3.0: deciphering microRNA function with experimental support. Nucleic Acids Res. 2015; 43: W460–6. doi: 10.1093/nar/gkv403 PMID: 25977294

46. Hsu S-D, Lin F-M, Wu W-Y, Liang C, Huang W-C, Chan W-L, et al. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res. 2011; 39: D163–9. doi: 10.1093/nar/gkq1107 PMID: 21071411

47. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. 2009; 37: D105–10. doi: 10.1093/nar/gkn851 PMID: 18996891

48. Sethupathy P, Megraw M, Hatzigeorgiou AG. A guide through present computational approaches for the identification of mammalian microRNA targets. Nat Methods. 2006; 3: 881–6. doi: 10.1038/nmeth954 PMID: 17060911

49. Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG. Lost in translation: an assessment and perspective for computational microRNA target identification. Bioinformatics. 2009; 25: 3049–55. doi: 10.1093/bioinformatics/btp565 PMID: 19789267

50. Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. Nature. 2008; 455: 58–63. doi: 10.1038/nature07228 PMID: 18668040

51. Fan X, Kurgan L. Comprehensive overview and assessment of computational prediction of microRNA targets in animals. Brief Bioinform. Oxford University Press; 2015; 16: 780–94. doi: 10.1093/bib/bbu044

52. Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. Nature. 2014; 505: 344–52. doi: 10.1038/nature12986 PMID: 24429633

53. Yang J, Li T, Gao C, Lv X, Liu K, Song H, et al. FOXO1 3'UTR functions as a ceRNA in repressing the metastases of breast cancer cells via regulating miRNA activity. FEBS Lett. Federation of European Biochemical Societies; 2014; 588: 3218–24. doi: 10.1016/j.febslet.2014.07.003

54. Yuan Y, Liu B, Xie P, Zhang MQ, Li Y, Xie Z, et al. Model-guided quantitative analysis of microRNA-mediated regulation on competing endogenous RNAs using a synthetic gene circuit. Proc Natl Acad Sci U S A. 2015; 112: 3158–63. doi: 10.1073/pnas.1413896112 PMID: 25713348

55. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995; 57: 289–300.

A.3 Publication 3

Preusse M, Marr C, Saunders S, Maticzka D, Lickert H, Backofen R and Theis FJ. *SimiRa: A tool to identify coregulation between microRNAs and RNA-binding proteins.* RNA Biol. 2015;12(9):998–1009.

RESEARCH PAPER

RNA Biology 12:9, 998–1009; September 2015; Published with license by Taylor & Francis Group, LLC

SimiRa: A tool to identify coregulation between microRNAs and RNA-binding proteins

Martin Preusse[1,2], Carsten Marr[1], Sita Saunders[3], Daniel Maticzka[3], Heiko Lickert[2,4], Rolf Backofen[3,5], and Fabian Theis[2,6,*]

[1]Helmholtz Zentrum München – German Research Center for Environmental Health; Institute of Computational Biology; Neuherberg, Germany; [2]Helmholtz Zentrum München – German Research Center for Environmental Health; Institute of Diabetes and Regeneration Research; Neuherberg, Germany; [3]Bioinformatics; Department of Computer Science; University of Freiburg; Freiburg, Germany; [4]Medical Faculty; Technische Universität München; Munich, Germany; [5]BIOSS Center for Biological Signaling Studies; Cluster of Excellence; University of Freiburg; Freiburg, Germany; [6]Technische Universität München; Center for Mathematics; Chair of Mathematical Modeling of Biological Systems; Garching, Germany

Keywords: coregulation, functional similarity, microRNAs, RNA-binding proteins, web application.

Abbreviations: miRNA, microRNA; RBP, RNA-binding protein

microRNAs and microRNA-independent RNA-binding proteins are 2 classes of post-transcriptional regulators that have been shown to cooperate in gene-expression regulation. We compared the genome-wide target sets of microRNAs and RBPs identified by recent CLIP-Seq technologies, finding that RBPs have distinct target sets and favor gene interaction network hubs. To identify microRNAs and RBPs with a similar functional context, we developed simiRa, a tool that compares enriched functional categories such as pathways and GO terms. We applied simiRa to the known functional cooperation between Pumilio family proteins and miR-221/222 in the regulation of tumor supressor gene p27 and show that the cooperation is reflected by similar enriched categories but not by target genes. SimiRa also predicts possible cooperation of microRNAs and RBPs beyond direct interaction on the target mRNA for the nuclear RBP TAF15. To further facilitate research into cooperation of microRNAs and RBPs, we made simiRa available as a web tool that displays the functional neighborhood and similarity of microRNAs and RBPs: http://vsicb-simira.helmholtz-muenchen.de.

Introduction

Post-transcriptional gene regulation

With the discovery of small regulatory RNAs the landscape of gene regulation changed dramatically: It became clear that the abundance of a gene's protein products is not only determined by mRNA processing and the resulting level of mRNA transcripts but also controlled by a whole new layer of regulatory elements.[1]

Post-transcriptional gene regulation has since been associated with almost all biological processes and diseases.[2] MicroRNAs (miRNAs) were the most prominently analyzed species of post-transcriptional regulators but recently microRNA-independent RNA-binding proteins (RBPs) came into focus.[3] Moreover, functional cooperation between miRNAs and RBPs has been shown in various processes such as cancer formation[4] and angiogenesis.[5]

Recent advances in elucidating the functional roles of both classes were supported by new experimental technologies, which extract RNA-protein complexes followed by sequencing of the RNA: HITS-CLIP,[6] PAR-CLIP,[7] iCLIP[8] and CLASH[9] (specific for miRNAs). These methods facilitate global identification of functional binding sites of miRNAs and RNA-binding proteins.

A broad overview of the targeting capabilities is necessary to decipher the complex network of post-transcriptional gene regulation and ultimately define the functional targets of miRNAs and RBPs. Moreover, the global perspective on targeting allows to deduce functional impact beyond regulation of single targets by analyzing effects on functional modules such as signaling pathways.

miRNAs

miRNAs are small endogenous RNAs that bind to target mRNAs and down-regulate the expression by translational repression or degradation of the mRNA.[1,10,11] It has been established that the majority of genes in most eukaryotes are post-transcriptionally regulated by miRNAs.[2] To bind and regulate target mRNAs, miRNAs are first integrated into an AGO protein, which is part of the RNA-induced silencing complex (RISC).

The most important issue in miRNA research is to determine their functional targets. It has been shown that complementary binding between miRNA and target mRNA occurs mostly between nucleotide 2 and 8 of the miRNA (seed region).[1,10-12] CLIP-Seq studies emphasized the importance of the seed region for a significant number of target sites but also demonstrated that non-canonical binding exists and accounts for a significant part of miRNA target sites.[9]

Experimental methods indicated that miRNAs have many (dozens to hundreds) targets and most mRNAs are bound by a

*Correspondence to: Fabian Theis; Email: fabian.theis@helmholtz-muenchen.de

Submitted: 03/24/2015; Revised: 06/16/2015; Accepted: 06/29/2015

http://dx.doi.org/10.1080/15476286.2015.1068496

miRNA at one stage. However, miRNAs regulate their targets only to a small extent and fine-tune protein expression.[13,14] In addition, some parts of the cellular interaction network, such as signaling pathways, are targeted more frequently than others.[15] The dynamics of miRNA-mediated down-regulation change over time[16] and activity of miRNAs depends on the tissue-specific expression of mRNAs[17] and competing binding sites.[18] Thus the complete miRNA-target interaction network is very difficult to predict and the functional classification of miRNAs is still challenging.

RNA-binding proteins

miRNAs are able to guide a functional protein complex to an mRNA target. However, mRNAs interact with a multitude of other miRNA independent RNA-binding proteins during their life cycle from transcription through processing, splicing, relocalization, translation and degradation.[3,19] While the involvement of regulatory proteins in mRNA biogenesis has long been known, CLIP-Seq studies expanded the genome-wide picture of RBP-mRNA interactions and protein occupancy of RNAs.[20-22]

Several hundred proteins are annotated with RNA-binding domains and therefore classified as RBPs.[20,21,23] Interestingly, CLIP-Seq studies identified new RBPs not predicted by protein domains or homology.[19] Many RBPs have thousands of targets although their biological function is not well understood. Similar to the difficulties in determining relevant miRNA targets, the binding mode and potential recognition sequences for RBPs are often not known. Secondary binding determinants such as stabilization by interaction partners or structure of the mRNA have been shown to be important for RBP binding.[24]

Cooperation of miRNAs and RNA-binding proteins

Interaction between miRNA and RBPs occurs via different modes of action: the 2 regulatory partners can act either cooperatively or competitively, directly or indirectly to change expression levels of their target. A cooperative regulation is where both regulatory partners work together, whereas a competitive regulation is where one regulator antagonizes the normal function of the other. A direct regulation occurs when both regulatory partners interact with the target simultaneously (usually with physically close binding sites on the RNA transcript). For the case of direct interactions, several computational studies analyzed the occupancy of mRNAs for proteins and miRNA/AGO complexes and showed that RBPs bind in close proximity to functional miRNA target sites.[25,26] Supporting this notion, there is both computational and experimental evidence that miRNA-binding sites cluster in close proximity leading to increased down regulation of the target mRNA.[27,28]

A well-studied example for a direct interaction with miRNAs are human RNA-binding Pumilio proteins. Downregulation of the tumor suppressor gene p27 by miR-221 and miR-222 has been shown to promote cancer cell proliferation.[29,30] Interestingly, the Pumilio protein PUM1 binds to p27 mRNA, which increases the accessibility of the target site of miR-221/222 by remodeling the mRNA structure.[31,32] Because of low Pumilio levels, quiescent cells have a stable expression of p27 despite high levels of miR-221/222. Thus, both regulators are necessary to promote cancer cell proliferation. The same PUM1 protein has also been shown to bind genes of the pluripotency network in embryonic stem cells (ESC) and facilitate differentiation.[33] Ablation of PUM1 hinders the exit from pluripotency and leads to severe defects in the differentiation process. In addition, there is growing evidence that miRNAs are necessary for ESC differentiation and regulation of the pluripotency network.[34-37] The combined regulation of pluripotency genes is a prime example for possible interactions between miRNAs and RBPs in the fine-tuning of a complex biological process. Moreover, Pumilio proteins have also been shown to be associated with the miRNA-based regulation of the E2F3 oncogenes.[38] There are also examples for competitive regulation where miRNA function is inhibited by RBPs. The RBP Dnd1 has been shown to inhibit the action of miR-21 on its target MSH2 and this regulation has also been implicated in in cancer devlopment.[39] More experimental evidence for miRNA-RBP interactions is reviewed in Ciafre 2013.[40]

An indirect regulation occurs when a previous regulatory effect by one regulatory partner causes a subsequent regulation of target transcript levels by the second regulatory partner. In addition, an important consideration is the cellular location of each regulatory partner. miRNA-mediated regulation always takes place in the cytoplasm, whereas some RBPs can also act in the nucleus. Due to their spatial separation, coregulation between nuclear RBPs and miRNAs must be indirect.

For example, if splicing of a transcript is regulated by RBPs within the nucleus and the same transcript is later regulated by miRNAs, the gene might be indirectly coregulated. Moreover, a RBP can influence the expression level of a miRNA and thereby indirectly affect the expression level of the miRNA's target genes. To our knowledge there is currently no experimental evidence for indirect interaction on the same gene while experimental evidence supports functional regulation of miRNAs by RBPs.[41]

Identification of interaction via functional similarity of miRNA and RBP targets

While some functional interactions might be identified by comparing target sets, most will be difficult to identify due to incomplete targeting data. Even though recent CLIP methods perform better than computational methods, it has been shown that target detection depends on target mRNA expression and binding affinity of the used antibody.[42] Moreover, the size of target sets can vary between replicates.[42]

Methods that analyze the functional context of target sets try to overcome these shortcomings by focusing on biological processes instead of individual target genes. In general, the most widely used techniques to define the functional context of gene sets are GO-term[43] and pathway enrichment.[15,17,44,45] They assume that the over-representation of genes in a pathways or GO term indicates a functional association. Next to enrichment methods, the challenge of deducing biological functions from miRNA/RBP target genes and binding sites was approached by inferring highly regulated targets based on binding site cooperativity[27,28] and integrating miRNA targets with other omics data sets.[46,47]

In this study, we analyzed the combined activity of miRNAs and RBPs to infer functional cooperation between both classes of regulators. We focus on pathway and GO term enrichment to highlight the functional role of miRNAs and RBPs. By comparing the enriched categories for RBPs and miRNAs, we identified regulators with a similar biological function.

To facilitate research into combined action of RBPs and microRNAs, we developed simiRa, a web application that allows to find similar regulators for given input sets of microRNAs and RBPs. It was developed to act as a hypothesis-generator for wet lab scientists that run into common limitations of microRNA research: miRNAs have environment-specific functions and act in concert. To find miRNAs that influence a biological process, over-expression of single miRNAs is usually not sufficient. SimiRa extends the analysis beyond miRNAs and detects similar RBPs which might be necessary for miRNA effects and explain complex functional regulation. SimiRa is available at http://vsicb-simira.helmholtz-muenchen.de.

Results

Dataset

In this study, we used miRNA and RBP target sets identified with biochemical methods based on cross-linking of RNA-protein complexes followed by immunoprecipitation and sequencing (CLIP-Seq). Data for human RBPs was extracted from the doRiNA database[48] and data for human miRNAs from StaRBase v2.[49] Our compiled data set contains 19 RBPs and 366 miRNAs and a total of 14356 unique gene targets. 268 genes are only targeted by RBPs, 1496 are unique for miRNAs and 12592 are targeted by both. In general, we find more targets for RBPs (892 to 7153) than for miRNAs (161 to 1588).

RNA-binding proteins are located in different cell compartments

RBPs can be classified by their cellular localization. In the nucleus, they cannot directly interact with miRNAs on a target mRNA. In the cytoplasm, they can directly cooperate with a miRNA in regulating an mRNA. We analyzed the GO-terms associated with the 19 RBPs in order to elucidate their cellular localization (see **Table 1** for an overview of relevant terms). The selected terms indicate the cellular localization either by biological process (e.g., splicing) or cellular component (nucleus or cytoplasm). We associated the 19 RBPs with their putative role in the mRNA life cycle (**Fig. 1A**). 13 are nuclear while 6 classify as cytoplasmic and all but 3 RBPs have been described in their function (**Fig. 1B**). We found no significant difference in the number of targets between nuclear and cytoplasmic RBPs (Wilcoxon rank-sum test, p-value 0.19).

RNA-binding proteins have distinct target sets

To quantify RBP target-set similarity, we use the Jaccard index (J) defined as the intersection of targets divided by their union (**Fig. 2B**, see methods). Even though some RBPs are characterized as global regulators of splicing (such as TARDBP), the target sets have a Jaccard index between $J = 0.05$ and $J = 0.65$, implying that many RBPs have distinct, non-overlapping target sets. We thus conclude that RBPs are likely to have different functional roles and are in this respect similar to miRNAs.

We performed a hierarchical clustering of the similarity between RBP target sets (**Fig. 2A**). Interestingly, nuclear and cytoplasmic RBPs were not clearly separated with respect to their target genes in the hierarchical tree. The respective groups did not cluster together, and did not show a high overlap of target genes.

Genes are targeted by more RBPs and miRNAs than expected

In order to compare the global targeting properties of RBPs and miRNAs, we analyzed the number of RBPs and miRNAs targeting each gene (**Fig. 3A**). Real target-number distributions were compared to random samplings of targets by constructing artificial target sets following the distributions of targets for real RBPs and miRNAs (see methods).

Interestingly, we find that many genes are targeted by more or fewer RBPs and miRNAs than expected by chance. While random samplings result in 0 to 8 RBPs and 0 to 25 miRNAs per gene (**Fig. 3C**), real RBPs and miRNAs show a wider distribution (**Fig. 3B**). Most importantly, 15% of the genes are targeted by both more than 8 distinct RBPs and more than 25 distinct miRNAs (not counting multiple target sites for a single miRNA/RBP). Genes are targeted by both nuclear and cytoplasmic RBPs. The distribution of targeting RBPs per gene is similar for both groups and the correlation to miRNAs does not change.

We performed a GO term analysis of the 2034 highly targeted genes (targeted by more than 25 miRNAs and more than 8 RBPs) to elucidate the functional role. Among the significantly enriched GO terms are many top-level processes essential for regulatory mechanisms and cell cycle: Chromatin modification (224 associated genes, multiple testing corrected p-value 0.00052), cell cycle (341 associated genes, multiple testing corrected p-value 0.022), protein transport (276 associated genes, multiple testing corrected p-value 0.031), transcriptional regulation (173 associated genes, multiple testing corrected p-value 0.040) and gene expression (512 associated genes, multiple testing corrected p-value 0.042).

RNA-binding proteins prefer to target network hubs

As shown above, many genes are targeted by more RBPs and miRNAs than expected. We hypothesized that highly regulated genes have an important role in the regulatory network of a cell as has been shown before.[50]

We therefore constructed the complete human protein-protein interaction network from the STRING database (**Fig. 3A**), one of the most comprehensive interaction databases.[51] We then calculated the degree (i.e., number of direct neighbors in the network) of all genes in the network and compared it to the number of RBPs (**Fig. 3D**) and miRNAs (**Fig. 3E**) targeting the gene. Interestingly, genes that are targeted by many RBPs have a significantly higher degree (Wilcoxon rank-sum test, p-value = 0, see Methods), while this is not the case for miRNAs.

Table 1. Overview of all RNA-binding proteins and their associated GO terms

RBP	Syn	Entrez ID	Cellular Component	Molecular Function	Biological Process	Ref	Method	# targets
ALKBH5	ABH5	54890	nucleus, nuclear speck	oxidative RNA demethylase activity	mRNA processing, mRNA export from nucleus	Baltz 2012	PAR-CLIP	918
C17ORF85	ELG	55421	—	nucleotide binding	-	Baltz 2012	PAR-CLIP	1419
C22ORF28	RTCB FAAP	51493	cytoplasm, tRNA-splicing ligase complex	ATP binding, metal ion binding, RNA ligase (ATP) activity	tRNA splicing	Baltz 2012	PAR-CLIP	3909
CAPRIN1	M11S1 GPIAP1	4076	cytoplasm, cytoplasmic mRNA processing body, cytosol	RNA binding	regulates translation	Baltz 2012	PAR-CLIP	3891
ELAVL1	HUR	1994	nucleus, nucleoplasm, cytoplasm, cytosol	RNA binding, mRNA binding, protein binding, AU-rich element binding, protein kinase binding, mRNA 3′-UTR AU-rich region binding	mRNA stabilization, positive regulation of translation	Lebedeva 2011	PAR-CLIP	4942
EWSR1	EWS	2130	nucelus, cytoplasm, membrane	RNA binding, protein binding, calmodulin binding, zinc ion binding, metal ion binding	regulation of transcription	Hoell 2011	PAR-CLIP	3400
FUS	TLS	2521	nucleus, nucleoplasm	nucleotide binding, RNA binding, protein binding, zinc ion binding, metal ion binding	RNA splicing, gene expression	Hoell 2011	PAR-CLIP	3981
hnRNPC	C1 C2	3183	nucleus, nucleoplasm, spliceosomal complex, ribonucleoprotein complex	nucleotide binding, protein binding, RNA binding	mRNA processing, RNA splicing, gene expression	König 2012	iCLIP	1428
IGF2BP1	IMP1 ZBP1	10642	nucleus, cytoplasm, cytosol, plasma membrane	RNA binding, protein binding, mRNA 3′-UTR binding	gene expression, regulation of translation, RNA localization, CRD-mediated mRNA stabilization	Hafner 2010	PAR-CLIP	7423
MOV10	gb110 fSAP113	4343	cytoplasm, cytosol, cytoplasmic mRNA processing body	RNA binding, protein binding, ATP binding, hydrolase activity	transcription, DNA-dependent, gene silencing by RNA, mRNA cleavage involved in gene silencing by miRNA	Sievers 2012	PAR-CLIP	3059
PTB	PTBP1 HNRPI	5725	nucleus, nucleoplasm	RNA binding, protein binding, pre-mRNA binding	RNA splicing, mRNA processing, gene expression	Xue 2009	CLIP-seq	1939
PUM2	PUMH2 PUML2	23369	cytoplasm, cytoplasmic stress granule	RNA binding, protein binding	regulation of translation	Hafner 2010	PAR-CLIP	4078
QKI	QK Hqk	9444	nucleus, cytoplasm	RNA binding, protein binding	mRNA processing, RNA splicing, mRNA transport, regulation of translation, mRNA transport	Hafner 2010	PAR-CLIP	1601
SFRS1	ASF SF2	6426	cytoplasm, nucleoplasm, nuclear speck, catalytic step 2 spliceosome	RNA binding, protein binding	gene expression, mRNA processing, mRNA splicing, termination of RNA polymerase II transcription	Sanford 2009	CLIP-seq	6340
TAF15	Npl3 RBP56	8148	nucleus	DNA binding, RNA binding, protein binding	positive regulation of transcription, DNA-dependent	Hoell 2011	PAR-CLIP	2329
TARDBP	ALS10 TDP-43	23435	nucleus	RNA binding, protein binding	RNA splicing, mRNA processing, 3′-UTR-mediated mRNA stabilization	Tollervey 2011	iCLIP	2918
TIA1	WDM TIA-1	7072	nucleus, cytoplasm	nucleotide binding, RNA binding, protein binding, poly(A) RNA binding, AU-rich element binding	negative regulation of translation, regulation of mRNA splicing, via spliceosome	Wang 2010	iCLIP	5217
TIAL1	TCBP TIAR	7073	nucleus, cytoplasm	RNA binding, AU-rich element binding	regulation of transcription from RNA polymerase II promoter	Wang 2010	iCLIP	6938
ZC3H7B	RoXaN	23264	nucleus	protein binding, metal ion binding	virus-host interaction	Baltz 2012	PAR-CLIP	5728

Overview of the 19 RBPs from our compiled data set. The target sets were acquired with different CLIP-Seq methods. GO terms relevant for the cellular localization are shown. RNA-binding proteins have high numbers of targets, ranging from 918 to 7423.

