Activity of microRNAs and transcription factors in Gene Regulatory Networks

Haroon Naeem

München 2011

Activity of microRNAs and transcription factors in Gene regulatory Networks

Haroon Naeem

Dissertation
an der Fakultät fur Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Haroon Naeem
aus Lahore

München, den 15.11.2011

Erstgutachter: Prof. Dr. Ralf Zimmer
Zweitgutachter: Prof. Dr. Hans-Werner Mewes
Tag der mündlichen Prüfung: 10.2.2012

Contents

List of Figures

List of Tables

Summary

In biological research, diverse high-throughput techniques enable the investigation of whole systems at the molecular level. The development of new methods and algorithms is necessary to analyze and interpret measurements of gene and protein expression and of interactions between genes and proteins. One of the challenges is the integrated analysis of gene expression and the associated regulation mechanisms.

The two most important types of regulators, transcription factors (TFs) and microRNAs (miRNAs), often cooperate in complex networks at the transcriptional and post-transcriptional level and, thus, enable a combinatorial and highly complex regulation of cellular processes. For instance, TFs activate and inhibit the expression of other genes including other TFs whereas miRNAs can post-transcriptionally induce the degradation of transcribed RNA and impair the translation of mRNA into proteins.

The identification of gene regulatory networks (GRNs) is mandatory in order to understand the underlying control mechanisms. The expression of regulators is itself regulated, i.e. activating or inhibiting regulators in varying conditions and perturbations. Thus, measurements of gene expression following targeted perturbations (knockouts or overexpressions) of these regulators are of particular importance. The prediction of the activity states of the regulators and the prediction of the target genes are first important steps towards the construction of GRNs.

This thesis deals with these first bioinformatics steps to construct GRNs. Targets of TFs and miRNAs are determined as comprehensively and accurately as possible. The activity state of regulators is predicted for specific high-throughput data and specific contexts using appropriate statistical approaches. Moreover, (parts of) GRNs are inferred, which lead to explanations of given measurements. The thesis describes new approaches for these tasks together with accompanying evaluations and validations. This immediately defines the three main goals of the current thesis:

1. The development of a comprehensive database of regulator-target relation.

   Regulators and targets are retrieved from public repositories, extracted from the literature via text mining and collected into the miRSel database. In addition, relations can be predicted using various published methods. In order to determine the activity states of regulators (see 2.) and to infer GRNs (3.) comprehensive and accurate regulator-target relations are required.
   It could be shown that text mining enables the reliable extraction of miRNA, gene, and protein names as well as their relations from scientific free texts. Overall, the miRSel contains about three times more relations for the model organisms human, mouse, and rat as compared to state-of-the-art databases (e.g. TarBase, one of the currently most used resources for miRNA-target relations).

2. The prediction of activity states of regulators based on improved target sets.

In order to investigate mechanisms of gene regulation, the experimental contexts have to be determined in which the respective regulators become active. A regulator is predicted as active based on appropriate statistical tests applied to the expression values of its set of target genes. For this task various gene set enrichment (GSE) methods have been proposed. Unfortunately, before an actual experiment it is unknown which genes are affected. The missing standard-of-truth so far has prevented the systematic assessment and evaluation of GSE tests. In contrast, the trigger of gene expression changes is of course known for experiments where a particular regulator has been directly perturbed (i.e. by knockout, transfection, or overexpression). Based on such datasets, we have systematically evaluated 12 current GSE tests. In our analysis ANOVA and the Wilcoxon test performed best.

3. The prediction of regulation cascades.

   Using gene expression measurements and given regulator-target relations (e.g. from the miRSel database) GRNs are derived. GSE tests are applied to determine TFs and miRNAs that change their activity as cellular response to an overexpressed miRNA. Gene regulatory networks can constructed iteratively. Our models show how miRNAs trigger gene expression changes: either directly or indirectly via cascades of miRNA-TF, miRNA-kinase-TF as well as TF-TF relations.

   In this thesis we focus on measurements which have been obtained after overexpression of miRNAs. Surprisingly, a number of cancer relevant miRNAs influence a common core of TFs which are involved in processes such as proliferation and apoptosis.

Zusammenfassung

In der biologischen Forschung machen diverse Hochdurchsatztechniken die Untersuchung ganzer Systeme auf molekularer Ebene möglich. Um Messungen von Gen- und Proteinexpression sowie den Interaktionen zwischen Genen und Proteinen analysieren und interpretieren zu können, ist die Entwicklung neuer Rechenmethoden und Algorithmen erforderlich. Eine der Herausforderungen ist die integrierte Analyze der Genexpression und den zugehörigen Regulationsmechanismen.

Die beiden wichtigsten bisher bekannten Typen von Regulatoren, Transkriptionsfaktoren (TFs) und microRNAs (miRNAs), wirken auf transkriptioneller und post-transkriptioneller Ebene häufig in komplexen Netzwerken zusammen zur kombinatorischen und hochkomplexen Steuerung der zellulären Prozesse. Z.B. können TFs die Expression anderer Gene und auch anderer TF aktivieren und inhibieren, miRNAs können post-transkriptionell den Abbau schon transkribierter mRNA fördern bzw. die Translation der mRNA in das zugehörige Protein behindern.

Deshalb ist die Aufklärung der genregulatorischen Netzwerke (GRNs) unumgänglich, um die unterliegenden Steuerungsprozesse zu verstehen. Unter variierenden Bedingungen und Perturbationen kann die Expression der Regulatoren selbst reguliert sein, zudem können Regulatoren aber auch aktiviert und inaktiviert werden. Deshalb sind Messungen der Genexpression nach gezielten Störungen (Knockout oder Überexpression) dieser Regulatoren besonders interessant. Die Aufklärung der Aktivität der Regulatoren und die Aufklärung der regulierten Gene, der sogenannten Zielgene (targets) sind erste wichtige Schritte zur Konstruktion von Genregulationsnetzwerken (GRNs).

Die vorliegende Arbeit befasst sich mit diesen ersten Schritten zur bioinformatischen Aufklärung von GRNs. Dazu werden Zielgene von TFs und miRNAs mit verschiedenen Methoden möglichst genau und umfassend bestimmt und daraus mittels geeigneter algorithmischer und statistischer Verfahren die Aktivität der Regulatoren in bestimmten Kontexten und für bestimmte Hochdurchsatzdaten vorhergesagt. Weitergehend wird dann versucht, (Teile der) GRNs zu inferieren, die die Messdaten von bestimmten Hochdurchsatzexperimenten erklären können. Für diese Aufgaben werden neue algorithmische Verfahren entwickelt und eingehend evaluiert. Daraus ergeben sich die folgenden drei Hauptziele der Arbeit:

1. Der Aufbau einer umfassenden Datenbank von Regulator-Zielgen Beziehungen.

   Regulatoren und Zielgene (targets) werden aus öffentlich verfügbaren Repositories entnommen, durch Textmining aus der Literatur extrahiert und in der Datenbank miRSel zusammengeführt. Zusätzlich können solche Beziehungen auch durch diverse veröffentlichte Programme vorhergesagt werden. Zur Analyze der Aktivität der Regulatoren (siehe 2.) und der genregulatorischen Netzwerke (3.)

werden möglichst vollständige und korrekte Regulator-Target Beziehungen benötigt.
Es konnte gezeigt werden, dass Textmining die zuverlässige Extraktion von miRNA-, Gen-und Proteinnamen sowie deren Beziehungen aus Texten ermöglicht. Insgesamt konnte mit miRSel die Anzahl der miRNA-Gen Assoziationen in Mensch, Maus und Ratte um mindestens das dreifache im Vergleich zu state-of-the-art Datenbanken (z.B. TarBase, eine der aktuell meistverwendeten Ressourcen für miRNA-Gen Beziehungen) gesteigert werden.