Taking into account that genes belonging to essential processes are more tightly regulated, the preference of RBPs for network hubs suggests that RBPs confer regulatory specificity that augments the more global fine-tuning activity of miRNAs. In summary, the RBP targetome shows evidence for specific regulation of essential biological processes.

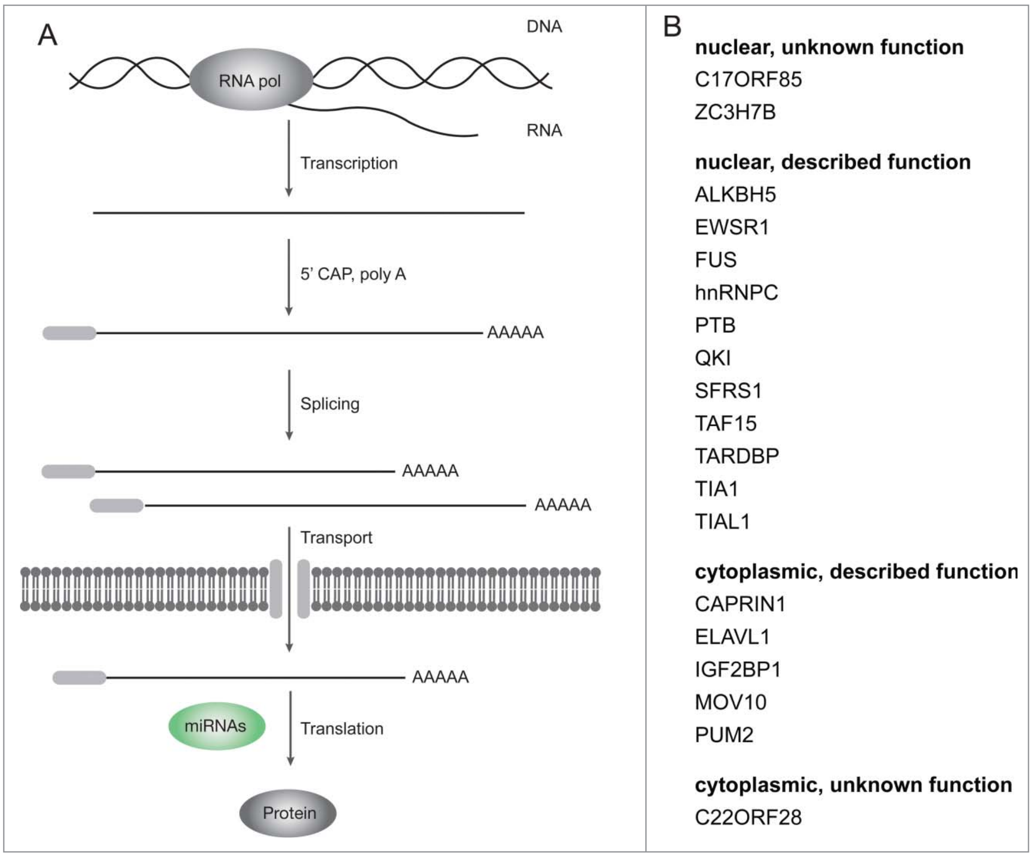


Figure 1. RNA-binding proteins act on all levels of the mRNA lifecycle. (**A**) The mRNA lifecycle from transcription to translation. Multiple steps of processing are necessary to produce mature mRNAs from nascent transcripts. miRNAs interact with their target mRNAs in the cytoplasm. (**B**) The 19 RNA-binding proteins used in this study are located on different levels of the mRNA lifecycle. Their putative localization was inferred from GO term associations and selected publications. 3 of the 19 RBPs have no described function.

SimiRa: miRNA-RBP cooperation revealed by pathway and GO term association

To further analyze combined activity of miRNAs and RBPs within their functional context, we developed simiRa, a web tool that compares not only genes but also functional categories associated to both classes of regulators. By extending the analysis beyond binding of single genes, we are able to capture putative interactions between nuclear and cytoplasmic RBPs that cannot be explained by joint binding of a target mRNA. SimiRa performs an enrichment analysis to find significant functional categories and subsequently compares miRNAs and RBPs (**Fig. 4A**). We used KEGG pathways[52] and GO-terms[43] as functional categories to identify the biological context of miRNA and RBP gene target sets.

We applied our compiled data set of 366 miRNAs and 19 RBPs on 285 KEGG pathways and 40624 GO terms, resulting in 15,749,965 comparisons. Of those, 16,582 are significant (with a multiple testing corrected p-value $<$0.05, see Methods). We compared miRNAs and RBPs by calculating the similarity of target genes and enriched categories using the Jaccard index (intersection divided by union, see **Fig. 2B** and Methods). The scatterplot of all gene similarities against all category similarities is shown in **Figure 4B**. Interestingly, the Pearson correlation between the target similarity and category similarity for all pairwise comparisons of RBPs and miRNAs is high (0.72). While there is a trend toward higher term similarity for increasing gene similarities, many outliers show a high similarity in either genes or terms. The correlation indicates a connection between targets and enriched categories but also highlights the fact that the category enrichment finds similarities that are less likely to be identified by similar target genes.

Since RBPs generally have more targets than miRNAs, the maximum Jaccard index between RBPs and miRNAs is lower than between members of each group. Indeed, distributions of similarities show that RBPs have a higher similarity with other RBPs than with miRNAs (**Fig. 4C**). For miRNA-miRNA and miRNA-RBP comparisons, the median gene similarity is higher than the median category similarity. miRNA-miRNA similarities show a distribution with low median and few very high similarities. Many miRNAs are grouped into families with similar seed-sequences and target binding characteristics (such as miR-221 and miR-222), thus explaining highly similar outliers not found for RBPs.

From the top 100 RBP-miRNA pairs in terms of similar enriched categories, only 53 are also in the top 100 in terms of similar target genes. The other 47 show a disparity between their target gene overlap and enriched functional categories. In summary, comparing enriched functional categories identifies new potential interactions between miRNAs and RBPs that are not obvious from gene targets.

To ease further research into this topic, we made simiRa available as a user-friendly web-tool that allows searching for similar miRNAs and RBPs based on common targets and common

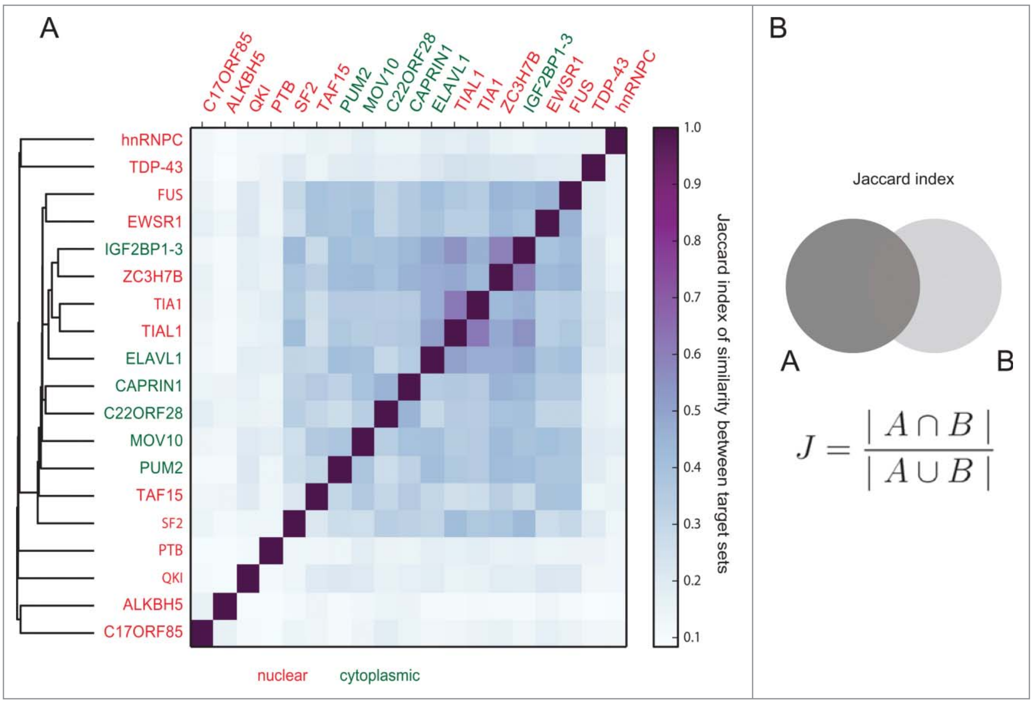


Figure 2. RNA-binding proteins have distinct sets of gene targets. (**A**) Pairwise comparison of the similarity of gene targets of all 19 RNA-binding proteins. While some RNA-binding proteins form clusters with a similarity of J = 0.65, many proteins have distinct target sets. This points toward functional differences and implicates that the proteins take part in different cellular processes. The similarity among nuclear (red) and cytoplasmic (green) RBPs is not different than between groups (Wilcoxon rank sum test, p-value = 0.42). (**B**) The similarity between 2 sets is calculated with the Jaccard index (size of intersection divided by the size of union).

enriched functional categories (**Fig. 5**). The basic workflow starts with the input of an miRNA or RBP. The result is presented as a network of similar miRNAs and RBPs. Search settings for the Jaccard index cut-off can be set individually for gene and category based similarity search. The default settings show term similarities with J > 0.2 and gene similarities with J > 0.3 (see **Fig. 4** for the distributions of Jaccard indexes). The edges of the presented network denote similar gene targets or enriched categories, respectively. The user can change the cutoff for similar miRNAs/RBP, leading to a dense or sparse similarity network.

In a next step, the user can select one or more nodes in the network view to see the targets and enriched categories for the selection. When only one node is selected, all targets/categories are shown. When more nodes are selected, the common targets/categories are shown. This allows for a fine-grained overview of the targeting and functional context for subsets of the similarity network. The network can be extended around single nodes. This gives the user the opportunity to find more interesting candidates.

Case study: the interaction of Pumilio and miR-221/222 is reflected by enriched categories

Pumilio family proteins (e.g., PUM1 and PUM2) are necessary for the regulatory function of miR-221/222 on the tumor suppressor gene p27. Upon binding of PUM1, the binding sites of miR-221 and miR-222 become accessible. PUM2 shows similar effects.[32] This cooperation is a prime example for combined activity of miRNAs and RBPs. The cooperation is not limited to p27: there is evidence for a deeper involvement of both Pumilio proteins and miR-221/222 in the cell cycle misregulation leading to cancer progression.[4,38]

In human, miR-221 and miR-222 have 90% identical targets with a total union of ~1200 targets. The dataset of 19 RBPs contains PUM2 with 4078 targets. PUM2 and miR-221/222 share only 632 target genes, a similarity of J = 0.16 and J = 0.17, respectively (**Fig. 6A**).

In order to compare PUM2 and miR-221/222 to other miRNAs and RBPs, we calculated the similarity of gene targets for all pairs of miRNAs and RBPs using the Jaccard index (see methods). The histogram of the distribution of all pairwise similarities between miRNAs and RBPs shows that most pairs have a Jaccard index < 0.2. Interestingly, PUM2/miR-221 and PUM2/miR-222 are not in the top quartile of miR-RBP pairs. Despite their known functional cooperation, they rank at at the 72.9 percentile of the distribution (**Fig. 6B**). When only considering target sets, PUM2 and miR-221/222 would likely not have been identified as candidates for an interaction. In comparison, the similarity of enriched functional categories between PUM2 and miR-221/222 is higher than the overlap of gene targets. 57 of 192 enriched categories are shared between PUM2 and at least one miRNA. 31 categories are shared by all 3 regulators (**Fig. 6C**). We compared all miRNA and RBP pairs for their overlap in enriched categories. In general, the similarity of enriched categories is slightly higher than for gene targets. Here, PUM2/miR-221 and PUM2/miR-222 are in the top 10% of all pairwise similarities between miRNAs and RBPs (**Fig. 6D**).

Thus, a comparison of functional categories renders PUM2 and miR-221/222 as potential candidates for a functional interaction that would likely be overlooked when only comparing individual target sets. A closer look at the shared categories also highlights the relevance for cancer: We find cancer pathways and signaling cascades commonly functional in the formation of cancer (**Fig. 6E**).

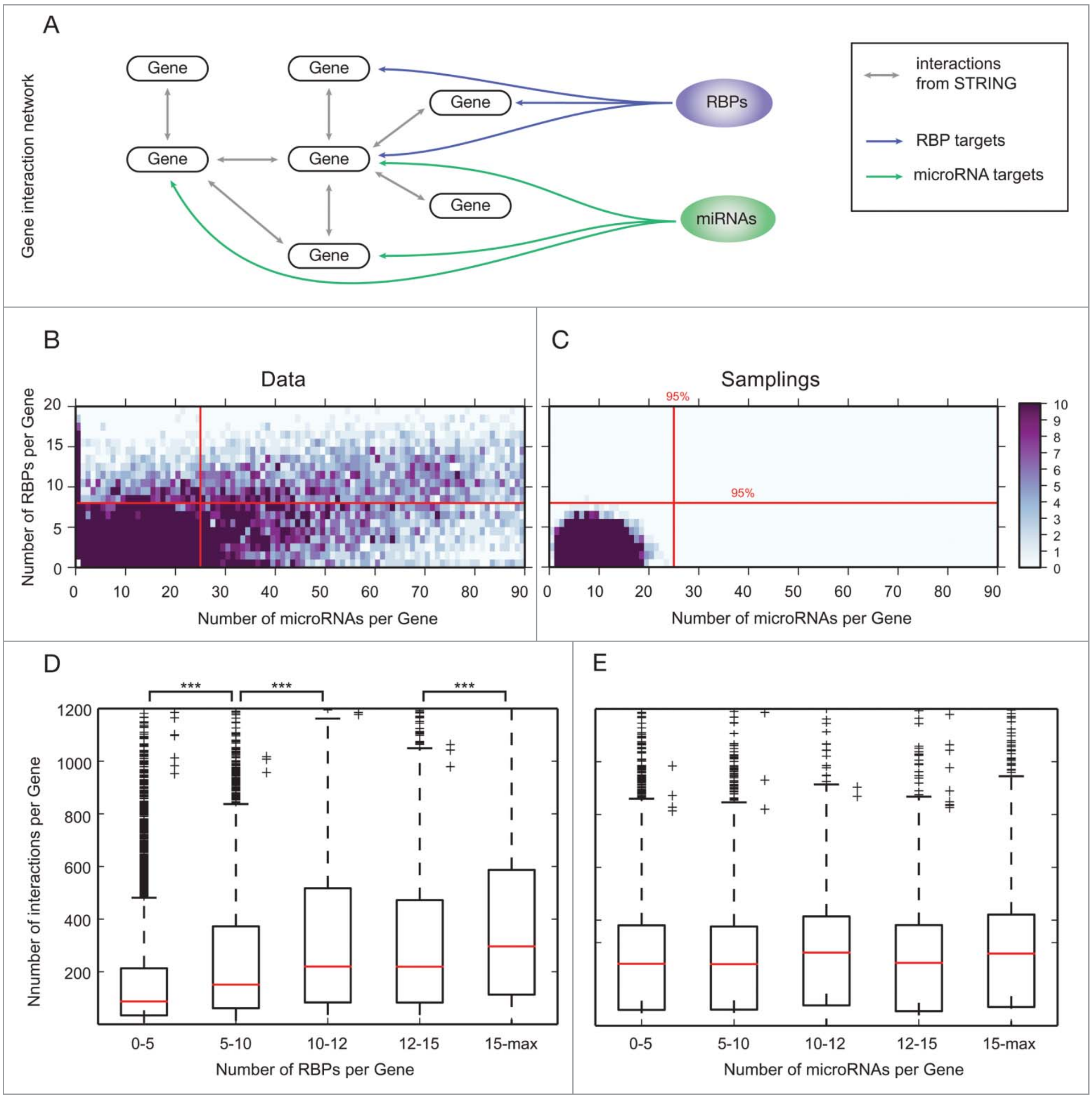


Figure 3. Genes are targeted by more miRNAs and RBPs than expected. (**A**) We mapped the gene target sets of all miRNAs and RBPs in our compiled data set onto a global gene interaction network constructed from STRING. (**B**) Number of targeting miRNAs and RBPs per gene with color coded density. Red lines indicate the 95 percentile from random samplings (**C**). 2034 genes are targeted by more miRNAs and RBPs than expected. Due to the lower number of RBPs in the data set, more genes are targeted only by miRNAs than vice versa. (**C**) Random samplings of gene targets for miRNAs and RBPs. The distribution is more narrow than found for real data. Less genes are targeted by high numbers of miRNAs and RBPs. Red lines show the 95 percentile located at 8 RBPs and 25 miRNAs per gene. (**D**) Network hubs are favored targets of RNA-binding proteins but not miRNAs. Genes were grouped by the number of targeting RBPs and miRNAs, respectively. We counted the number of protein-protein interactions of all genes in the groups. Genes that are targeted by many RBPs show an increased number of network interactions (denoted by ***, one sided Wilcoxon rank sum test, p-value = 0, see Methods). (**E**) For miRNAs, there is no correlation between the number of targeting entities and interactions within the gene interaction network.