2. Die Vorhersage der Aktivität von Regulatoren auf Basis der verbesserten Target Mengen.

   Um die Mechanismen der Genregulation zu untersuchen, müssen die experimentellen Kontexte bestimmt werden, unter denen die jeweiligen Regulatoren aktiv werden. Ein Regulator wird als aktiv vorhergesagt, indem mittels statistischer Tests die Expressionswerte seiner Zielgene analysiert werden. Dafür wurden sogenannte "gene set enrichment" (GSE) Tests vorgeschlagen. Allerdings weiß man vor einem Experiment nicht, welche Gene betroffen sein werden. Das Fehlen eines zuverlässigen standard-of-truth hat bisher die systematische Auswahl und Bewertung der GSE Tests verhindert. Im Gegensatz dazu ist der betroffene Regulator natürlich bekannt, wenn er direkt (z.B. durch Knockout, Transfektion oder Überexpression) experimentell beeinflusst wurde. Für solche Datensätze wurde hier eine systematische vergleichende Bewertung von insgesamt 12 aktuellen GSE-Tests durchgeführt. In unserer Analyze zeigen ANOVA und der Wilcoxon-Test die besten Ergebnisse.

3. Die Vorhersage von Regulationskaskaden.

   Aus Genexpressionsmessungen und gegebenen Regulator-Target Beziehungen (wie z.B. aus der miRSel Datenbank) sollen regulatorische Netzwerke abgeleitet werden. Dazu werden GSE Tests angewendet, um TFs und miRNAs zu bestimmen, die ihre Aktivität als zelluläre Antwort auf die überexprimierten miRNAs ändern. Iterativ können so Regulationsnetzwerke rekonstruiert werden. Diese Modelle zeigen, wie miRNAs die Genexpression beeinflussen können: entweder direkt oder indirekt über Kaskaden von miRNA-TF, miRNA-Kinase-TF sowie TF-TF Beziehungen.

   In dieser Arbeit konzentrieren wir uns auf Messungen, die nach Überexpression von Krebs-relevanten miRNAs durchgeführt wurden. Überraschenderweise stellt sich heraus, dass eine Reihe von verschiedenen Krebs-relevanten miRNAs einen gemeinsamen Kern von TFs beeinflussen, die an Prozessen wie Proliferation oder Apoptose beteiligt sind.

1. Introduction

1. Introduction

Recent technological advances that led to 'omics' revolution have enabled large-scale data generation in different areas of biology. Thousands of high-throughput datasets are available that contain the expression levels of all genes of an organism under various experimental conditions. Expression of each gene is a complex process that requires coordination of many factors for maintaining the basic mechanisms of development and controlled by two important classes of regulators: microRNAs (miRNAs) and transcription factors (TFs) (*Nestler et al.,* 2004; *Hobert*, 2008). TFs are regulatory proteins that bind to promoter regions of target genes to regulate their levels of expression (*Chen et al.,* 2007). miRNAs are small (~22-nucleotide) non-coding RNAs (ncRNAs) that post-transcriptionally regulate the levels of a potentially large number of proteins by base-pairing to messenger RNAs (mRNAs) (*Brodersen et al.,* 2009). Perfect or near-perfect complementarity to the target RNA promotes cleavage and degradation of the RNA, while imperfect base-pairing impairs translation of the target mRNA (*Orom et al.,* 2009). Functional studies implicate effects of miRNAs on a wide range of cellular and developmental processes such as cell cycle control, cell growth, apoptosis, embryo development, stress response, metabolism or morphogenesis and in various diseases including cancer (*Li et al.,* 2010). Several miRNAs were found differentially expressed during brain development, neuronal differentiation including neurological syndromes such as Schizophrenia, Huntington and Parkinson disease (*Kuss* and *Chen,* 2008).

Since the discovery of the first miRNA, lin-4 in *Caenorhabditis elegans* (*C. elegans*) (*Lee et al.,* 1993), thousands of miRNAs have been identified in vertebrates, flies, worms and plants and even in viruses (*Griffiths-Jones et al.,* 2008). Tens of thousands of gene targets have been predicted mostly by the use of automatic prediction algorithms (*Mazière* and *Enright*, 2007; *Kertesz et al.,* 2007; *Wang et al.,* 2008). So far, the targets of only a handful of these miRNAs have been experimentally validated (*Ritchie et al.,* 2009; *Papadopoulos et al.,* 2009). Recently developed databases provide resources of miRNA nomenclature, sequence data, genomic localization and functional annotation in human, mouse, rat and other organisms (*Griffiths-Jones et al.,* 2008) similar to the established gene-specific databases (*Maglott et al.,* 2011; *Bruford et al.,* 2008; *Bult et al.,* 2008). Several web-based tools integrate predicted miRNA targets e.g. miRNAMap 2.0 (*Hsu et al.,* 2008), miRGator (*Nam et al.,* 2008), miRGen (*Alexiou et al.,* 2009). The databases such as miR2Disease (*Jiang et al.,* 2009), miRecords (*Xiao et al.,* 2009) and TarBase (*Papadopoulos et al.,* 2009) collect target genes of the miRNAs in different organisms.

Several databases of predicted and/or experimentally validated TF regulatory interactions have been developed such as TRANSFAC (*Wingender et al., 2000*; *Matys et al.,* 2006). RegulonDB (*Gama-Castro et al.,* 2011) contains experimentally validated and manually curated TF-gene regulatory relationships in Escherichia coli (*E. coli*). *MacIsaac et al.* (2006) examine ChIP-chip data to determine the TF binding sites (TFBS) in Saccharomyces cerevisiae (*S. cerevisiae*). *Liu et al.* (2008) implement a phylogenetic footprinting approach to identify the TFBS in mammalian gene promoters. Several transcription profiling studies of miRNA and TF overexpression or deletion experiments have been performed to investigate the influence of regulators on transcript levels (*Hu et al.,* 2007; *Selbach et al.,* 2008; *Baek et al.,* 2008). Some of

which have been incorporated into Many Microbe Microarrays Database (M3D) developed by *Faith et al.* (2008). The M3D database collects Affymetrix microarrays for *E. coli* and *S. cerevisiae*.

The expression profiling studies show that regulators such as miRNAs exert a widespread impact on the regulation of their target genes and (potentially mediated via TFs) on non-target genes (*Tu et al.,* 2009). TFs have been found enriched among miRNA targets in plants (*Rhoades et al.,* 2002) and insects (*Enright et al.,* 2003), suggesting that these two classes of regulators could be linked in gene regulatory networks (*Skipper,* 2008). However, the determination of the conditions where given regulators become active is difficult as regulators themselves are frequently regulated on the protein level (e.g. by phosphorylation) that is not immediately detectable by transcriptional profiling. On the other hand, transcriptional effects of regulators are in general expected to be small and could easily be obscured by noise in the measurements. The detection of active regulators thus requires very sensitive approaches that rely on indirect evidence rather than the expression of the regulators themselves.

The identification of miRNAs and TFs activity changes is important to understand the regulation of gene expression and dynamic cellular mechanisms (*Tu et al.,* 2009; *Hu,* 2010). It is challenging to measure the regulators activity (*Boorsma et al.,* 2008). It may not be directly observed but can be determined by analyzing the activities of genes they regulate (*Farh et al.,* 2005; *Hu et al.,* 2007). Hence, the target genes of a given miRNA or TF are considered a gene set. If more genes than expected by chance for such a set exhibit significant fold changes, the miRNA/TF is assumed active. For the analysis of such gene set, several enrichment tests have been proposed. *Sohler et al.* (2005), *Essaghir et al.* (2010) and *Liu et al.* (2010) independently proposed the hypergeometric test to detect the active TFs. Analogously, statistical tests such as the hypergeometric test were applied to detect expression changes of miRNAs based on the expression of their target gene set (*Sood et al.,* 2006; *Arora et al.,* 2008; *Cheng et al.,* 2009; *Volinia et al.,* 2010; *Ott et al.,* 2011).