Case study: candidates for functional interactions between nuclear RBPs and miRNAs

The TAF15 protein is an interesting candidate for functional analysis: Together with FUS and EWS it constitutes the FET (FUS/EWS/TAF15) protein family[53] that was first discovered as genes frequently translocating in human sarcomas and leukemias.[54] Later, the family members have been shown to participate in the transcriptional machinery as well as various steps of mRNA processing, such as splicing and transport.[55,56] While their exact role remains unclear, recent publications point toward cell-type specific expression and function as well as differences between FUS, EWS and TAF15.[57]

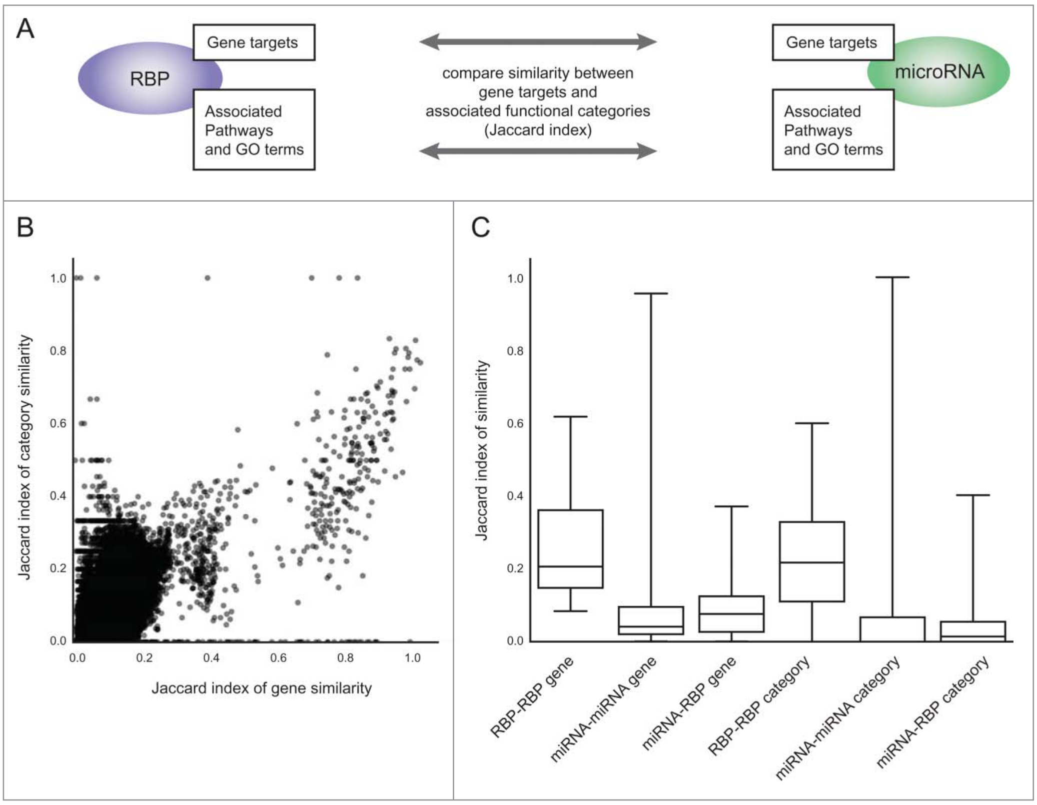


Figure 4. simiRa compares target gene and category similarities of miRNAs and RBPs. (**A**) simiRa compares RBPs and miRNAs based on the similarity (Jaccard index) of significantly enriched functional categories and gene targets. (**B**) Scatterplot of the Jaccard indexes for target gene similarity against category similarity of all pairwise comparisons between miRNAs and RBPs. (**C**) Distributions of pairwise similarities separated by RBP/RBP, miRNA/miRNA and miRNA/RBP comparisons for both target gene and category similarity.

TAF15 is necessary for the cell cycle and proliferation but the mechanism remains elusive. While direct targets have not been validated outside of CLIP-Seq studies, it has recently been reported that TAF15 cooperates indirectly with miR-17-5p and miR-20a-5p to repress the cell-cycle gene CDKN1A/p21 by increasing expression levels of the mature miRNAs, which subsequently downregulate CDKN1A/p21.[41] Upon depletion of TAF15, the levels of the miRNAs decrease, CDKN1A/p21 increases and proliferation is impaired. Again, we found that TAF15 and both miRNAs are similar in terms of enriched categories (~0.22, rank 94%) and less similar in terms of targets (~0.15, rank 83%). Notably, other miRNAs have even higher similarities to TAF15. Those miRNAs are candidates that either collaborate with TAF15 in an indirect fashion like miR-17-5p/miR-20a-5p or they could act on the same targets as TAF15, leading to either cooperative activity or competitive inhibition of the miRNA and TAF15. The top miRNAs showing similar functional categories as TAF15 are miR-590-3p ($J = 0.33$) and miR-495-3p ($J = 0.29$). MiR-495 has been shown to inhibit differentiation of human mesenchymal stem cells[58] and mouse embryonic stem cells,[36] pointing toward a similar regulatory loop as for miR-17-5p/miR-20a-5p. MiR-590-3p, on the other hand, is so far not associated with cell cycle progression and is thus a highly interesting candidate for functional studies in combination with TAF15.

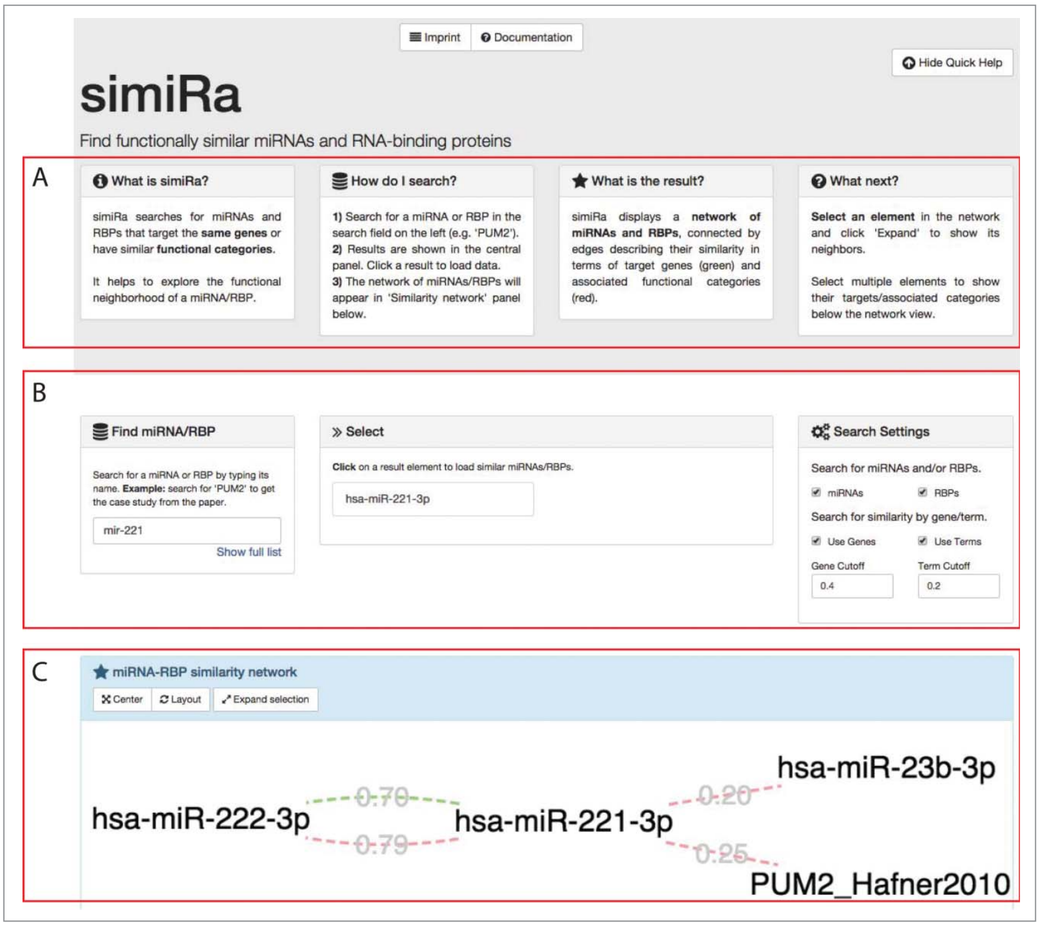
Ballarino et al.[41] found candidate miRNAs for an interaction with TAF15 by manually screening the small set of validated binding sites from miRTarBase.[59] Functionally similar miRNAs identified by our large-scale approach are interesting candidates to extend the TAF15/miRNA interaction network by direct and indirect cooperation.

Discussion

The field of RBPs is growing rapidly since CLIP-Seq studies identified global binding sites. Recently, such an approach identified 300 new and previously uncharacterized RBPs.[21] It is still unclear to what extend RBPs carry out specific regulatory functions. Some RBPs might be housekeeping genes that mostly have a structural role in e.g., transport or decay of mRNAs. To answer this question and provide first insight into global targeting properties, we showed that genes are regulated by very different numbers of RBPs. Moreover, RBPs target network hubs. This indicates that they indeed have a more specific rather than global house-keeping function.

To provide a basis for experiments investigating the combined activity of multiple miRNAs and/or RBPs, we have developed simiRa. The intuitive interface allows for easy exploration of the functional neighborhood of a miRNA or RBP. We expect that most users will start the search with a set of miRNAs/RBPs they are investigating in the biological context of interest. From this starting point, simiRa provides useful candidates for functional cooperation partners which might act in concert to carry out a biological function. For example, Ballarino et al.[41] identified candidate miRNAs for combined activity with TAF15 by manually screening the small set of validated binding sites from miRTarBase.[59] Using our large-scale approach, we are able to identify a lot more potential partners that might function in the same fashion as miR-17-5p and miR-20a-5p.

When comparing PUM2 and miR-221/222, the analysis of enriched functional categories points toward a combined activity in a cancer context that has been shown experimentally. For TAF15, we find miRNAs that might cooperate in an indirect regulatory loop. Considering combinations of multiple miRNAs and RBPs with a similar functional background could prove beneficial in experimental settings where researches look for new regulators of a biological process and single miRNAs did not show the desired effects. By either using more miRNAs or adding RBPs to the experimental set-up, researchers could potentially identify new regulatory elements.



simiRa

Find functionally similar miRNAs and RNA-binding proteins

What is simiRa?

simiRa searches for miRNAs and RBPs that target the **same genes** or have similar **functional categories**.

It helps to explore the functional neighborhood of a miRNA/RBP.

How do I search?

1) Search for a miRNA or RBP in the search field on the left (e.g. 'PUM2').
2) Results are shown in the central panel. Click a result to load data.
3) The network of miRNAs/RBPs will appear in 'Similarity network' panel below.

What is the result?

simiRa displays a **network of miRNAs and RBPs**, connected by edges describing their similarity in terms of target genes (green) and associated functional categories (red).

What next?

Select an element in the network and click 'Expand' to show its neighbors.

Select multiple elements to show their targets/associated categories below the network view.

Find miRNA/RBP

Search for a miRNA or RBP by typing its name. **Example:** search for 'PUM2' to get the case study from the paper.

mir-221

Select

Click on a result element to load similar miRNAs/RBPs.

hsa-miR-221-3p

Search Settings

Search for miRNAs and/or RBPs.

☑ miRNAs ☑ RBPs

Search for similarity by gene/term.

☑ Use Genes ☑ Use Terms

Gene Cutoff: 0.4

Term Cutoff: 0.2

miRNA-RBP similarity network

Figure 5. SimiRa – a web application to identify similar miRNAs and RBPs. (**A**) Introduction and quick help for simiRa is provided on the front page. (**B**) The user starts by searching for an miRNA or RBP in the search field in the 'Find miRNA/RBP' panel on the left. The 'Show full list' button opens a list of all miRNAs and RBPs. A fuzzy search is carried out upon typing of a miRNA/RBP name and results are shown in the 'Select' panel in the center. Clicking on a miRNA/RBP loads the network view of similar miRNAs/RBPs. Settings can be adjusted in the 'Search settings' panel on the right. (**C**) The resulting similar miRNAs/RBPs are displayed in a network visualization in the 'miRNA-RBP similarity network' panel. Similarity in gene targets is indicated by green edges, common enriched categories are denoted by red edges. The user can zoom by scrolling and pan by dragging. Targets and enriched categories of selected nodes are shown below the network panel. The network can be extended by selecting a node and clicking 'Expand selection'. This allows for the stepwise exploration of the functional neighborhood of a miRNA/RBP of interest.

The next step in analyzing combined activity of different post-transcriptional regulators is functional testing: Researches working with miRNAs could benefit from identifying RBPs as potential interaction partners. Cell-type specific miRNA activity has been explained by expression of competing endogenous RNAs (ceRNAs) that fish miRNAs and thereby repress their function on a specific cellular environment.[18] RBPs could be another way of creating tissue-specific effects. If a miRNA requires a RBP to function or if the regulatory effect is increased in the presence of a RBP, the expression of this RBP confers specificity to the miRNA function. Similar to the PUM1/miR-221/222 regulation of p27, RBPs could explain variance in target regulation between different cell types.

We have previously developed miTALOS, a web-tool to analyze the signaling pathways associated to single miRNAs.[17] SimiRa extends the functionality of miTALOS by not only considering a single miRNA and their function but rather allowing to explore the functional neighborhood of a single regulatory component. It thus extends our tool box of miRNA-related applications that aim at providing the functional context of miRNAs and new candidates for functional testing.

The study presented here addresses an unresolved issue: How is the complex process of post-transcriptional gene regulation structured? miRNAs have hundreds of targets and only small effect sizes. A miRNA does not have a unique function but is part of a dense regulatory network whose output depends on the cellular environment. The more we know about the elements and connections within this network, the better our predictions of miRNA function become. By adding RNA-binding proteins to the mix, we extend the regulatory network with a new type of node. Comparing miRNAs and RBPs by their enriched categories takes a step back from individual target relationships and reveals the global picture of miRNA/RBP co-targeting.

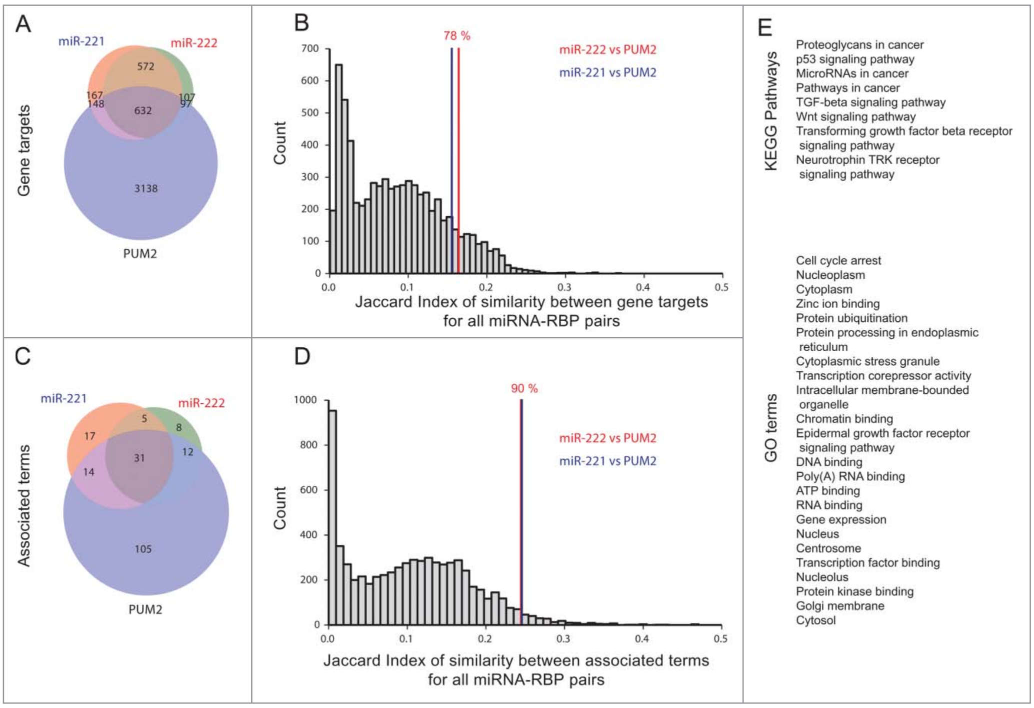


Figure 6. SimiRa case study. The interaction of miR-221/222 and Pumilio is reflected by enriched pathways but not gene targets. (**A**) The overlap of gene targets of miR-221/222 and Pumilio Protein 2 (PUM2) is 632, containing only one fifth of all targets of PUM2. (**B**) The pairwise overlaps of miR-221/PUM2 and miR-222/PUM2 rank at 78% of the overall distribution of miRNA/RBP target similarities. (**C**) When considering enriched terms (Pathways and GO terms), the similarity between miR-221/222 and PUM2 is larger compared to gene targets. (**D**) The pairwise similarities of miR-221/PUM2 and miR-222/PUM2 rank in the top 10%, indicating a functional relationships beyond their gene targets. (**E**) Significantly enriched terms for miR-221/222 and PUM2 (corrected p-value <0.05, see methods). The terms are associated with cancer, cancer signaling and transcriptional activity (terms are sorted by p-value). The genes associated with miR-221/222 and PUM2 can be retrieved from the simiRa web application.

Methods

CLIP-Seq data sets

We used miRNA targets provided by starBase v2.0,[49] a database that collects and integrates CLIP-Seq experiments. We downloaded the complete set of human miRNA target sites with the minimal requirement of one supporting experiment. The data set contains 366 miRNAs with 536888 miRNA-mRNA interactions (i.e., binding sites). RBP binding sites were extracted from the doRiNA database.[48]

We calculated the enrichment of gene sets (miRNA and RBP targets) on gene sets from 285 KEGG pathways[52] and 40624 GO terms.[43] KEGG pathways were obtained via the KEGG REST API (http://www.kegg.jp/kegg/rest/). GO terms were downloaded from http://www.geneontology.org/GO.downloads.ftp.cvs.shtml.

Similarity between miRNAs and RBPs

We define the similarity of 2 non-empty sets A and B using the Jaccard index (number of elements in intersection divided by number of elements in union, [0,1]).

Generation of target set null model

To compare distributions of miRNAs and RBPs targeting genes we sampled artificial target sets from all human genes (as defined in the NCBI Gene database) in the same number as real miRNAs and RBPs in the respective data set. To avoid degree bias, we constructed a bipartite graph linking miRNAs/RBPs to genes and resampled the edges while preserving the degree of miRNA/RBP nodes. Thus, distribution of the number of target per entity resembles real miRNAs and RBPs. We performed 100 sampling runs and averaged over all results.

Protein-protein interaction network

We used protein-protein interaction data from he STRING 9.1 database. Data was downloaded from http://string-db.org. We used all interactions with a combined score >0.75. For a description of the database and score see Von Mering et al.[60] and Szklarczyk et al.[61]

Statistics

Enrichment of a miRNA/RBP (X) in a GO term or pathway (C) was calculated by constructing a 2×2 cross table

	Category C	
miRNA/RBP *X*	*XC*	*Xn*
	Cn	*U*

where *XC* is the number of gene targets of X in *C*, *Cn* is the number of genes in *C* not targeted by *X*, X*n* is the number of targets of X not in *C* and the background *U* is the union of all target genes and all genes in the tested category without *XC*, *Xn* and *Cn*.

The enrichment score *E* is calculated as the odds ratio of *X* and *C*. *E* describes the dependence of variables *X* and C, $E > 1$ indicates an over-representation of targets of *X* in the category *C*:

$$\mathrm{E(X,\ C) = (XC//Cn)//(Xn//U)}$$

P-values for the enrichment were obtained with Fisher's exact test[62] using the 'stats.fisher_exact' module from the SciPy Python package (v0.14.1). To control the false discovery rate (rate of type I errors) in the enrichment analysis, all p-values were corrected with the Benjamini-Hochberg procedure[63] using the 'sandbox.stats.multicomp.multipletests' module from the statsmodels Python package (v0.5.0). Results with a an enrichment sore $E > 1$ and a corrected p-value < 0.05 were considered enriched.