1.1 Problem identification

This thesis aims to improve the understanding of the gene expression regulation controlled by two important regulators: miRNAs and TFs. To analyze the regulatory mechanisms of gene expression, we need to determine the experimental conditions where these regulators become active. For this purpose, we have to choose the appropriate methods for the detection of activity of regulators based on indirect evidence such as the target gene set. In turn, this requires a collection of regulator target gene sets. This section gives a brief introduction to three crucial, mutually dependent problems.

1.1.1 miRNA-target gene associations in databases

The comprehensive collection of miRNA-gene associations is important for the development of miRNA target prediction tools and the analysis of regulatory networks. Most miRNA-target associations contained in databases are derived from the large scale experiments where a detailed experimental validation of individual pairs has not been performed. For instance, TarBase and miRecords report 1031 and 776 miRNA-target pairs in human, respectively. Out of these, 769 and 447 have been

obtained from the supplementary material of just two publications (*Lim et al.,* 2005; *Selbach et al.,* 2008) and (*Lim et al.,* 2005; *Calin et al.,* 2008). miRecords additionally collects 158 rat miRNA-target gene pairs, 140 out of these pairs are extracted from a single publication (*Jeyaseelan et al.,* 2008). A more detailed analysis has been performed by *Ritchie et al.,* (2009). They found that only 48 miRNA-target pairs of miRecords are sufficiently validated by experiments and, as a consequence, they conclude that benchmarks for the evaluation of miRNA target prediction algorithms cannot be constructed from the available databases.

1.1.2 Assessment of gene set enrichment scores

The interpretation of gene expression studies reporting mRNA levels for a high number of genes or other expressed sequences is difficult. Instead of individual genes, it has been proposed to analyze gene sets corresponding to biological processes. The Gene Ontology (GO, *Harris et al.,* 2004) is an example source for biological process definitions and process associated gene sets. The analysis of expression data in the context of such gene sets can be performed by many different enrichment or over-representation tests (see section 2.2). These tests aim to detect gene sets exhibiting significant levels of differential expression. However, it is difficult to decide *a priori* which biological processes will be affected in a given gene expression experiment. This lack of a dependable standard of truth has prevented an objective selection and evaluation of enrichment tests on real data. None of the studies provide a comprehensive comparative analysis of the tests evaluated against real data.

1.1.3 Analysis of miRNA regulated TFs

Several expression datasets of miRNA overexpression experiments are available to analyze the regulatory mechanisms downstream of miRNA effects. The miRNA induced regulatory effects can be propagated via TFs. *Sohler et al.* (2005), *Essaghir et al.* (2010) and *Liu et al.* (2010) applied the hypergeometric test to determine the activity changes of TFs. According to our analysis, the hypergeometric test is not sensitive enough to pick up the small expression changes caused by miRNAs. To assemble regulatory cascades from experiments where cancer related miRNAs have been over-expressed, *Tu et al.* (2009) suggest linear models to detect miRNA regulated TFs. They extracted two layered networks where TFs mediate miRNA initiated regulatory effects. The time complexity of their approach substantially limited the set of detected TFs. On average, only two active TFs were identified per miRNA overexpression experiment. Thus to understand miRNA-mediated gene regulation, a more detailed gene regulatory network analysis is needed.

1.2 Outline of the thesis

In this thesis we focus on several problems to enhance our understanding of the regulation of gene expression as described above. We address these problems with new methods including new databases and appropriate evaluations.

In the **Background** chapter, we present a brief description and discussion of previously developed methods. We focus on various approaches for predicting or experimentally validating the target genes of miRNAs and their performance reviews in the literature (section 2.1). We introduce gene set enrichment (GSE) tests originally

proposed for analyzing gene sets associated with biological processes (section 2.2). Furthermore, we provide an overview of biomedical named entity recognition (NER) discussing miRNA, gene and protein naming convention, approaches to identify biomedical named entities in texts, Aho-Corasick string matching algorithm (*Aho* and *Corasick,* 1975) and quality assessment tests (section 2.3).

In the **Databases** chapter, we provide an introduction to state-of-the-art publically available databases related to the field. The databases include miRNAs (section 3.1), miRNA-target gene interactions (section 3.2), TF-target gene relations (section 3.3), gene/protein nomenclature (section 3.4), protein-protein interactions (section 3.5), taxonomy (section 3.6), predefined gene sets and pathways (section 3.7) including gene expression (section 3.8) and scientific literature repositories (section 3.9).

In the **Methods** chapter, we describe the new methods together with new databases to address the challenge of improving the understanding of the gene regulation (section 1.1). Consequently, the entire chapter is explained via three mutually dependent subsections.

First, we explain and discuss the implementation of a new miRNA-target gene association database, namely, miRSel that combines the text mining results with existing databases and computational predictions (*Naeem et al.,* 2010, section 4.1). We focus on a dictionary-based approach for biomedical NER with application to the detection of miRNA-gene associations including miRNA-target relations in texts.

Second, we discuss the first comprehensive comparison and rigorous assessment of 12 statistical enrichment tests for analyzing gene sets (*Naeem et al.,* 2011, section 4.2). We applied state-of-the-art statistical methods such as ANOVA, Wilcoxon, Kolmogorov-Smirnov as well as the hypergeometric test to decide whether or not to reject the null hypothesis, i.e. that expression changes in regulator target sets might be due to random fluctuations in the data. Before getting into the details of enrichment tests we will explain how the standard-of-truth is derived, and how the sign annotations are treated to assess the up and down regulation of gene sets and consistency of each test statistic.

Finally, we discuss the method MIRTFnet to determine the experimental conditions where certain regulators like miRNAs become active and how they regulate the transcriptome via cascades of other miRNAs, TFs or kinases (*Naeem et al.,* 2011, section 4.3).

In the **Results** chapter, we provide the results for each developed method/database accordingly.

Initially, we explain the results of our developed database miRSel. Comparison to previously developed state-of-the-art resources for miRNA-gene relationships e.g. TarBase, miRSel increased the number of miRNA-gene associations by at least three-fold (section 5.1).

Then, we discuss the results of 12 different tests to detect the activity changes of miRNAs and TFs based on our improved set of regulator target genes (section 5.2). We focused on comparing the performance and consistency of each statistical test across different species datasets including *E. coli*, *S. cerevisiae* and human. Subsequently, we discussed the ranking of enrichment tests based on our findings. Additionally, we elaborated that combining evaluated tests into a consensus is a particularly robust choice for the analysis of novel scenarios.

Finally, we describe the results of our developed method, MIRTFnet that explains the observed expression changes via models rooted at perturbed miRNAs. We discussed a range of different miRNAs that induce activity changes in a common core of TFs involved in cancer related processes (section 5.3).

In the **Conclusion and discussion** chapter, we discuss the results, draw some conclusions and hint to future direction of this work.

2. Background

2. Background

This section gives a brief introduction of the existing methods and techniques that are related to the field.

1) We describe a number of different approaches to predict and validate the miRNA-target gene interactions (section 2.1).
2) We discuss a range of enrichment methods originally proposed to determine the differential expression of gene sets representing biological processes (section 2.2).
3) We discuss the methods and approaches to named entity recognition (NER) with application to the detection of biomedical entities of interest (section 2.3).

2.1 Computational methods for prediction of miRNA-target gene relations

Since the detection of the first miRNA (*Lee et al.,* 1993), many more miRNAs have been identified in animals, plants and even in viruses (*Griffiths-Jones et al.,* 2008). These pre-miRNAs are known to be processed into mature miRNAs and integrated into the RNA-induced silencing complex (RISC) to regulate the expression of target messenger RNA (mRNA) genes (*Bartel*, 2009). Several studies hae been conducted to examine the miRNA targeting and biogenesis mechanisms (*Bartel,* 2009, Figure 1). The targeting mechanisms can be divided into several classes based on the level of complementarity to the target mRNA such as 5′-dominant canonical and 5′-dominant seed (2-8 nucleotides at the 5′ site of a miRNA) (*Mazière* and *Enright*, 2007).