The Wilcoxon rank-sum test was employed to test for difference in distributions of 2 samples,[64] using the 'stats.ranksum' module from the SciPy Python package (v0.14.1). P-values of 0 occur due to occur due to the limits in floating point precision and represent p-values smaller than 10^{-238}.

simiRa web-tool

The simiRa web frontend is implemented with the AngularJS framework and Cytoscape.js for the network view. The backend is implemented in Python using the SciPy stack for calculations and the Flask web framework for the REST API. A neo4j 2.2.2 community edition database is used to integrate data for miRNA/RBP targets and pathways/GO terms.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Funding

This work was funded by the European Research Council (Latent Causes, 259294); the Deutsche Forschungsgemeinschaft (InKoMBio, SPP 1395); the Helmholtz Alliance ICEMED-Imaging and Curing Environmental Metabolic Diseases; the Initiative and Networking Fund of the Helmholtz Association.

References

1. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009; 136(2):215-33; PMID:19167326; http://dx.doi.org/10.1016/j.cell.2009.01.002]
2. Friedman RC, Farh KK-H, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 2009; 19(1):92-105; PMID:18955434; http://dx.doi.org/10.1101/gr.082701.108
3. Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett 2008; 582(14):1977-86; PMID:18342629; http://dx.doi.org/10.1016/j.febslet.2008.03.004
4. Van Kouwenhove M, Kedde M, Agami R. MicroRNA regulation by RNA-binding proteins and its implications for cancer. Nat Rev Cancer 2011; 11(9):644-56; PMID:21822212; http://dx.doi.org/10.1038/nrc3107
5. Chang S-H, Hla T. Gene regulation by RNA binding proteins and microRNAs in angiogenesis. Trends Mol Med 2011; 17(11):650-658; PMID:21802991; http://dx.doi.org/10.1016/j.molmed.2011.06.008
6. Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature 2009; 460(7254):479-86; PMID:19536157
7. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 2010; 141 (1):129-41; PMID:20371350; http://dx.doi.org/10.1016/j.cell.2010.03.009
8. König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol 2010; 17 (7):909-15; PMID:20601959; http://dx.doi.org/10.1038/nsmb.1838
9. Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell 2013; 153 (3):654-65; PMID:23622248; http://dx.doi.org/10.1016/j.cell.2013.03.043
10. Brodersen P, Voinnet O. Revisiting the principles of microRNA target recognition and mode of action. Nat Rev Mol Cell Biol 2009; 10(2):141-8; PMID:19145236; http://dx.doi.org/10.1038/nrm2619
11. Fabian MR, Sonenberg N, Filipowicz W. Regulation of mRNA translation and stability by microRNAs. Annu Rev Biochem 2010; 79:351-79; PMID:20533884; http://dx.doi.org/10.1146/annurev-biochem-060308-103103
12. Thomas M, Lieberman J, Lal A. Desperately seeking microRNA targets. Nat Struct Mol Biol 2010; 17 (10):1169-74; PMID:20924405; http://dx.doi.org/10.1038/nsmb.1921
13. Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. Nature 2008; 455(7209):64-71; PMID:18668037; http://dx.doi.org/10.1038/nature07242
14. Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. Nature 2008; 455 (7209):58-63; PMID:18668040; http://dx.doi.org/10.1038/nature07228
15. Kowarsch A, Marr C, Schmidl D, Ruepp A, Theis FJ. Tissue-specific target analysis of disease-associated microRNAs in human signaling pathways. Morris RJ, ed. PLoS One 2010; 5(6):e11154; PMID:20614023; http://dx.doi.org/10.1371/journal.pone.0011154
16. Hock S, Ng Y-K, Hasenauer J, Wittmann D, Lutter D, Trümbach D, Wurst W, Prakash N, Theis FJ. Sharpening of expression domains induced by transcription and microRNA regulation within a spatio-temporal model of mid-hindbrain boundary formation. BMC Syst Biol 2013; 7:48; PMID:23799959; http://dx.doi.org/10.1186/1752-0509-7-48
17. Kowarsch A, Preusse M, Marr C, Theis FJ. miTALOS: analyzing the tissue-specific regulation of signaling pathways by human and mouse microRNAs. RNA 2011; 17(5):809-19; PMID:21441347; http://dx.doi.org/10.1261/rna.2474511
18. Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. Nature 2014; 505(7483):344-52; PMID:24429633; http://dx.doi.org/10.1038/nature12986
19. Müller-McNicoll M, Neugebauer KM. How cells get the message: dynamic assembly and function of mRNA-protein complexes. Nat Rev Genet 2013; 14 (4):275-87; PMID:23478349; http://dx.doi.org/10.1038/nrg3434
20. Baltz AG, Munschauer M, Schwanhäusser B, Vasile A, Murakawa Y, Schueler M, Youngs N, Penfold-Brown D, Drew K, et al. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Mol Cell 2012; 46(5):674-90; PMID:22681889; http://dx.doi.org/10.1016/j.molcel.2012.05.021
21. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell 2012; 149(6):1393-406; PMID:22658674; http://dx.doi.org/10.1016/j.cell.2012.04.031
22. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature 2013; 499 (7457):172-7; PMID:23846655; http://dx.doi.org/10.1038/nature12311
23. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet 2014; 15 (12):829-845; PMID:25365966; http://dx.doi.org/10.1038/nrg3813
24. Mitchell SF, Parker R. Principles and properties of eukaryotic mRNPs. Mol Cell 2014; 54(4):547-58; PMID:24856220; http://dx.doi.org/10.1016/j.molcel.2014.04.033
25. Jacobsen A, Wen J, Marks DS, Krogh A. Signatures of RNA binding proteins globally coupled to effective microRNA target sites. Genome Res 2010; 20 (8):1010-9; PMID:20508147; http://dx.doi.org/10.1101/gr.103259.109
26. Jiang P, Singh M, Coller H a. Computational assessment of the cooperativity between RNA binding proteins and MicroRNAs in Transcript Decay. PLoS Comput Biol 2013; 9(5):e1003075; PMID:23737738; http://dx.doi.org/10.1371/journal.pcbi.1003075
27. Broderick JA, Salomon WE, Ryder SP, Aronin N, Zamore PD. Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing. RNA 2011:1858-1869; PMID:21878547; http://dx.doi.org/10.1261/rna.2778911
28. Rinck A, Preusse M, Laggerbauer B, Lickert H, Engelhardt S, Theis FJ. The human transcriptome is enriched for miRNA-binding sites located in cooperativity-permitting distance. RNA Biol 2013; 10 (7):1125-35; PMID:23696004; http://dx.doi.org/10.4161/rna.24955
29. Galardi S, Mercatelli N, Giorda E, Massalini S, Frajese GV, Ciafrè SA, Farace MG. miR-221 and miR-222 expression affects the proliferation potential of human prostate carcinoma cell lines by targeting p27Kip1. J Biol Chem 2007; 282(32):23716-24; PMID:17569667; http://dx.doi.org/10.1074/jbc.M701805200
30. Le Sage C, Nagel R, Egan DA, Schrier M, Mesman E, Mangiola A, Anile C, Maira G, Mercatelli N, Ciafrè SA, et al. Regulation of the p27(Kip1) tumor suppressor by miR-221 and miR-222 promotes cancer cell proliferation. EMBO J 2007; 26(15):3699-708; PMID:17627278; http://dx.doi.org/10.1038/sj.emboj.7601790
31. Galgano A, Forrer M, Jaskiewicz L, Kanitz A, Zavolan M, Gerber AP. Comparative analysis of mRNA targets for human PUF-family proteins suggests extensive interaction with the miRNA regulatory system. PLoS One 2008; 3(9):e3164; PMID:18776931; http://dx.doi.org/10.1371/journal.pone.0003164

32. Kedde M, van Kouwenhove M, Zwart W, Oude Vrielink JAF, Elkon R, Agami R. A Pumilio-induced RNA structure switch in p27-3′ UTR controls miR-221 and miR-222 accessibility. Nat. Cell Biol 2010; 12 (10):1014-20; PMID:20818387; http://dx.doi.org/10.1038/ncb2105
33. Leeb M, Dietmann S, Paramor M, Niwa H, Smith A. Genetic exploration of the exit from self-renewal using haploid embryonic stem cells. Cell Stem Cell 2014; 14 (3):385-93; PMID:24412312; http://dx.doi.org/10.1016/j.stem.2013.12.008
34. Gangaraju VK, Lin H. MicroRNAs: key regulators of stem cells. Nat Rev Mol Cell Biol 2009; 10(2):116-25; PMID:19165214; http://dx.doi.org/10.1038/nrm2621
35. Xu N, Papagiannakopoulos T, Pan G, Thomson JA, Kosik KS. MicroRNA-145 regulates OCT4, SOX2, and KLF4 and represses pluripotency in human embryonic stem cells. Cell 2009; 137(4):647-58; PMID:19409607; http://dx.doi.org/10.1016/j.cell.2009.02.038
36. Yang D, Lutter D, Burtscher I, Uetzmann L, Theis FJ, Lickert H. miR-335 promotes mesendodermal lineage segregation and shapes a transcription factor gradient in the endoderm. Development 2014; 141(3):514-25; PMID:24449834; http://dx.doi.org/10.1242/dev.104232
37. Kumar RM. Deconstructing transcriptional heterogeneity in pluripotent stem cells. Nature 2014; 516:55-61; PMID:25471879; http://dx.doi.org/10.1038/nature13920
38. Miles WO, Tschöp K, Herr A, Ji J-Y, Dyson NJ. Pumilio facilitates miRNA regulation of the E2F3 oncogene. Genes Dev 2012; 26(4):356-68; PMID:22345517; http://dx.doi.org/10.1101/gad.182568.111
39. Bhandari A, Gordon W, Dizon D, Hopkin AS, Gordon E, Yu Z, Andersen B. The Grainyhead transcription factor Grhl3/Get1 suppresses miR-21 expression and tumorigenesis in skin: modulation of the miR-21 target MSH2 by RNA-binding protein DND1. Oncogene 2013; 32(12):1497-507; PMID:22614019; http://dx.doi.org/10.1038/onc.2012.168
40. Ciafrè SA, Galardi S. microRNAs and RNA-binding proteins: a complex network of interactions and reciprocal regulations in cancer. RNA Biol 2013; 10(6):934-942; PMID:23696003; http://dx.doi.org/10.4161/rna.24641
41. Ballarino M, Jobert L, Dembélé D, de la Grange P, Auboeuf D, Tora L. TAF15 is important for cellular proliferation and regulates the expression of a subset of cell cycle genes through miRNAs. Oncogene 2013; 32 (39):4646-55; PMID:23128393; http://dx.doi.org/10.1038/onc.2012.490
42. Jungkamp AC, Stoeckius M, Mecenas D, Grün D, Mastrobuoni G, Kempa S, Rajewsky N. In vivo and transcriptome-wide identification of RNA binding protein target sites. Mol Cell 2011; 44:828-840; PMID:22152485; http://dx.doi.org/10.1016/j.molcel.2011.11.009
43. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium Nat Genet 2000; 25(1):25-9; PMID:10802651
44. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009; 37(1):1-13; PMID:19033363; http://dx.doi.org/10.1093/nar/gkn923
45. Ulitsky I, Laurent LC, Shamir R. Towards computational prediction of microRNA function and activity. Nucleic Acids Res 2010; 38(15):e160; PMID:20576699; http://dx.doi.org/10.1093/nar/gkq570
46. Sass S, Dietmann S, Burk U, Brabletz SS, Lutter D, Kowarsch A, Mayer KF, Brabletz T, Ruepp A, Theis FJ, et al. MicroRNAs coordinately regulate protein complexes. BMC Syst Biol 2011; 5(1):136; PMID:21867514; http://dx.doi.org/10.1186/1752-0509-5-136
47. Sass S, Buettner F, Mueller NS, Theis FJ. A modular framework for gene set analysis integrating multilevel omics data. Nucleic Acids Res 2013:1-12; PMID:23143271
48. Anders G, Mackowiak SD, Jens M, Maaskola J, Kuntzagk A, Rajewsky N, Landthaler M, Dieterich C. doRiNA: a database of RNA interactions in post-transcriptional regulation. Nucleic Acids Res 2012; 40 (Database issue):D180-6; PMID:22086949; http://dx.doi.org/10.1093/nar/gkr1007
49. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res 2013:1-6; PMID:23143271
50. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011; 12(1):56-68; PMID:21164525; http://dx.doi.org/10.1038/nrg2918
51. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 2013; 41(Database issue): D808-15; PMID:23203871; http://dx.doi.org/10.1093/nar/gks1094
52. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000; 28(1):27-30; PMID:10592173; http://dx.doi.org/10.1093/nar/28.1.27
53. Bertolotti A, Lutz Y, Heard DJ, Chambon P, Tora L. hTAF(II)68, a novel RNA/ssDNA-binding protein with homology to the pro-oncoproteins TLS/FUS and EWS is associated with both TFIID and RNA polymerase II. EMBO J 1996; 15(18):5022-31; PMID:8890175
54. Riggi N, Cironi L, Suvà M-L, Stamenkovic I. Sarcomas: genetics, signalling, and cellular origins. Part 1: the fellowship of TET. J Pathol 2007; 213(1):4-20; PMID:17691072; http://dx.doi.org/10.1002/path.2209
55. Law WJ, Cann KL, Hicks GG. TLS, EWS and TAF15: a model for transcriptional integration of gene expression. Brief Funct Genomic Proteomic 2006; 5(1):8-14; PMID:16769671; http://dx.doi.org/10.1093/bfgp/ell015
56. Jobert L, Argentini M, Tora L. PRMT1 mediated methylation of TAF15 is required for its positive gene regulatory function. Exp Cell Res 2009; 315(7):1273-86; PMID:19124016; http://dx.doi.org/10.1016/j.yexcr.2008.12.008
57. Andersson MK, Stȧhlberg A, Arvidsson Y, Olofsson A, Semb H, Stenman G, Nilsson O, Aman P. The multifunctional FUS, EWS and TAF15 proto-oncoproteins show cell type-specific expression patterns and involvement in cell spreading and stress response. BMC Cell Biol 2008; 9:37; PMID:18620564; http://dx.doi.org/10.1186/1471-2121-9-37
58. Yang D, Wang G, Zhu S, Liu Q, Wei T, Leng Y, et al. MiR-495 suppresses mesendoderm differentiation of mouse embryonic stem cells via the direct targeting of Dnmt3a. Stem Cell Res 2014; 12(2):550–61.
59. Hsu S-D, Lin F-M, Wu W-Y, Liang C, Huang W-C, Chan W-L, Tsai WT, Chen GZ, Lee CJ, Chiu CM, et al. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res 2011; 39(Database issue):D163-9; PMID:21071411; http://dx.doi.org/10.1093/nar/gkq1107
60. Von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 2005; 33(Database issue):D433; PMID:15608232; http://dx.doi.org/10.1093/nar/gki005
61. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C.et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 2011; 39(Database issue):D561-8; PMID:21045058; http://dx.doi.org/10.1093/nar/gkq973
62. Fisher RA. On the interpretation of χ2 from contingency tables, and the calculation of P. J R Stat Soc 1922; 85(1):87-94; http://dx.doi.org/10.2307/2340521
63. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 1995; 57:289-300
64. Bauer DF. Constructing Confidence Sets Using Rank Statistics. J Am Stat Assoc 1972; 67(339):687.

A.4 Publication 4

Matthes M*, **Preusse M***, Zhang J*, Schechter J, Mayer D, Lentes B, Theis FJ, Prakash N, Wurst W, Trümbach D. Mouse IDGenes: a reference database for genetic interactions in the developing mouse brain. *Database (Oxford). 2014;2014(0): bau083 – bau083.*

DATABASE
The Journal of Biological Databases and Curation

Database, 2014, 1–16
doi: 10.1093/database/bau083
Original article

Original article

Mouse IDGenes: a reference database for genetic interactions in the developing mouse brain

Michaela Matthes[1,2,†], **Martin Preusse**[3,4,†], **Jingzhong Zhang**[1,†], **Julia Schechter**[1], **Daniela Mayer**[1], **Bernd Lentes**[1], **Fabian Theis**[4,5], **Nilima Prakash**[1,*], **Wolfgang Wurst**[1,6,7,8,9,*], **Dietrich Trümbach**[1,6,*]

[1]Institute of Developmental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany, [2]Technische Universität München-Weihenstephan, Lehrstuhl für Genetik, Emil-Ramannstr. 8, 85354 Freising, Germany, [3]Institute of Diabetes and Regeneration Research, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany, [4]Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany, [5]Technische Universität München, Zentrum Mathematik, Boltzmannstr. 3, 85747 Garching, Germany, [6]Max-Planck-Institute of Psychiatry, Kraepelinstr. 2-10, 80804 München, Germany, [7]Deutsches Zentrum für Neurodegenerative Erkrankungen e. V. (DZNE), Standort München, Schillerstr. 44, 80336 München, Germany, [8]Technische Universität München-Weihenstephan, Lehrstuhl für Entwicklungsgenetik, c/o Helmholtz Zentrum München, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany and [9]Munich Cluster for Systems Neurology (SyNergy), Adolf-Butenandt-Institut, Ludwig-Maximilians-Universität München, Schillerstr. 44, 80336 München, Germany

*Corresponding author: Tel: ++49-(0)89-3187-3341; Fax: ++49-(0)89-3187-3099; Email: dietrich.truembach@helmholtz-muenchen.de

Correspondence may also be addressed to Wolfgang Wurst. Tel: ++49-(0)89-3187-4111; Fax: ++49-(0)89-3187-3099; Email: wurst@helmholtz-muenchen.de and Nilima Prakash. Tel: ++49-(0)89-3187-2275; Fax: ++49-(0)89-3187-3099; Email: nilima.prakash@helmholtz-muenchen.de

[†]These authors contributed equally to this work.

Citation details: Matthes,M., Preusse,M., Zhang,J., *et al*. Mouse IDGenes: a reference database for genetic interactions in the developing mouse brain. *Database* (2014) Vol. 2014: article ID bau083; doi:10.1093/database/bau083

Received 27 May 2014; Revised 18 July 2014; Accepted 23 July 2014

Abstract

The study of developmental processes in the mouse and other vertebrates includes the understanding of patterning along the anterior–posterior, dorsal–ventral and medial–lateral axis. Specifically, neural development is also of great clinical relevance because several human neuropsychiatric disorders such as schizophrenia, autism disorders or drug addiction and also brain malformations are thought to have neurodevelopmental origins, i.e. pathogenesis initiates during childhood and adolescence. Impacts during early neurodevelopment might also predispose to late-onset neurodegenerative disorders, such as Parkinson's disease. The neural tube develops from its precursor tissue, the neural plate, in a patterning process that is determined by compartmentalization into morphogenetic units, the action of local signaling centers and a well-defined and locally restricted expression of genes and their interactions. While public databases provide gene expression data with spatio-temporal resolution, they usually neglect the genetic interactions that govern neural development. Here, we introduce Mouse IDGenes, a reference database for genetic interactions in the developing mouse brain. The database is highly curated and offers detailed information about gene expressions and the genetic interactions at the developing mid-/hindbrain boundary. To showcase the predictive power of interaction data, we infer new Wnt/β-catenin target genes by machine learning and validate one of them experimentally. The database is updated regularly. Moreover, it can easily be extended by the research community. Mouse IDGenes will contribute as an important resource to the research on mouse brain development, not exclusively by offering data retrieval, but also by allowing data input.

Database URL: http://mouseidgenes.helmholtz-muenchen.de.

Introduction

Brain formation during vertebrate development is a complex process that has been studied for decades. The understanding of neuronal development is a prerequisite for the fight not only against neurodegenerative diseases, e.g. Parkinson's disease, but also toward neuropsychiatric disorders in particular schizophrenia, autism disorders and drug addiction.

The emergence of the neural tube from the neural plate and the patterning of these structures along their anterior–posterior, dorsal–ventral and medial–lateral axes are fundamental processes during vertebrate neural development. The formation of forebrain, midbrain, hindbrain and spinal cord is determined by well-defined and locally restricted expression of genes and their gene regulatory networks (1). Whereas the patterning of the dorso–ventral axis depends on the relative amounts of dorsalizing and ventralizing factors such as the bone morphogenetic protein (BMP) and Sonic hedgehog (Shh), respectively, the patterning along the anterior–posterior axis is usually accomplished by local signaling centers such as the isthmic organizer (IsO) (2). The IsO, which is necessary and sufficient for the development of mesencephalic and metencephalic structures, is located at the boundary between midbrain and hindbrain and is, therefore, also referred to as the mid-/hindbrain boundary (MHB). The IsO also controls the generation of clinically highly relevant cell populations such as the ventral midbrain dopaminergic neurons, which are involved in Parkinson's disease, schizophrenia and drug addiction, or the rostral hindbrain serotonergic neurons, which take part in mood disorders and depression. Therefore, the MHB or IsO is not only of developmental importance but also of high clinical relevance and thus subject of intense investigations (1–10). Up to now, four stages are thought to be necessary for the development of the MHB: (i) positioning and establishment, (ii) induction, (iii) maintenance and (iv) morphogenesis (1, 2, 7).

Positioning of the future MHB is almost exclusively achieved by the cross-inhibitory interaction of orthodenticle homolog 2 (Otx2) and gastrulation brain homeobox 2 (Gbx2), two transcription factors initially expressed in the anterior and posterior part of the developing embryo, respectively. The inductive mechanism for these two and other factors of the IsO in the neural plate are still unknown. Wingless-type MMTV integration site family member 1 (Wnt1) and fibroblast growth factor 8 (Fgf8) are two factors secreted from the anterior and posterior region of the MHB, respectively. Wnt1 is required for the maintenance of the MHB, and Fgf8 is necessary for the patterning of the midbrain and rostral hindbrain. The engrailed genes En1 and En2 as well as the paired box transcription factors Pax2 and Pax5 act up- and downstream of Wnt1 and Fgf8, mediating their maintenance as well as patterning function at the MHB (1, 2).

Advances in understanding the signaling cascades that give rise to distinct neuronal populations open new prospects for clinical therapies, like stem cell–based treatments. On the other hand, it allows clinicians to classify malformations of the brain more precisely, as with the help of embryology and genetics the major categories of a classification are the causative genes and their pathways and not exclusively the clinical phenotype (8–10).

The gene expression in neural development has been subject to many large-scale studies, and the results were stored in publically available databases. The most important of these resources were recently reviewed (11) and in the following a few will be exemplified. The mouse gene expression database developed by Mouse Genome

Informatics (MGI) is a community resource for gene expression information from the laboratory mouse (12). It is designed as a database to collect and integrate raw expression data from a wide range of sources, such as RNA *in situ* hybridization, immunohistochemistry, western blots, northern blots and RT-PCR. Other databases focus on *in situ* hybridization data: The Allen Developing Mouse Brain Atlas [part of the Allen Brain Atlas (13)], GenePaint.org (14) and the e-Mouse Atlas of Gene Expression (15). Further, there is the Mouse Atlas of Gene Expression, which collects expression data based on serial analysis of gene expression (SAGE) (16). SAGE is more quantitative than *in situ* hybridization but lacks the high spatial resolution. While experimental methods like western and northern blots as well as RT-PCR are not suitable to derive information about exact spatial gene expression (e.g. within single cell populations), the main disadvantage of immunohistochemistry often represents the lack of a functional antibody. Notably, for many of the genes expressed at the MHB suitable antibodies are not available.

To facilitate, for example, dynamic modeling approaches, which necessitate a priori knowledge, i.e. highly curated data, a comprehensive collection of known genetic interactions containing spatial and temporal information is essential (17). These dynamic approaches are means to provide valuable insight into biological problems (17–20). Another interesting field for which integration of interaction and expression data was applied represents the prediction of new transcription factor binding sites (TFBSs) by using statistical models (21, 22). Although, plenty of gene expression data for the developing mouse brain are publicly accessible, interaction databases such as STRING (23), IntAct (24) or BioGRID (25) do not provide high spatial resolution on a developmental time scale. Thus, additional information about the interaction type [IEXP (inferred from experiment and/or expression pattern), 'direct', 'direct signaling', 'indirect', 'indirect signaling' or 'maintenance', which means to keep a gene active/inactive if it was already turned on/off] and mode (i.e. activation or repression) is not yet available for the specific genetic interaction network at the MHB and other brain regions.

We thus developed Mouse IDGenes, which represents a manually curated reference database for genetic interactions in the developing mouse brain focusing on the MHB, but with the possibility to add gene expression and interaction data of the central nervous system (CNS) with the help of a graphical user interface. The freely available database can be accessed via a Web interface through the URL http://mouseidgenes.helmholtz-muenchen.de. The Web interface offers detailed information about the expression of genes and their genetic interactions in the developing mid-/hindbrain region. Stored data were already used in part to understand regulatory gene interactions on the systems level (26, 27). Therefore, the resource provides the possibility for the simulation of the processes occurring at the MHB, which is a unique feature of the presented Web page. The Mouse IDGenes project is conceived for a continuous expansion of stored gene expression and interaction data. Users can enter new data in the database via the Web interface. Currently, 89 spatio-temporally resolved *in vivo* gene expression data sets and 145 genetic interaction data sets from 154 original publications assigned to different anatomical regions at mouse embryonic developmental stages E8.5, E10.5 and E12.5 (Theiler Stages 13, 17 and 20) are available from the database.

Brain regionalization model

We introduced a CNS regionalization model, which covers developmental stages E8.5, E10.5 and E12.5 (Theiler Stages 13, 17, and 20) representing three crucial stages in the development of the murine MHB and mid-/hindbrain region (Figure 1, Supplementary Table S1). The development of the MHB and of the mid-/hindbrain region initiates after gastrulation is finished. At E10.5, the establishment of the IsO at the MHB is completed, and its function at this stage is well characterized, meaning that most of the known interactions are taking place at this time point. The neural tube already exists at this stage, whereas specific neuronal populations have not developed yet. These neuronal populations, however, are first identifiable at around E12.5. Our model about the mouse anatomy is based on data reviewed from literature (28–32) and the MGI database (http://www.informatics.jax.org). To comply with the Edinburgh Mouse Atlas Project (EMAP), ontologies of mouse developmental anatomy, which provides a standard nomenclature for the description of normal and mutant mouse anatomy (33) and, therefore, to allow reusability of the data, we provide EMAP identifiers and descriptions for the defined brain regions (Supplementary Table S1). Because EMAP ontologies were recently updated to the EMAPA ontology (34), we mapped the presented brain regions also to these identifiers (Supplementary Table S1). The regionalization model was kept as general as possible, but as exact as necessary and covers initially the anterior–posterior compartmentalization of the neural tube into the brain vesicles and spinal cord. These brain vesicles correspond to the regions of the brain, which have already developed at a given developmental stage. Within these compartments, we further divided the regions on the anterior–posterior (i.e. longitudinal) axis into lateral and medial or dorsal and ventral parts depending on the corresponding developmental stage. Therefore, we developed a tripartition of a respective

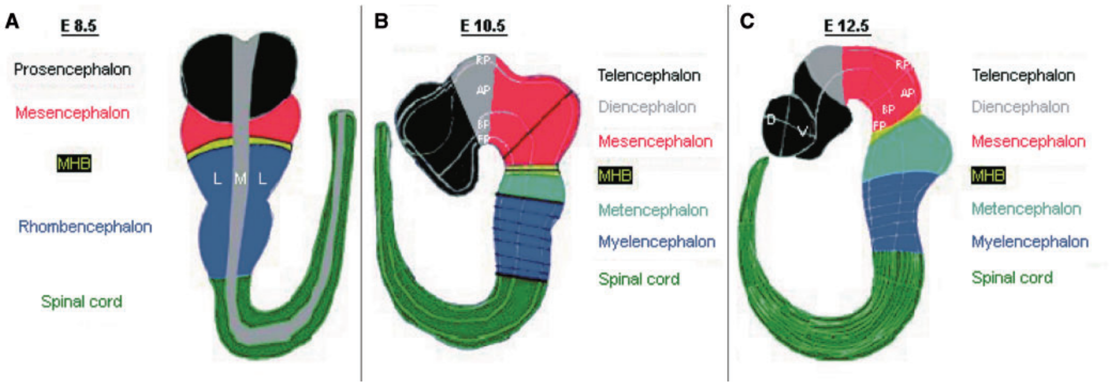


Figure 1. CNS regionalization of the mouse embryo at different developmental stages as used for the database structure. (**A**) Developmental stage E8.5: The whole embryo is divided into five regions along the anterior–posterior axis: prosencephalon, mesencephalon, MHB, rhombencephalon, spinal cord; along the medial–lateral axis the embryo is divided into medial and lateral, and the region in between is considered as mediolateral boundary. (**B**) Developmental stage E10.5: The embryo regionalization along the anterior–posterior and along the dorsal–ventral (previous medial–lateral) axis is as follows: telencephalon, diencephalon, mesencephalon, MHB, metencephalon (r1), myelencephalon (r2-r8) and spinal cord; for all CNS regions, the new regionalization along the dorsal–ventral axis is RP, AP, ABB, BP and FP. (**C**) Developmental stage E12.5: The mouse embryo is regionalized along the anterior–posterior axis as follows: telencephalon (anterior, posterior), diencephalon (anterior hypothalamus, posterior hypothalamus, prethalamus, thalamus), mesencephalon (anterior, posterior), MHB (anterior, posterior), metencephalon (r1), myelencephalon (r2-r8) and spinal cord (cervical, thoracic, lumbar, sacral, caudal); the telencephalon, diencephalon and MHB are subdivided along the dorsal–ventral axis into RP, AP, ABB, BP and FP; the mesencephalon, metencephalon and myelencephalon are subdivided into dorsal and ventral regions and/or neuronal populations along the dorsal–ventral axis; the spinal cord is subdivided into roof plate, dI1, dI2, dI3, dI4, dI5, dI6, v0, v1, v2, v3, v4, mn and floor plate, where dI1 to dI6 describe the dorsal interneurons, and v0 to v3 and mn denote the ventral interneurons (not shown).

CNS region, which has the general structure of vesicle→anterior–posterior localization→medial–lateral or dorsal–ventral localization (column ‘Vesicle’, ‘Structure AP’, ‘Structure ML/DV’ in Supplementary Table S1, respectively). Whenever no further division was made, the description ‘all’ was used.

The CNS regions defined at developmental stage E8.5 are the prosencephalon, mesencephalon, MHB, rhombencephalon and spinal cord. Because the 2D neural plate has not yet folded up to give rise to the 3D neural tube, we only defined an anterior–posterior axis and a medial–lateral axis at this stage. At E10.5, the 3D neural tube undergoes further regionalizations on the anterior–posterior axis as well as on the dorso–ventral (previous medial–lateral) axis. For all CNS regions, the new regionalization on the dorso–ventral axis is roof plate (RP), alar plate (AP), alar basal boundary (ABB), basal plate (BP) and floor plate (FP). The prosencephalon separates along the anterior–posterior axis into the telencephalon and the diencephalon, the developing mesencephalon separates into anterior and posterior, the rhombencephalon splits into met- and myelencephalon and further into eight rhombomeres. The metencephalon was defined as consisting only of rhombomere one (r1).

Because these subdivisions are not defined as sharply as before and distinct neuronal populations have already arisen or are arising in the mid-/hindbrain region of the E12.5 mouse embryo, we additionally defined individual neuronal populations and new subdivisions of the regions (compartments) tel-, di-, mesencephalon, MHB, met- and myelencephalon and also spinal cord. At this stage, the telencephalon as well as the diencephalon show additional subdivisions along the dorsal–ventral axis, and the mesencephalon is subdivided into regions or neuronal populations (or both) also along the dorsal–ventral axis. While at E12.5 the MHB is still subdivided along the anterior–posterior and dorsal–ventral axis, in the met- and myelencephalon as well as spinal cord, dorsal and ventral regions or neuronal populations (or both) are defined along the dorsal–ventral axis. Along the anterior–posterior axis, the diencephalon is now subdivided into anterior hypothalamus, posterior hypothalamus, prethalamus and thalamus. In addition, at E12.5, the developing spinal cord is subdivided into five anterior–posterior regions, namely cervical, thoracic, lumbar, sacral and caudal.

Database and Web page

Mouse IDGenes was implemented as a relational database using PostgreSQL (http://www.postgresql.org). Gene expression and interaction data were manually extracted from literature and stored in the database including references. Currently, the database contains 89 expression data sets and 145 genetic interactions. Genetic interactions as well as expression data sets are assigned to different anatomical regions at the mouse embryonic developmental stages E8.5, E10.5 and E12.5, as described before. To assess the quality of our data, we compared all interactions with the STRING database (v 9.1), a comprehensive resource of protein–protein interactions (23). All interactions but one in the current Mouse IDGenes data set are present in STRING with a high-confidence score (>0.7). The database was made accessible online through a Java Web interface, which was implemented using the Java Servlet class and runs on an Apache Tomcat Server. The Web interface allows the user to browse the Mouse IDGenes database for expression data and genetic interactions. By subscribing to a mailing list, it facilitates communication with other users of the database, as well as with the developers of the database and the Web page. Data can be retrieved in a legible PDF file format and as tab-delimited flat files (by navigating to the 'Download' tab on the Web page). By allowing external user an easy input (using the 'Input Data' tab on the Web page), which will be curated and evaluated by the authors, Mouse IDGenes will be continuously updated and thereby stay an up-to-date research tool.

Detailed search

Navigation on the Mouse IDGenes Web page by the 'Detailed Search' tab allows users to search for gene expression and interaction data in specific regions of the embryonic mouse CNS at mouse embryonic developmental stages E8.5, E10.5 and E12.5 (radio button: 'display genes according to the chosen region'). CNS regions and developmental stages (according to Supplementary Table S1) can be selected with the help of combo boxes. Having set a specific developmental stage and an anatomical structure, users are further able to analyze whether a specific gene of interest is expressed during that specific developmental stage in the selected anatomical structure.

Search for interactions

On the 'Search for Interactions' tab, the Mouse IDGenes Web page allows users to search for specific interactions between two genes of interest and for all interactions in which one specific gene of interest is involved (by the 'Search for specific interactions' button). Information about the region and at which stage the specific interaction takes place in the embryonic mouse CNS can be retrieved.

Users can also search whether two genes of interest display an overlapping gene expression (by the 'Search for overlapping gene expressions' button).

By selecting 'all' in either both or only one of the gene selection boxes, users can search for all stored interactions either of the whole database or in which a specific gene of interest is involved.

The displayed interactions follow an overall scheme consisting of the attributes 'effect', 'type' and 'name' for a genetic interaction. The attribute 'effect' can be either activation, i.e. turning on gene expression, or repression, which is defined as shutting down gene expression. The attribute 'type' of an interaction is defined by the following six values:

- direct: We define a direct interaction in case interaction partner 1 binds directly to the promoter of interaction partner 2.
- direct signaling: In case of a ligand that activates a signaling cascade or any other component of a signaling cascade that does not directly interact with (or binds to) the promoter of a target gene of this signaling pathway, a direct signaling interaction refers to the activation/ repression of a direct target gene of this signaling pathway.
- indirect: Interaction partner 1 does not bind directly to the promoter of interaction partner 2, and signaling or genetic interaction cascades have to occur between the two interaction partners.
- indirect signaling: In case of a ligand that activates a signaling cascade or any other component of a signaling cascade that does not directly interact with (or binds to) the promoter of a target gene of this signaling pathway, an indirect signaling interaction refers to the activation/ repression of an indirect target gene by a direct target gene of this signaling pathway.
- maintenance: Interaction partner 1 is not required to activate (i.e. turn on) or to repress (i.e. turn off) the promoter (or expression) of interaction partner 2, but to keep this promoter (or expression) activated ('on') or repressed ('off') over (a longer period of) time.
- IEXP: Inferred from experiment (e.g. loss of function/ gain of function) and/or expression pattern

Currently, the database contains these general interaction schemes: direct activation, direct signaling activation, IEXP activation, maintenance activation, indirect activation, indirect signaling activation, direct repression, direct signaling repression, maintenance repression, IEXP repression, indirect repression and indirect signaling repression. The attribute 'name' of an interaction, which is displayed on the Mouse IDGenes Web page, is composed of the official gene symbol according to MGI of interaction partner 1, followed by an arrow symbol from the third column of Table 1 and finally the gene symbol of interaction partner 2.

As pointed out before, the maintenance activation occurs over a longer period to cause a downstream effect. Time-wise, such an interaction can occur over several developmental stages as, for example, is the case of the development of midbrain dopaminergic (mDA) neurons. There, initially a Wnt1-regulated network together with a Shh-controlled genetic cascade establishes the mDA

Table 1. General scheme for interactions used at the Mouse IDGenes Web page is listed; interactions are divided into a type, which can be 'direct', 'direct signaling', 'indirect', 'indirect signaling', 'IEXP' and 'maintenance', as well as an effect, namely 'activation' or 'repression' of gene expression

Interaction type	Interaction effect	Symbol
Direct	Activation	- >
Direct signalling	Activation	- >
Indirect	Activation	- - >
Indirect signalling	Activation	- - >
IEXP	Activation	- >
Maintenance	Activation	- >
Direct	Repression	- |
Direct signalling	Repression	- |
Indirect	Repression	- - |
Indirect signalling	Repression	- - |
IEXP	Repression	- |
Maintenance	Repression	- |

Direct: a transcription factor (interaction partner 1) directly binds to the promoter of a gene (interaction partner 2); direct signaling: an activation/repression of a direct target gene of a specific pathway initiated by a ligand that activates the signaling cascade of this pathway; indirect: interaction partner 1 does not bind directly to the promoter of interaction partner 2, and signaling or genetic interaction cascades have to occur between the two interaction partners; indirect signaling: an activation/repression of an indirect target gene through a direct target gene of a specific pathway initiated by a ligand that activates the signaling cascade of this pathway; IEXP: inferred from experiment (e.g. loss of function/gain of function) and/or expression pattern; maintenance: interaction partner 1 is not required to activate (i.e. turn on) or to repress (i.e. turn off) the promoter (or expression) of interaction partner 2 but to keep this promoter (or expression) activated ('on') or repressed ('off') over (a longer) time.

progenitor domain, by maintaining Otx2 expression in the ventral midbrain (35). Another example is Lmx1b, which is known to be necessary for the initiation of Fgf8 expression and for the maintenance of several other genes including Engrailed 1 (En1), En2 and Wnt1 (3). Lmx1b, therefore, falls into the interaction scheme 'maintenance activation' for En1, En2 and Wnt1. On our Web page, these interactions are depicted in the following way: Lmx1b → En1, Lmx1b → En2 and Lmx1b →Wnt1, meaning that Lmx1b keeps the expression of En1, En2 and Wnt1, respectively, on overtime.

Furthermore, cooperative interactions, i.e. interactions with more than two interaction partners, can be stored in the database, for example, if two transcription factors or one transcription factor and another cofactor will bind on the promoter of a target gene, i.e. interaction partner 2.