Stark et al., (2003) developed the first computational miRNA target prediction program focused on Drosophila melanogaster. Since then, several target prediction programs have been proposed such as miRanda (*Enright et al.,* 2003), TargetScan (*Lewis et al.,* 2003), RNAhybrid (*Rehmsmeier et al.,* 2004), TargetScanS (*Lewis et al.,* 2005), PicTar (*krek et al.,* 2005), TargetBoost (*Saetrom et al.* 2005), MicroTar (*Thadani* and *Tammi*, 2006), RNA22 (*Miranda et al.,* 2006), NBmiRTar (*Yousef et al.,* 2007) and PITA (*Kertesz et al.,* 2007) and DIANA-microT (*Maragkakis et al.,* 2009). These methods predict the target genes of a miRNA based on important features of miRNA-mRNA duplexes such as seed region complementarity and conservation, and thermodynamic stability (binding free energy) using machine-learning based approaches (*Mazière* and *Enright*, 2007; *Bartel,* 2009; *Dai* and *Zhou,* 2010) (Table 1).

2.1.1 Seed-complementarity-based methods

The miRanda algorithm proposed by *Enright et al.* (2003) and improved by *John et al.* (2004), implements a dynamic programming method to identify the miRNA binding sites in the complementarity regions (3′ UTRs) of the target genes. TargetScan implements a strict criteria that requires a perfect base pairing between the seed region of a miRNA and the 3′-UTR of its target mRNA (*Lewis et al.,* 2003). TargetScanS is a customized version of the TargetScan and reduces the false positive predictions by restricting the miRNA-mRNA pairing to a 6-nucleotide seed region plus an additional adenosine anchor match (*Lewis et al.,* 2005; *Mazière* and *Enright*, 2007). PicTar combines the features such as seed match, RNA duplex free energy and evolutionary conservation together with maximum likelihood fit approach to score each miRNA-target gene interaction (*krek et al.,* 2005). Using PicTar, *Grün et al.*

(2005) and *Lall et al.* (2006) score and rank miRNA targets in several *Drosophila* and nematode species. *Robins et al.* (2005) and *Miranda et al.* (2006) integrate mRNA secondary structure features and pattern based approach to predict the target genes of a specific miRNA. HuMiTar combines the seed and outside-seed miRNA-mRNA pairing scores to predict the common targets of miRNAs (*Ruan et al.,* 2008).