Data input

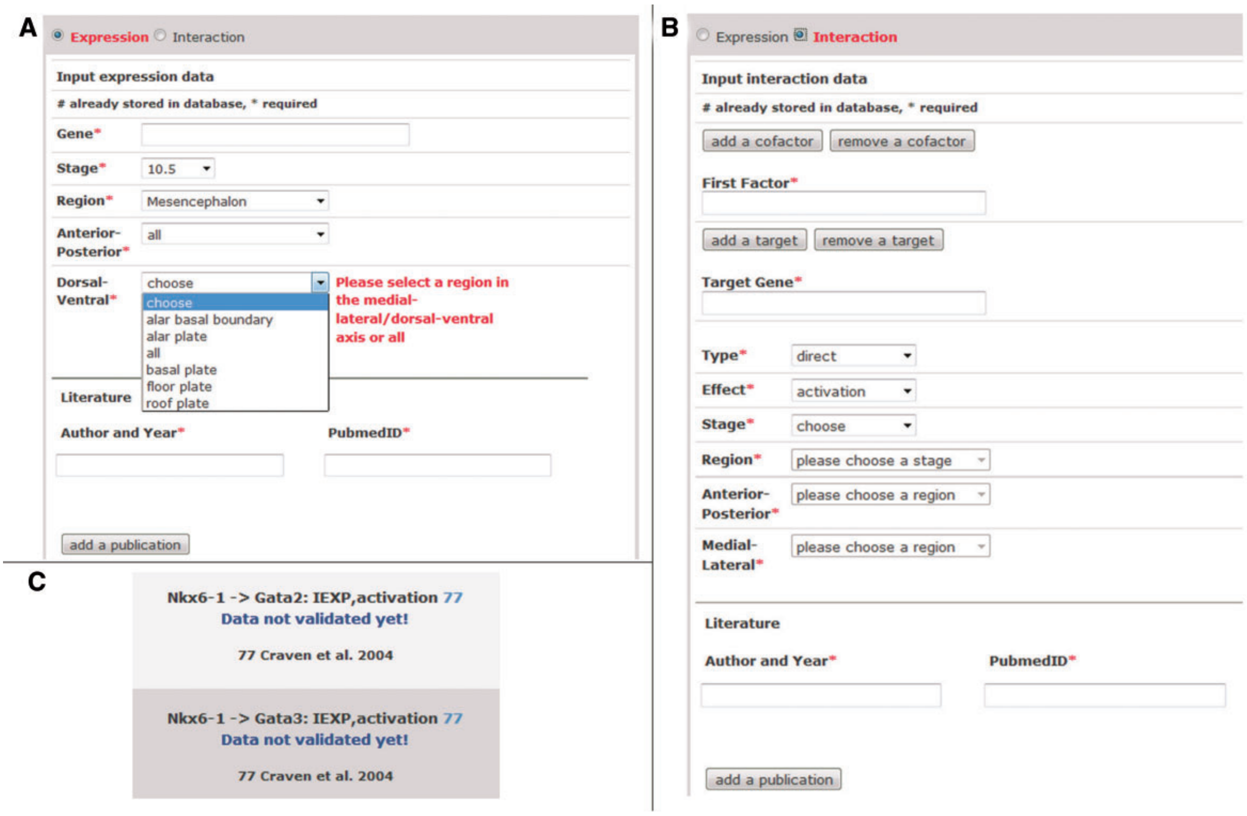
One of the core functions of Mouse IDGenes is the possibility for data input by the user. This feature allows the permanent update of data by the respective experts in the field of developmental neurosciences. To input new gene expression and interaction data into the database, the user is asked to maintain the overall CNS regionalization scheme (according to Supplementary Table S1), by choosing and subsequently storing specifications given by combo boxes, which follow our CNS model (Figure 2A). The database so far has stored data exclusively about the developmental stages E8.5 to E12.5, as these are the crucial stages in the establishment of the MHB and development of the mid-/hindbrain region, which is the authors' main research interest. To maintain the high degree of curation in the current database, it is necessary to also introduce at least one relevant publication as well as the corresponding PubMed ID.

Users can input gene expression data as well as interaction data into the database by selecting the respective type of data on the Web page. All fields that are mandatory for data input as well as storage in the database are labeled by an asterisk (Figure 2A and B), which ensures the completeness of expression and interaction data sets. In addition, it is automatically controlled from the PubMed abstract by using the link to PubMed whether the given year of the publication and the entered PubMed ID are consistent (compare with the fields 'Author and Year' and 'PubmedID' in Figure 2A and B); otherwise, a warning message is displayed, and storage of the data is prevented. Further, gene symbols from MGI are auto-completed after typing some letters in the field 'Gene' (Figure 2A) or 'First Factor' or 'Target Gene' (Figure 2B), and they are internally stored via MGI identifiers. The use of predefined lists by combo boxes for the selection of e.g. a specific brain region and/or the interaction type helps to comply with our brain regionalization model as well as the model for genetic interactions and therefore ensures data consistency. In case of extending the database for an interaction, it is possible to input also cofactors and more targets of an interaction (Figure 2B). After completion of the data set, the user is asked to review the input before final submission to the database. Constraints in the PostgreSQL database prevent data from being duplicated when stored.

To control for incorrect input, the data will be regularly curated by the authors. Before curation, new data will be distinguishable on the Web page from already curated data by labeling the not yet validated gene expression or interaction data on the Web page (Figure 2C).

Confirmation of new data is performed by

- reading the given publications,
- comparing the indicated gene expression and/or genetic interaction,
- comparing the developmental stage as well as the brain regions with the data entered into the Mouse IDGenes database.



Nkx6-1 -> Gata2: IEXP,activation 77
Data not validated yet!
77 Craven et al. 2004

Nkx6-1 -> Gata3: IEXP,activation 77
Data not validated yet!
77 Craven et al. 2004

Figure 2. Dialogue of the data input and output of not yet validated data as found on the Mouse IDGenes Web interface. Input dialogue for (**A**) expression data and (**B**) interaction data. (**C**) The output of not yet validated data.

A mailing list has been set up for users to discuss their plan to submit new data.

After verification of new data, the flag 'Data not validated yet!' is removed from the database and the Web page; otherwise, if the data could not be verified, they are deleted from the database.

Output

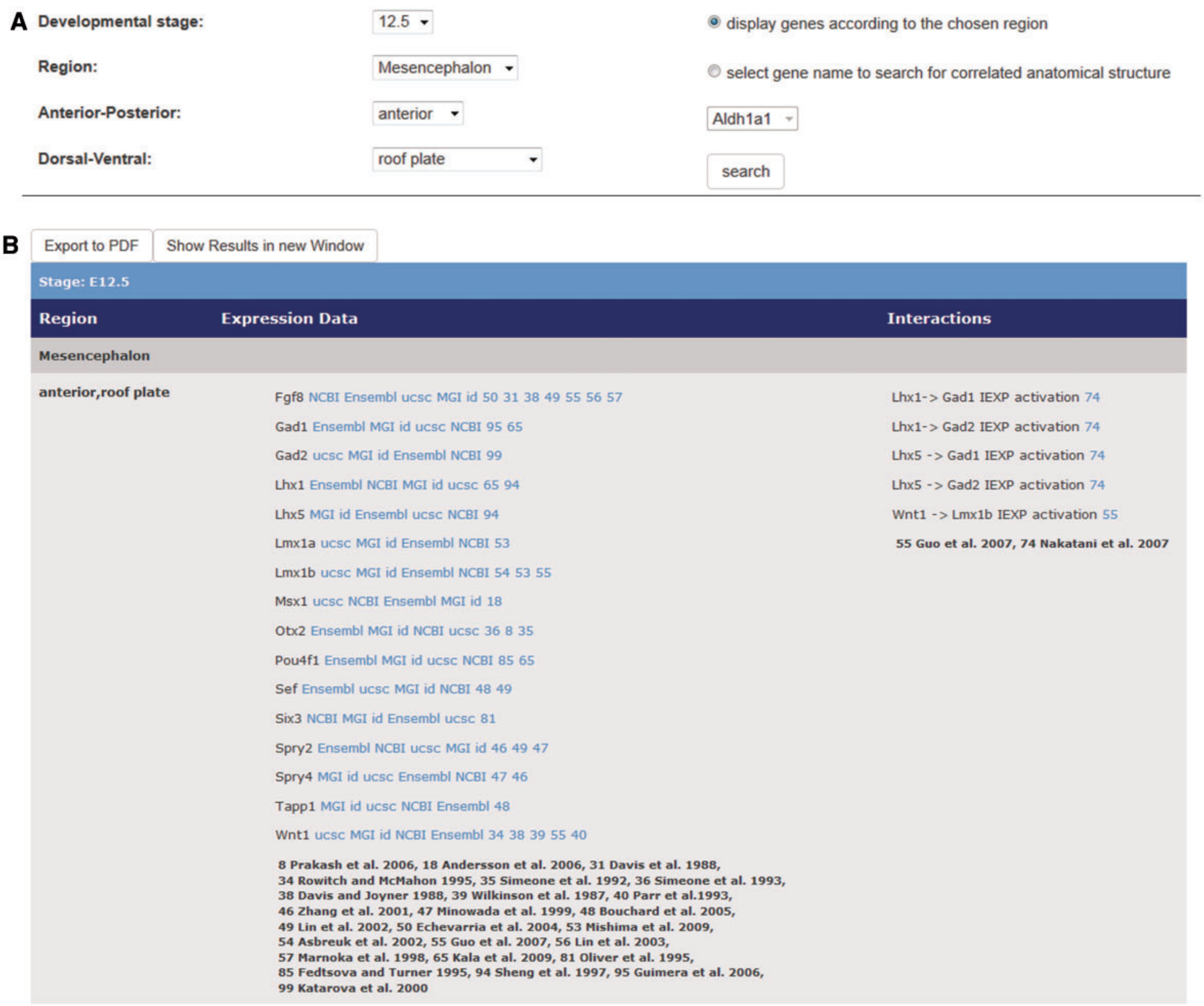
The output of a specific search either in one of the gene expression or the interaction menu items appears on the same page underneath the user selection (Figure 3B), organized in a table. In case of the gene expression data (radio button: 'display genes according to the chosen region' (Figure 3A) in the 'Detailed Search' menu item), the output table indicates the developmental stage the user is looking at. The output further shows the specific subdivision of the chosen anatomical area and which genes are expressed there in the column 'Region'. Additionally, links to the public databases NCBI, MGI, UCSC and Ensembl are given in the column 'Expression Data', where further general information about a displayed gene can be retrieved. Most importantly, links to literature references are indicated as evidence of gene expression information. The last column 'Interactions' of the output shows the interaction data, which can be retrieved for the actual brain region at the actual developmental stage with the corresponding literature references.

Application of the database

Prediction of a new Wnt1 target based on Mouse IDGenes

To demonstrate the usefulness of the Mouse IDGenes database for other research applications, we chose an example from our own scientific interests focused on the role of the Wnt signaling pathway in the development of the mid-/hindbrain region and of neuronal populations located in this region, such as the ventral midbrain dopaminergic neurons. In this context, the Wnt signaling pathway plays a crucial role because it participates in the regulation of regional patterning, cell cycle, cell fate specification, cell differentiation and cell survival. It is also involved in various human diseases (36).

The manually curated, and thus, highly reliable data set of Mouse IDGenes provides an ideal basis to further analyze, for example, the complex gene regulatory network at the MHB and in the ventral midbrain in which Wnt1/β-catenin signaling has so far been implicated (37, 38).



Developmental stage: 12.5

Region: Mesencephalon

Anterior-Posterior: anterior

Dorsal-Ventral: roof plate

● display genes according to the chosen region

○ select gene name to search for correlated anatomical structure

Aldh1a1

Stage: E12.5

Region	Expression Data	Interactions
Mesencephalon		
anterior,roof plate	Fgf8 NCBI Ensembl ucsc MGI id 50 31 38 49 55 56 57	Lhx1-> Gad1 IEXP activation 74
	Gad1 Ensembl MGI id ucsc NCBI 95 65	Lhx1-> Gad2 IEXP activation 74
	Gad2 ucsc MGI id Ensembl NCBI 99	Lhx5 -> Gad1 IEXP activation 74
	Lhx1 Ensembl NCBI MGI id ucsc 65 94	Lhx5 -> Gad2 IEXP activation 74
	Lhx5 MGI id Ensembl ucsc NCBI 94	Wnt1 -> Lmx1b IEXP activation 55
	Lmx1a ucsc MGI id Ensembl NCBI 53	55 Guo et al. 2007, 74 Nakatani et al. 2007
	Lmx1b ucsc MGI id Ensembl NCBI 54 53 55	
	Msx1 ucsc NCBI Ensembl MGI id 18	
	Otx2 Ensembl MGI id NCBI ucsc 36 8 35	
	Pou4f1 Ensembl MGI id ucsc NCBI 85 65	
	Sef Ensembl ucsc MGI id NCBI 48 49	
	Six3 NCBI MGI id Ensembl ucsc 81	
	Spry2 Ensembl NCBI ucsc MGI id 46 49 47	
	Spry4 MGI id ucsc Ensembl NCBI 47 46	
	Tapp1 MGI id ucsc NCBI Ensembl 48	
	Wnt1 ucsc MGI id NCBI Ensembl 34 38 39 55 40	

8 Prakash et al. 2006, 18 Andersson et al. 2006, 31 Davis et al. 1988,
34 Rowitch and McMahon 1995, 35 Simeone et al. 1992, 36 Simeone et al. 1993,
38 Davis and Joyner 1988, 39 Wilkinson et al. 1987, 40 Parr et al.1993,
46 Zhang et al. 2001, 47 Minowada et al. 1999, 48 Bouchard et al. 2005,
49 Lin et al. 2002, 50 Echevarria et al. 2004, 53 Mishima et al. 2009,
54 Asbreuk et al. 2002, 55 Guo et al. 2007, 56 Lin et al. 2003,
57 Marnoka et al. 1998, 65 Kala et al. 2009, 81 Oliver et al. 1995,
85 Fedtsova and Turner 1995, 94 Sheng et al. 1997, 95 Guimera et al. 2006,
99 Katarova et al. 2000

Figure 3. Output window of expression data and interaction data by the use of the 'Detailed Search' option on the Mouse IDGenes Web page. (**A**) Search dialogue on the menu item 'Detailed Search'. In this example, expression and interaction data for the roof plate of the anterior mesencephalon at embryonic day 12.5 are requested. (**B**) Output for gene expression data as well as interaction data for the request according to (A).

We used the Mouse IDGenes database to predict novel direct or indirect targets of the Wnt1/β-catenin signaling pathway that might be involved in the development of the mid-/hindbrain region and of associated neuronal populations and validated our results experimentally.

We obtained known targets with the interaction type 'direct signaling' and 'indirect signaling' of the Wnt1/β-catenin pathway from the Mouse IDGenes Web page by choosing Wnt1 as search term in the 'Search for Interactions' menu as well as from literature searches (Figure 4A) (35, 39–52). One hallmark of the Wnt1/β-catenin signaling pathway is the stabilization and nuclear translocation of cytoplasmic β-catenin (Figure 5). In the absence of a Wnt1 signal, the Lef1/Tcf transcription factors are bound to the promoter regions of the direct Wnt1 target genes together with other co-repressors, thereby inhibiting the activation of these genes. In the presence of a Wnt1 signal, the replacement of these co-repressors and binding of β-catenin to the Lef1/Tcf transcription factors activates the transcription of the direct Wnt1 target genes. The frequent presence of, in particular, evolutionary conserved Lef1/Tcf TFBSs in the promoter region of a gene is therefore indicative that this gene might be a direct target gene of the Wnt1/β-catenin signaling pathway. An indirect target gene of the Wnt1/β-catenin pathway was defined as a gene that is upregulated on Wnt1/β-catenin signaling activity but is not directly bound by β-catenin and Lef1/Tcf transcription factors and thus requires another mediator, i.e. an intermediate gene regulatory step.

We performed an *in silico* promoter analysis for Wnt1 target genes of the interaction type 'direct signaling' and 'indirect signaling' in the training data set (Figure 4A) as

A training set for the SVM classification

Gene Symbol	*Number of conserved binding sites*	*Average matrix similarity*	*Class*	*Ref*
Otx2	2	0.901	1	35;39;40
Lmx1a	1	0.911	1	39;40
En1	0	0.844	1	41-43
Ccnd1	1	0.831	1	44;45;40
Fgf20	7	0.914	1	46
Dkk1	4	0.923	1	46
Sp5	8	0.934	1	47
Msx1	0	0,96	2	39;40
Pitx3	0	0.827	2	39
Gbx2	0	0,8	2	48;49
Shh	1	0.948	2	50
Pou4f1	0	0.903	2	51
Six3	1	0.843	2	52

B test set and predicted classes

Gene symbol	*Number of conserved binding sites*	*Average matrix similarity*	*Predicted class*	*Ref*
Fgf8	0	0.763	2	53;54
Lef1	2	0.933	1	55;56
Islet1	0	0.82	2	57-59
Dkk3	2	0.914	1	*

* Zhang, J. and Prakash, N. et al., unpublished in vivo data

C SVM classification plot

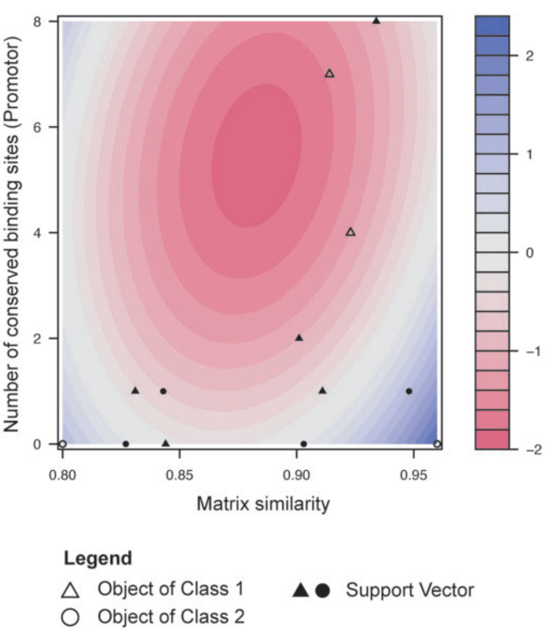


Figure 4. Training and test data for SVM classification. (**A**) Training data containing experimentally validated Wnt1 target genes. Genes were extracted from Mouse IDGenes, and number of conserved binding sites and average matrix similarity were computed with Genomatix MatInspector. Class 1 contains Wnt1 targets of the type 'direct signaling' (i.e. direct Lef1/Tcf target genes), whereas class 2 includes Wnt1 targets of the type 'indirect signaling'. (**B**) Result of classification of selected genes from the test set. Direct Lef1/Tcf targets have more conserved binding sites and a higher average matrix similarity. (**C**) SVM classification (contour) plot showing training data. Filled objects indicate support vectors, blank objects remaining data points. Red color indicates decision values of class 1 (i.e. direct Lef1/Tcf binding), while blue color indicates decision values of class 2 (indirect Lef1/Tcf binding).

well as in the test set (Figure 4B) consisting of interesting target genes for which the interaction type is not yet known in the CNS (53–59). The MatInspector (Genomatix) program was used to identify Lef1/Tcf binding sites in the promoter sequences of these target genes with help of predefined position weight matrices (PWMs) (60). For classification of the interaction type, we assigned 'direct signaling' target genes of Wnt1/β-catenin pathway to class 1 and 'indirect signaling' targets to class 2. We computed two parameters for each gene: (i) the number of evolutionary conserved Lef1/Tcf TFBSs (V$LEFF). Orthologous promoter sequences from human, chimp, mouse, rat, dog, horse, cow and opossum were taken into account, and only binding sites present in at least two species were considered. (ii) The average matrix similarity of all Lef1/Tcf binding sites in the promoter, a score indicating the similarity of a predicted binding site with the consensus matrix of the TF (60, 61). It lies in [0, 1] and reaches 1 only if the predicted sequence corresponds to the most conserved nucleotide at each position. While the matrix similarity allows the assessment of the structural quality of a binding site, the number of conserved binding sites supports the possibility that these binding sites might be functionally relevant (62). It is assumed that in case of direct interactions the matrix similarity is higher and conserved binding sites are more frequent than in case of indirect interactions.

We trained a support vector machine (SVM) with 'direct signaling' targets (class 1) and 'indirect signaling' targets (class 2) using both parameters (Figure 4C). Classification via SVM has been successfully used not only for feature selection of microarray data (63–65) but also to integrate expression as well as genomic data, e.g. evolutionary conservation or binding site clusters for the improvement of TFBS prediction (21, 22). The statistical model was calculated with help of the ksvm function from the kernlab package in R statistical software by using a polynomial kernel matrix similarity (degree = 2), a cost parameter C = 1 and a 13-fold cross validation

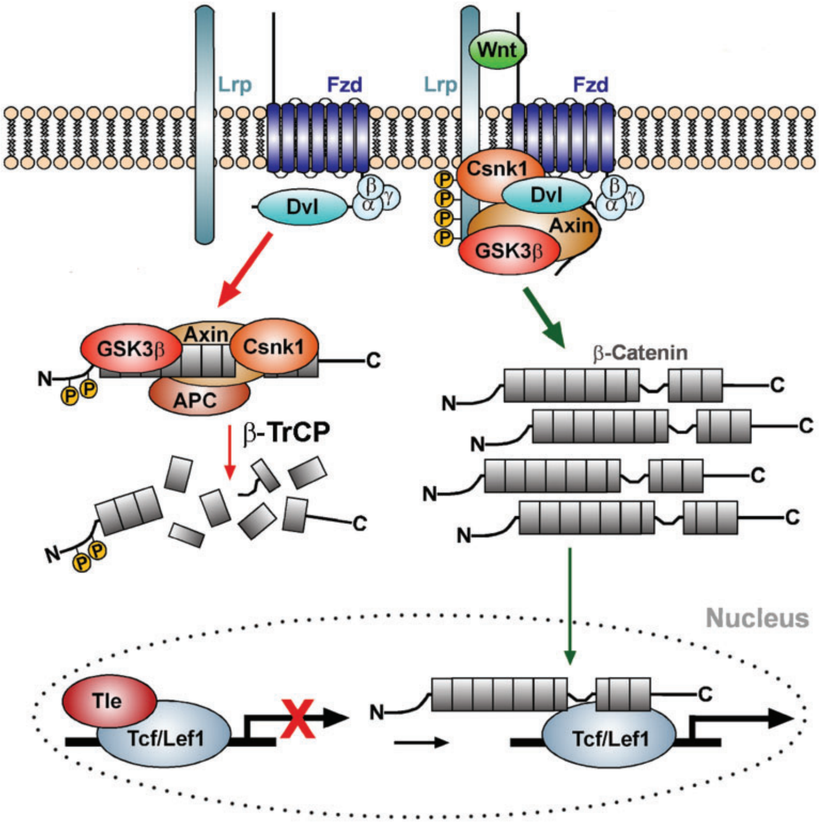


Figure 5. The Wnt/β-catenin signaling pathway. Left (red arrows): In the absence of Wnt ligand, β-catenin is bound by the destruction complex consisting of the scaffolding proteins Axin and Adenomatosis polyposis coli (APC), and the protein kinases Glycogen synthase kinase 3 beta (GSK3β) and Casein kinase I (Csnk1), and sequentially phosphorylated by these kinases. Phosphorylated β-catenin binds to and is ubiquitinated by the E3 ubiquitin ligase β-TrCP, thereby targeting it for proteasomal degradation. In the absence of Wnt ligand, lymphoid enhancer binding factor 1 (Lef1) or T cell-specific (TCF) transcription factors are bound to the promoters of Wnt target genes in the cell nucleus together with co-repressors of the Groucho/transducin-like enhancer of split (Tle) family proteins, thereby repressing their expression. Right (green arrows): On binding of Wnt ligand to the Frizzled (Fzd) receptor and low-density lipoprotein receptor-related protein (Lrp) co-receptor complex, Axin and GSK3β are recruited to the cell membrane via Dishevelled (Dvl) and the destruction complex falls apart. Unphosphorylated β-catenin accumulates in the cytosol and translocates into the nucleus, where it binds to the Lef1/TCF transcription factors and activates Wnt target genes by displacing the co-repressors and recruiting co-activators to this complex. Properties of Lef1/Tcf binding sites in the promoters of known Wnt target genes, i.e. the number of conserved Lef1/Tcf binding sites as well as the averaged matrix similarity, were used to train a classifier and to predict direct or indirect interactions of potentially new target genes and Lef1/Tcf transcription factors in the Wnt/β-catenin signaling pathway.

(leave-ne-out cross validation). Class 1 consists of Otx2, Lmx1a, En1, Ccnd1, Dkk1 and Sp5 and class 2 consists of Msx1, Pitx3, Gbx2, Shh, Pou4f1 (Brn3a) and Six3 (Figure 4A). Additionally, Fgf20 was included in class 1. Interaction data for this gene cannot be retrieved from Mouse IDGenes, as expression of Fgf20 in the mouse neural tube starts only after E12.5 (66). We assigned En1 to class 1 because it was shown that En1 expression under the control of the Wnt1 enhancer in mice rescues the Wnt1-/- mid-/hindbrain phenotype (41, 42), thus indicating that En1 is the downstream target of Wnt1 signaling in mid-/hindbrain development, and because a direct interaction of the promoter of the homologous gene engrailed-2 (En2) with the Lef1/Tcf transcription factor was observed in the frog (43). For our statistical model, we obtained a training error of 7.69% and a cross-validation error of 46.15%.

To further elucidate the Wnt1-controlled gene regulatory network at the MHB, we applied SVM prediction by using the number of conserved Lef1/Tcf binding sites and the average matrix similarity in the promoters of four interesting genes, Fgf8, Lef1, Islet1 and Dkk3, representing the test set (from Figure 4B) to predict whether they are directly or indirectly activated by the Wnt1/β-catenin pathway. For these four genes, it is not known whether they are direct or indirect targets of Wnt1 in neural tissues, but it was observed that Lef1 and Dkk3 are co-expressed with Wnt1 in the midbrain (Götz, S. et al., unpublished data), whereas Fgf8 and Islet1 are not co-expressed with Wnt1 but depend on Wnt1 expression in the mid-/hindbrain region (53, 54, 58, 59). Using the SVM on the test set (Figure 4B), our analysis predicts that Lef1 and Dkk3 are direct targets of Lef1/Tcf-mediated Wnt1/β-catenin signaling, whereas Fgf8 and Islet1 are not direct target genes of this signaling pathway and interact indirectly with the Wnt1 signaling cascade. Fgf8 and Islet1 therefore would represent Wnt1/β-catenin target genes that are most likely activated by other genes that in turn are activated by Wnt1/β-catenin signaling. Direct binding of Lef1/Tcf transcription factors to the Lef1 promoter was shown in colon cancer/lymphocytes (55) as well as in HEK 293 cells (56), which is in accordance with our prediction. A direct interaction of Lef1/Tcf transcription factors with a larger promoter region of the Islet1 gene was shown in embryonic heart tissue but not in neural tissues (57). However, direct activation of Fgf8 and Islet1 in the CNS by Lef1/Tcf binding sites was up to now never observed although it was inferred from mutant mouse embryo analyses (53, 54, 58, 59).

Experimental validation of the predicted Wnt1 target gene Dkk3

To show the predictive power of our approach, we experimentally validated the so far unknown Wnt1 target gene Dkk3 as a direct target gene of the Lef1/Tcf-mediated Wnt1/β-catenin signaling cascade *in vitro*.

To identify conserved Lef1/Tcf binding sites in ~700 bp extended promoter region (as used for the SVM prediction) of the Dkk3 gene of five different mammalian species (mouse, rat, cow, pig and opossum), we applied the DiAlign TF program in the Genomatix software suite GEMS Launcher to evaluate the overall promoter similarity and to identify conserved Lef1/Tcf binding sites in these regions. For the alignment, we chose the five most conserved Dkk3 promoter sequences among 14 organisms from Genomatix homology group Hg3927. Four Lef1/Tcf binding sites were predicted in the putative mouse Dkk3 promoter, of which one [binding site ‘c’, the most proximal Lef1/Tcf binding site to the transcription start site (TSS)] was highly conserved among all five species (Figure 6A). To determine whether these predicted Lef1/Tcf binding sites control the Wnt1/β-catenin and Lef1/Tcf-mediated activation of the murine Dkk3 promoter, we cloned a 744-bp-long fragment of this promoter containing three of the four predicted Lef1/Tcf binding sites into a promoter-less luciferase reporter vector (Figure 6E). Co-transfection of increasing amounts of rat Lef1 complementary DNA (67) or a constitutively active β-catenin [ΔN-β-catenin, which mimics the activation of Wnt1 signaling, (68)] into ‘Wnt-responsive’ HEK293T cells [exhibiting a basal level of Wnt1/β-catenin signaling activity, Prakash, N. et al., unpublished data, (69)] led to a dose-dependent activation of luciferase expression mediated by this mouse Dkk3 (mDkk3) promoter fragment (Figure 6F and G), indicating that the promoter of the mDkk3 gene is a direct target of Lef1-mediated Wnt1/β-catenin signaling in this *in vitro* context. Additional *in vivo* evidence indicates that the murine Dkk3 gene is also a direct target of Lef1-mediated Wnt1/β-catenin signaling in the mouse ventral midbrain (Zhang, J. and Prakash, N., unpublished data).

To evaluate whether the predicted Lef1/Tcf binding sites in the murine Dkk3 promoter are functional, we mutagenized each of these binding sites either individually (mutation of a single Lef1/Tcf binding site) or altogether (mutation of all three Lef1/Tcf binding sites) within the mDkk3 promoter fragment such that they cannot be recognized by Lef1/Tcf transcription factors anymore (Figure 6B–E) (46). Site-directed mutagenesis of single or all three Lef1/Tcf binding sites in the mDkk3 promoter/reporter constructs revealed that

- luciferace activity was significantly decreased relative to the wild-type mDkk3 promoter after co-transfection of Lef1 cDNA (Figure 6H),
- the activation of the mDkk3 promoter/reporter construct carrying a mutagenized Lef1/Tcf binding site ‘c’ by Lef1 was completely abolished relative to the pcDNA3.1 control, in contrast to a still significant activation of the mDkk3 promoter/reporter constructs carrying a mutagenized Lef1/Tcf binding site ‘a’ or Lef1/Tcf binding site ‘b’ (Figure 6I).

This result strongly suggests that the most conserved (by position and sequence similarity among five mammalian species) and proximal (relative to the TSS) Lef1/Tcf binding site ‘c’ is the functionally most important of the three Lef1/Tcf binding sites for Lef1-mediated activation of the murine Dkk3 gene by Wnt1/β-catenin signaling.

Lef1/Tcf transcription factors were predicted to directly activate the mouse Dkk3 gene in the context of Wnt1/β-catenin signaling by our SVM analyses. Therefore, the experimental validation of Lef1/Tcf binding site ‘a’ and Lef1/Tcf binding site ‘c’ with a matrix similarity of 0.867 and 0.961, respectively, confirm the result of our SVM classification, indicating that direct Lef1/Tcf targets have in general more conserved binding sites than indirect targets (Figure 4B). Altogether, our experimental results indicated that the activation of the mouse Dkk3 gene is mediated at least in part by the predicted Lef1/Tcf binding sites in its promoter region and therefore highlight the importance and predictive power of a database combining both expression and interaction data.

Future directions

So far, the Mouse IDGenes database offers spatially resolved and manually registered data about the developmental stages E8.5, E10.5 and E12.5 of the mouse mainly

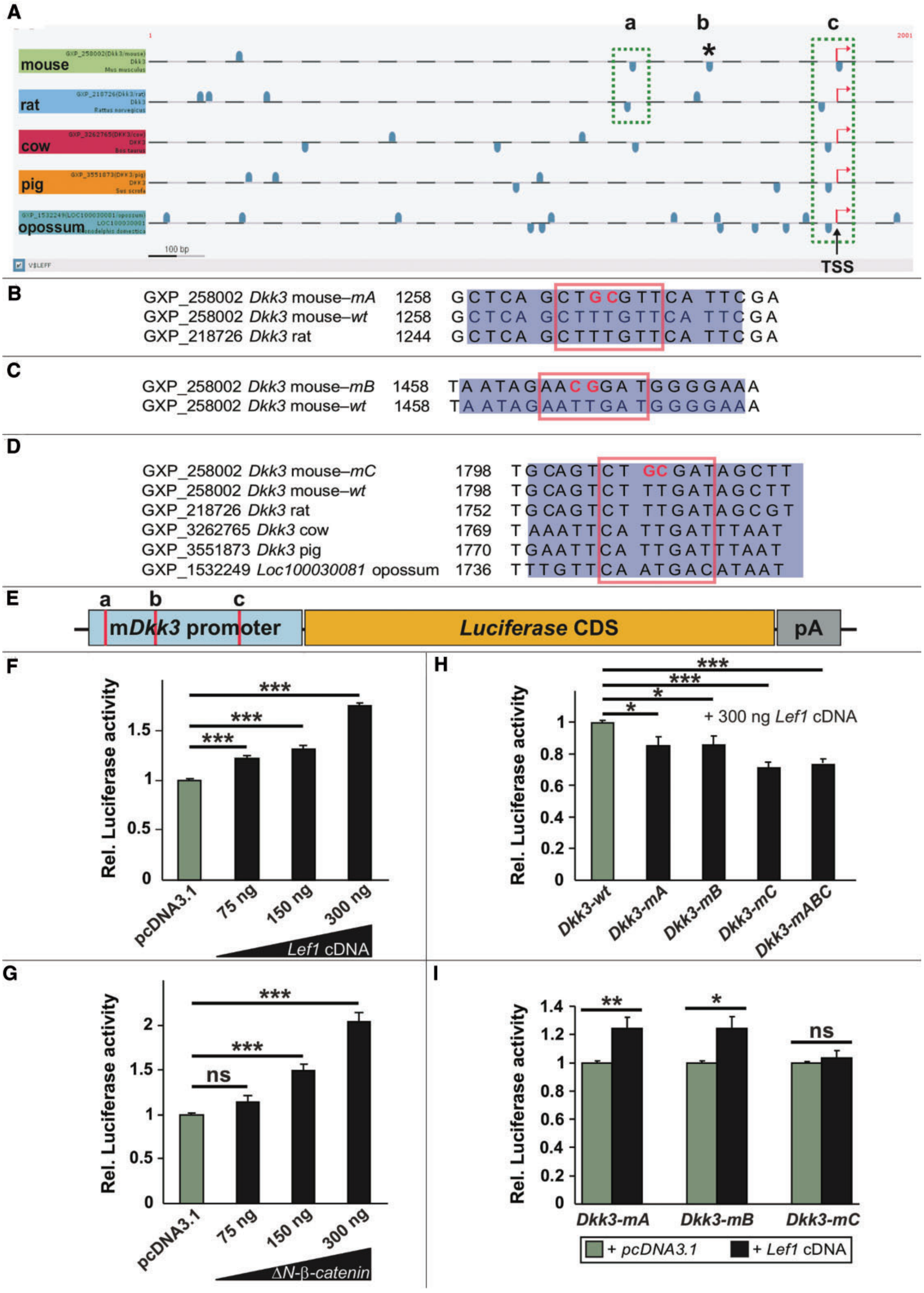


Figure 6. Mouse Dkk3 is a direct target gene of Lef1-mediated Wnt/β-catenin signaling. (**A**) Representation of the putative Dkk3 promoter (5′ proximal) regions from mouse, rat, cow, pig and opossum and of the predicted Lef1/Tcf binding sites on the sense (upper blue boxes) and antisense

(continued)

with gene expressions and interactions important for the development of the mid-/hindbrain region because it harbors important neuronal populations that are implicated in several neurodevelopmental human diseases (1, 2, 7). We aim to enlarge the data sets to more entries that would provide a good representation of the known gene expression patterns and interactions in the developing mouse CNS. This aim is already facilitated by allowing users to input data to the database. With this, we intend to attract experts in different fields of developmental neurosciences to update the existing platform so that Mouse IDGenes becomes the data source of choice, for experimental and *in silico* analyses related to gene expression and interaction data in the developing murine CNS. Additionally, a broader data set will lead to improvement of dynamic modeling projects and to more precise prediction methods (26, 27). As demonstrated, the genetic interaction data stored in the Mouse IDGenes database can be used for the prediction of Wnt target genes, but with this database in combination with publicly available PWMs (70), it is also possible to predict new targets of other signaling pathways (e.g. Shh, Fgf8 or BMPs), which are equally important for CNS development. With our mailing list, we seek to develop an open platform, which eases the communication between neuroscientists.

Materials and methods

Bioinformatics prediction of Lef1/Tcf binding sites in the promoter regions of Wnt/*β*-catenin target genes

To discriminate interactions of the type 'direct signaling', i.e. genes bound by Lef1/Tcf transcription factors activated by Wnt1 signaling from correlations of the type 'indirect signaling' and to predict the interaction type of new target genes, a support vector machine (SVM) was applied. The data matrix of the SVM is composed of two variables (columns) for each object (row), i.e. the frequency as well as the averaged matrix similarity of conserved TFBSs for each promoter sequence of a specific gene. The classification of the interaction type in training data set for each object (gene), meaning whether Lef1/Tcf transcription factor binds directly or indirectly to the promoter sequence, was derived from the Mouse IDGenes database. Promoter sequences for Wnt/*β*-catenin target genes were derived from the ElDorado genome database (Genomatix/Germany), versions 12-2010 and 08-2011. Orthologous promoter sequences from different mammalian species of the Genomatix homology group (human, chimp, mouse, rat, dog, cow, pig and opossum) were analyzed using the MatInspector program (with the Matrix Family Library Version 8.4) from Genomatix to identify potential Lef1/Tcf binding sites and to extract matrix similarities. Conserved binding sites were determined by using the DiAlign TF program (Genomatix) and by searching for common Lef1/Tcf sites occurring at the same position in (aligned) orthologous promoter sequences of each Wnt/*β*-catenin target gene. The length of the promoter regions used for the detection of Lef1/Tcf sites to derive the total number and the average matrix similarity of binding sites both included in the SVM algorithm were generally in the range of 600–1400 bp. In addition, a longer promoter region of mouse Dkk3 mRNA (with the Genbank identifier AK013054, firstly detected in whole body of 10- and 11-day-old mouse embryos) was defined as 1800 bp upstream, including the proximal region, and 200 bp downstream of the TSS. Dkk3 promoter sequences of 2000 bp length from

(lower blue boxes) strands within these Dkk3 promoter regions. The most conserved (by sequence similarity and position) and proximal (relative to the TSS) predicted Lef1/Tcf binding sites were designated as 'a' and 'c' (green dotted boxes). The Lef1/Tcf binding site 'b' is only predicted in the mouse Dkk3 promoter (asterisk). (**B–D**) Sequence alignments of the mutated (m) and wild-type (wt) Lef1/Tcf binding site 'a' in the mouse and rat Dkk3 promoter regions (B), 'b' in the mouse Dkk3 promoter region only (C) and 'c' in the mouse, rat, cow, pig and opossum Dkk3 promoter regions (D). The blue rectangles delimit the sequence, and the red boxes frame the core sequence of the corresponding Lef1/Tcf binding site. The red bold letters indicate the mutagenized nucleotides in the corresponding Lef1/Tcf binding site core sequence. (**E**) Schematic drawing of the murine Dkk3 (mDkk3) promoter/luciferase reporter construct used in the following experiments, and of the approximate position of the three predicted and partly conserved proximal Lef1/Tcf binding sites within this promoter fragment (red bars). CDS, coding sequence; pA, polyadenylation signal. (**F–I**) Luciferase reporter assays in HEK293T cells using the wild-type and mutated mDkk3 promoter/reporter construct depicted in (E). (F) Co-transfection of increasing amounts of Lef1 cDNA led to a dose-dependent activation of the wild-type mDkk3 promoter relative to the 'empty' (pcDNA3.1) vector control. (Rel. luciferase activities: pcDNA3.1, 1.0 ± 0.01; 75 ng Lef1 cDNA, 1.21 ± 0.03; 150 ng Lef1 cDNA, 1.32 ± 0.035; 300 ng Lef1 cDNA, 1.74 ± 0.04). (G) Co-transfection of increasing amounts of a constitutively active *β*-catenin (ΔN-*β*-catenin) led to a dose-dependent activation of the wild-type mDkk3 promoter relative to the 'empty' (pcDNA3.1) vector control. (Rel. luciferase activities: pcDNA3.1, 1.0 ± 0.01; 75 ng ΔN-*β*-catenin, 1.14 ± 0.07; 150 ng ΔN-*β*-catenin, 1.49 ± 0.06; 300 ng ΔN-*β*-catenin, 2.04 ± 0.11). (H) Site-directed mutagenesis of single and of all three predicted Lef1/Tcf binding sites ('a', 'b', 'c', 'abc') within the mDkk3 promoter fragment (Dkk3-mA, Dkk3-mB, Dkk3-mC, Dkk3-mABC) resulted in a site-specific and significant decrease of luciferase activity relative to the wild-type mDkk3 promoter (Dkk3-wt) after co-transfection of 300 ng Lef1 cDNA. (Rel. luciferase activities: Dkk3-wt, 1.0 ± 0.01; Dkk3-mA, 0.85 ± 0.05; Dkk3-mB, 0.86 ± 0.06; Dkk3-mC, 0.71 ± 0.04; Dkk3-mABC, 0.73 ± 0.03). (I) Site-directed mutagenesis of the most conserved (across species) and proximal (relative to the TSS) Lef1/Tcf binding site 'c' in the mDkk3 promoter completely abolished the activation of this promoter after co-transfection of 300 ng Lef1 cDNA relative to the 'empty' vector control (pcDNA3.1). (Rel. luciferase activities: Dkk3-mA: pcDNA3.1, 1.0 ± 0.01; Lef1 cDNA, 1.24 ± 0.08; Dkk3-mB: pcDNA3.1, 1.0 ± 0.01; Lef1 cDNA, 1.24 ± 0.08; Dkk3-mC: pcDNA3.1, 1.0 ± 0.01; Lef1 cDNA, 1.04 ± 0.05). **P* < 0.05; ***P* < 0.01; ****P* < 0.001; ns, not significant.

five different mammalian species (mouse, rat, cow, pig and opossum) were analyzed with the MatInspector to predict Lef1/Tcf binding sites.

Cloning of a mouse Dkk3 (mDkk3) promoter/reporter vector

A 744-bp fragment of the putative mDkk3 promoter (Entrez Gene ID: 50781; chromosome 7, strand: –, position: 112 158 266 to 112 159 009 bp, NCBI build 38) was amplified from C57BL/6 mouse genomic DNA by PCR using the forward primer 5′-*ctcgag*TGACCAGATCCAGCTTGCA-3′ and reverse primer 5′-*aagctt*CCTCCTGAGGGTAGTTGAGA-3′ that included an *Xho*I and *Hind*III restriction site (underlined sequences in italics), respectively. The amplified fragment was cloned into the pCR®II TOPO TA vector (TOPO® TA Cloning® Kit, Life Technologies/Germany) and sequenced throughout its entire length (Sequiserve/Germany). The mDkk3 promoter fragment was excised from the pCR®II TOPO TA vector by *Xho*I/*Hind*III digestion and subcloned into an *Xho*I/*Hind*III-digested pGL3-Basic Vector (Promega/USA).

Site-directed mutagenesis of the mDkk3 promoter fragment

Site-directed mutagenesis of the most conserved and proximal (relative to the TSS) Lef1/Tcf binding sites predicted in the 744-bp mDkk3 promoter fragment was done using the QuickChange Lightning Multi Site-Directed Mutagenesis Kit (Agilent Technologies/USA) according to the manufacturer's instructions. Mutagenic primers were the following: Dkk3-mA: 5′-ccagcttgcagctcag*ct**gc**gtt*cattcgaattgggtg-3′; Dkk3-mB: 5′-gtccaagagatcccagtaatag*aa**cg**gat*ggggaaatagtaaaggaa-3′; Dkk3-mC: 5′-ggtggtcctgcagtc*t**gc**gat*agctttccgggac-3′ (mutagenized nucleotides in bold; core sequence of the corresponding Lef1/Tcf binding site in italics). Mutated promoter fragments were confirmed by sequencing (Sequiserve).

Cell culture, transfections and luciferase reporter assays

HEK-293T cells were kept at 37°C and 5% CO_2 in DMEM medium + 10% fetal calf serum/glutamine (Life Technologies). HEK-293T cells (1.25×10^5 cells/well of a 24-well plate) were co-transfected with 300 ng/well pGL3-mDkk3 promoter/reporter vectors (wild-type and mutagenized sequences), 30 ng/well pRL-SV40 (as internal transfection control, Promega) and 150–225 ng/well pcDNA3.1 (Life Technologies) 'empty' vector, alone or together with 75 ng/well, 150 ng/well or 300 ng/well constitutively active ΔN-*β*-catenin (68) or rat Lef1 cDNA (67) using Lipofectamine 2000 (Life Technologies). The total amount of plasmid DNA transfected in each well was 630 ng. Cells were lysed in Passive Lysis Buffer (Promega) after 24 h, and Firefly and Renilla Luciferase luminescence were measured in a Centro LB 960 luminometer (Berthold Technologies/Germany) using the Dual-Luciferase® Reporter Assay System (Promega) according to the manufacturer's instructions. Firefly luminescence was normalized against Renilla luminescence for each well. Assays were performed in triplicates, and data are derived from three independent experiments.

Statistical analyses

All values shown are mean ± SEM. Statistical significance between groups was assessed by two-tailed independent-samples *t* tests using the SPSS 18.0 software (SPSS Inc./USA). A value of $P < 0.05$ was considered significant.

Supplementary Data

Supplementary data are available at *Database* Online.

Acknowledgments

The authors thank Sebastian Götz for extensive testing of the current version of the database. Their genuine thanks go further to Dr Regina Peis for her help regarding the *in silico* promoter analysis.

Funding

The 'Deutsche Forschungsgemeinschaft' ('German Research Foundation') within the framework of the Munich Cluster for Systems Neurology [EXC 1010 SyNergy]; the EU grant 'Systems Biology of Stem Cells and Reprogramming' [FP7-Health-F4-2010-242129]; the 'Bundesministerium für Bildung und Forschung' in the consortium 'From Disease Genes to Protein Pathways' [FKZ01GS0858]. Funding for open access charge: Helmholtz Zentrum München – Deutsches Forschungszentrum für Gesundheit und Umwelt (HMGU).

Conflict of interest. None declared.

References

1. Prakash,N. and Wurst,W. (2004) Specification of midbrain territory. *Cell Tissue Res.*, 318, 5–14.
2. Wurst,W. and Bally-Cuif,L. (2001) Neural plate patterning: upstream and downstream of the isthmic organizer. *Nat. Rev. Neurosci.*, 2, 99–108.
3. Hoekstra,E.J., Mesman,S., de Munnik,W.A. *et al.* (2013) LMX1B Is Part of a transcriptional complex with PSPC1 and PSF. *PLoS One*, 8, e53122.
4. Crespo-Enriquez,I., Partanen,J., Martinez,S. *et al.* (2012) Fgf8-related secondary organizers exert different polarizing planar instructions along the mouse anterior neural tube. *PLoS One*, 7, e39977.

5. Sato,T. and Joyner,A.L. (2009) The duration of Fgf8 isthmic organizer expression is key to patterning different tectal-isthmo-cerebellum structures. *Development*, 136, 3617–3626.
6. Guo,C., Qiu,H.-Y., Huang,Y. *et al.* (2007) Lmx1b is essential for Fgf8 and Wnt1 expression in the isthmic organizer during tectum and cerebellum development in mice. *Development*, 134, 317–325.
7. Dworkin,S. and Jane,S.M. (2013) Novel mechanisms that pattern and shape the midbrain-hindbrain boundary. *Cell. Mol. Life Sci.*, 70, 3365–33674.
8. Merlini,L., Fluss,J., Dhouib,A. *et al.* (2013) Mid-hindbrain malformations due to drugs taken during pregnancy. *J. Child Neurol.*, 29, 538–544.
9. Barkovich,A.J., Millen,K.J. and Dobyns,W.B. (2009) A developmental and genetic classification for midbrain-hindbrain malformations. *Brain*, 132, 3199–3230.
10. Doherty,D., Millen,K.J. and Barkovich,A.J. (2013) Midbrain and hindbrain malformations: advances in clinical diagnosis, imaging, and genetics. *Lancet Neurol.*, 12, 381–393.
11. Pollock,J.D., Wu,D.-Y. and Satterlee,J.S. (2014) Molecular neuroanatomy: a generation of progress. *Trends Neurosci.*, 37, 106–123.
12. Finger,J.H., Smith,C.M., Hayamizu,T.F. *et al.* (2011) The mouse gene expression database (GXD): 2011 update. *Nucleic Acids Res.*, 39, D835–D841.
13. Lein,E.S., Hawrylycz,M.J., Ao,N. *et al.* (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445, 168–176.
14. Visel,A., Thaller,C. and Eichele,G. (2004) GenePaint.org: an atlas of gene expression patterns in the mouse embryo. *Nucleic Acids Res.*, 32, D552–D556.
15. Christiansen,J.H., Yang,Y., Venkataraman,S. *et al.* (2006) EMAGE: a spatial database of gene expression patterns during mouse embryo development. *Nucleic Acids Res.*, 34, D637–D641.
16. Siddiqui,A.S., Khattra,J., Delaney,A.D. *et al.* (2005) A mouse atlas of gene expression: Large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. *Proc. Natl Acad. Sci. UUSA*, 102, 18485–18490.
17. Aswani,A., Keranen,S., Brown,J. *et al.* (2010) Nonparametric identification of regulatory interactions from spatial and temporal gene expression data. *BMC Bioinformatics*, 11, 413.
18. von Dassow,G., Meir,E., Munro,E.M. *et al.* (2000) The segment polarity network is a robust developmental module. *Nature*, 406, 188–192.
19. Ma,W., Lai,L., Ouyang,Q. *et al.* (2006) Robustness and modular design of the *Drosophila* segment polarity network. *Mol. Syst. Biol.*, 2, 70.
20. Turning,A.M. (1952) The chemical basis of morphogenesis. *Philos. Trans. R. Soc. Lond. B*, 237, 37–72.
21. Holloway,D.T., Kon,M. and DeLisi,C. (2005) Integrating genomic data to predict transcription factor binding. *Genome Inform.*, 16, 83–94.
22. Nykter,M., Lähdesmäki,H., Rust,A. *et al.* (2009) A data integration framework for prediction of transcription factor targets. *Ann. N. Y. Acad. Sci.*, 1158, 205–214.
23. Franceschini,A., Szklarczyk,D., Frankild,S. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, 41, D808–D815.
24. Kerrien,S., Aranda,B., Breuza,L. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, 40, D841–D846.
25. Chatr-aryamontri,A., Breitkreutz,B.-J., Heinicke,S. *et al.* (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, 41, D816–D823.
26. Wittmann,D.M., Blöchl,F., Trümbach,D. *et al.* (2009) Spatial analysis of expression patterns predicts genetic interactions at the mid-hindbrain boundary. *PLoS Comput. Biol.*, 5, e1000569.
27. Hock,S., Ng,Y.-K., Hasenauer,J. *et al.* (2013) Sharpening of expression domains induced by transcription and microRNA regulation within a spatio-temporal model of mid-hindbrain boundary formation. *BMC Syst. Biol.*, 7, 48.
28. Puelles,L., Harrison,M., Paxinos,G. *et al.* (2013) A developmental ontology for the mammalian brain based on the prosomeric model. *Trends Neurosci.*, 36, 570–578.
29. Vieira,C., Pombero,A., García-Lopez,R. *et al.* (2010) Molecular mechanisms controlling brain development: an overview of neuroepithelial secondary organizers. *Int. J. Dev. Biol.*, 54, 7–20.
30. His,W. (1892) Zur allgemeinen morphologie des gehirns. *Arch. Anat. EntwGesch.*, 1892, 346–383.
31. Jessell,T.M. (2000) Neuronal specification in the spinal cord: inductive signals and transcriptional codes. *Nat. Rev. Genet.*, 1, 20–29.
32. Caspary,T. and Anderson,K.V. (2003) Patterning cell types in the dorsal spinal cord: what the mouse mutants say. *Nat. Rev. Neurosci.*, 4, 289–297.
33. Bard,J.B.L., Kaufman,M.H., Dubreuil,C. *et al.* (1998) An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. *Mech. Dev.*, 74, 111–120.
34. Hayamizu,T., Wicks,M., Davidson,D. *et al.* (2013) EMAP/EMAPA ontology of mouse developmental anatomy: 2013 update. *J. Biomed. Semantics*, 4, 15.
35. Prakash,N. and Wurst,W. (2006) Genetic networks controlling the development of midbrain dopaminergic neurons. *J. Physiol.*, 575, 403–410.
36. Clevers,H. and Nusse,R. (2012) Wnt/*β*-catenin signaling and disease. *Cell*, 149, 1192–1205.
37. van Amerongen,R. and Nusse,R. (2009) Towards an integrated view of Wnt signaling in development. *Development*, 136, 3205-3214.
38. Prakash,N. and Wurst,W. (2007) A Wnt signal regulates stem cell fate and differentiation *in vivo*. *Neurodegener. Dis.*, 4, 333–338.
39. Chung,S., Leung,A., Han,B.-S. *et al.* (2009) Wnt1-lmx1a forms a novel autoregulatory loop and controls midbrain dopaminergic differentiation synergistically with the SHH-FoxA2 pathway. *Cell Stem Cell*, 5, 646–658.
40. Omodei,D., Acampora,D., Mancuso,P. *et al.* (2008) Anterior-posterior graded response to Otx2 controls proliferation and differentiation of dopaminergic progenitors in the ventral mesencephalon. *Development*, 135, 3459–3470.
41. McMahon,A.P., Joyner,A.L., Bradley,A. *et al.* (1992) The midbrain-hindbrain phenotype of Wnt-1−Wnt-1− mice results from

stepwise deletion of engrailed-expressing cells by 9.5 days postcoitum. *Cell*, 69, 581–595.

42. Danielian,P.S. and McMahon,A.P. (1996) Engrailed-1 as a target of the Wnt-1 signalling pathway in vertebrate midbrain development. *Nature*, 383, 332–334.
43. McGrew,L.L., Takemaru,K.-I., Bates,R. *et al.* (1999) Direct regulation of the Xenopus engrailed-2 promoter by the Wnt signaling pathway, and a molecular screen for Wnt-responsive genes, confirm a role for Wnt signaling during neural patterning in Xenopus. *Mech. Dev.*, 87, 21–32.
44. Shtutman,M., Zhurinsky,J., Simcha,I. *et al.* (1999) The cyclin D1 gene is a target of the *β*-catenin/LEF-1 pathway. *Proc. Natl Acad. Sci. USA*, 96, 5522–5527.
45. Tetsu,O. and McCormick,F. (1999) [beta]-Catenin regulates expression of cyclin D1 in colon carcinoma cells. *Nature*, 398, 422–426.
46. Chamorro,M.N., Schwartz,D.R., Vonica,A. *et al.* (2005) FGF-20 and DKK1 are transcriptional targets of [beta]-catenin and FGF-20 is implicated in cancer and development. *EMBO J.*, 24, 73–84.
47. Fujimura,N., Vacik,T., Machon,O. *et al.* (2007) Wnt-mediated down-regulation of Sp1 target genes by a transcriptional repressor Sp5. *J. Biol. Chem.*, 282, 1225–1237.
48. Ellisor,D., Rieser,C., Voelcker,B. *et al.* (2012) Genetic dissection of midbrain dopamine neuron development *in vivo*. *Dev. Biol.*, 372, 249–262.
49. Li,B., Kuriyama,S., Moreno,M. *et al.* (2009) The posteriorizing gene Gbx2 is a direct target of Wnt signalling and the earliest factor in neural crest induction. *Development*, 136, 3267–3278.
50. Joksimovic,M., Yun,B.A., Kittappa,R. *et al.* (2009) Wnt antagonism of Shh facilitates midbrain floor plate neurogenesis. *Nat. Neurosci.*, 12, 125–131.
51. Lee,H.-Y., Kléber,M., Hari,L. *et al.* (2004) Instructive role of Wnt/ß-catenin in sensory fate specification in neural crest stem cells. *Science*, 303, 1020–1023.
52. Braun,M.M., Etheridge,A., Bernard,A. *et al.* (2003) Wnt signaling is required at distinct stages of development for the induction of the posterior forebrain. *Development*, 130, 5579–5587.
53. Lee,S.M., Danielian,P.S., Fritzsch,B. *et al.* (1997) Evidence that FGF8 signalling from the midbrain-hindbrain junction regulates growth and polarity in the developing midbrain. *Development*, 124, 959–969.
54. Yang,J., Brown,A., Ellisor,D. *et al.* (2013) Dynamic temporal requirement of Wnt1 in midbrain dopamine neuron development. *Development*, 140, 1342–1352.
55. Hovanes,K., Li,T.W.H., Munguia,J.E. *et al.* (2001) [beta]-catenin-sensitive isoforms of lymphoid enhancer factor-1 are selectively expressed in colon cancer. *Nat. Genet.*, 28, 53–57.
56. Filali,M., Cheng,N., Abbott,D. *et al.* (2002) Wnt-3A/*β*-catenin signaling induces transcription from the LEF-1 promoter. *J. Biol. Chem.*, 277, 33398–33410.
57. Lin,L., Cui,L., Zhou,W. *et al.* (2007) *β*-Catenin directly regulates Islet1 expression in cardiovascular progenitors and is required for multiple aspects of cardiogenesis. *Proc. Natl Acad. Sci. USA*, 104, 9313–9318.
58. Mastik,G.S., Fan,C.-M., Tessier-Lavigne,M. *et al.* (1996) Early deletion of neuromeres in Wnt-l-/- mutant mice: evaluation by morphological and molecular markers. *J. Comp. Neurol.*, 374, 246–258.
59. Porter,J.D. and Baker,R.S. (1997) Absence of oculomotor and trochlear motoneurons leads to altered extraocular muscle development in the Wnt-1 null mutant mouse. *Dev. Brain Res.*, 100, 121–126.
60. Cartharius,K., Frech,K., Grote,K. *et al.* (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, 21, 2933–2942.
61. Quandt,K., Frech,K., Karas,H. *et al.* (1995) Matlnd and Matlnspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, 23, 4878–4884.
62. Cohen,C.D., Klingenhoff,A., Boucherot,A. *et al.* (2006) Comparative promoter analysis allows de novo identification of specialized cell junction-associated proteins. *Proc. Natl Acad. Sci. USA*, 103, 5682–5687.
63. Zhou,X. and Tuck,D.P. (2007) MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. *Bioinformatics*, 23, 1106–1114.
64. Augustin,R., Lichtenthaler,S.F., Greeff,M. *et al.* (2011) Bioinformatics identification of modules of transcription factor binding sites in Alzheimer's disease-related genes by in silico promoter analysis and microarrays. *Int. J. Alzheimers Dis.*, 2011, 154325.
65. Trümbach,D., Graf,C., Pütz,B. *et al.* (2010) Deducing corticotropin-releasing hormone receptor type 1 signaling networks from gene expression data by usage of genetic algorithms and graphical Gaussian models. *BMC Syst. Biol.*, 4, 159.
66. Hajihosseini,M.K. and Heath,J.K. (2002) Expression patterns of fibroblast growth factors-18 and -20 in mouse embryos is suggestive of novel roles in calvarial and limb development. *Mech. Dev.*, 113, 79–83.
67. Kobielak,K., Kobielak,A. and Trzeciak,W.H. (1999) Cloning of the lymphoid enhancer binding factor-1 (Lef-1) cDNA from rat kidney: homology to the mouse sequence. *Acta. Biochim. Pol.*, 46, 885–888.
68. Chenn,A. and Walsh,C.A. (2002) Regulation of cerebral cortical size by control of cell cycle exit in neural precursors. *Science*, 297, 365–369.
69. Nakamura,R., Hunter,D., Yi,H. *et al.* (2007) Identification of two novel activities of the Wnt signaling regulator Dickkopf 3 and characterization of its expression in the mouse retina. *BMC Cell Biol.*, 8, 52.
70. Mathelier,A., Zhao,X., Zhang,A.W. *et al.* (2013) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 42, D142–D147.