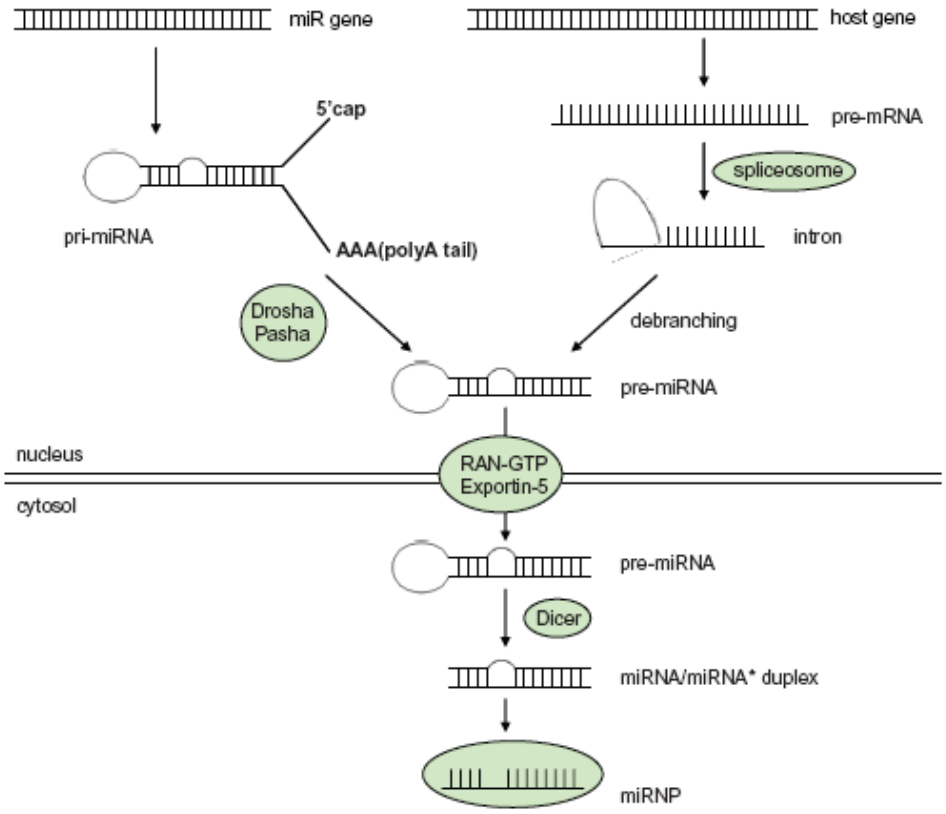

Figure 1: miRNA biogenesis in animals. In the cell nucleus, miRNAs are usually transcribed by the RNA polymerase II (Pol II) as primary transcripts (called as pri-miRNAs) with a cap and polyadenylated with multiple adenosines (a poly (A) tail). These pri-miRNAs are then processed into 70-nucleotide stem-loop structure (called as pre-miRNAs) via protein complex including nuclease Drosha and RNA binding protein Pasha. These pre-miRNA hairpins are then processed into mature miRNAs in the cytoplasm by interaction with the RNase III enzyme Dicer, which also commences the formation of RNA-induced silencing complex (RISC, also known as a miRNA ribonucleoprotein (miRNP) complex). These mature miRNAs together with miRNP complex responsible for the gene silencing activities. Figure taken from *Dai* and *Zhou* (2010).

2.1.2 Thermodynamic-based methods

Kiriakidou et al. (2004) developed Diana-microT that uses computational and experimental approaches combine with dynamic programming to calculate the miRNA-mRNA binding free energy. *Rehmsmeier et al.* (2004) proposed RNAhybrid that extends the RNA secondary structure prediction algorithm (*Zuker* and *Stiegler,* 1981) to two sequences. RNAhybrid can identify binding sites in the 3′-UTR of the target mRNA that can form thermodynamically stable duplexes with a miRNA (*Mazière* and *Enright,* 2007). A number of experimental studies suggest that target site accessibility is an important feature for repression (*Robins et al.,* 2005; *Zhao et al.,* 2005). *Kertesz et al.* (2007) developed PITA that uses the accessibility of the target sites.

2.1.3 Machine learning-based methods

Yousef et al. (2007) proposed NBmiRTar that uses a machine learning-based approach (naïve Bayes classifier) for predicting the miRNA targets. They combine multiple features extracted from the validated target gene sequences and miRNA-mRNA duplexes (*Yousef et al.,* 2007). *Saetrom et al.* (2005) developed a machine-learning-based algorithm TargetBoost that uses sequence information and binding site characteristics for the prediction of miRNA-target genes.

2.1.4 Motif-mining and gene expression-based methods

Several motif mining approaches have been proposed by searching the overrepresented mRNA sequence motifs in miRNA seed regions to predict the target genes (*Mazière* and *Enright*, 2007; *Xie et al.,* 2005). *Giraldez et al.* (2006) analyze the motifs in the 3′-UTR of the differentially expressed genes and conclude that motifs of 6-nucleotide bases could explain the majority of the validated targets.

Chi et al. (2009) use Argonaute (Ago) HITS-CLIP (high-throughput sequencing of RNAs isolated by cross linking immunoprecipitation) to determine the Ago-miRNA-mRNA interactions in the mouse. *Zhang et al.* (2007) identified over 3,000 mRNA genes by immunoprecipitation (IP) of the RISC components including AIN-1 and AIN-2 proteins in *C. elegans*. *Hammell et al.* (2008) developed miRWIP, a miRNA-mRNA target prediction tool based on these AIN-IP dataset features in *C. elegans*.

Huang et al. (2007) developed Bayesian learning analysis algorithm, GenMiR++ that uses the paired miRNA and mRNA expression profiles including sequence complementarity features to identify functional miRNA-mRNA pairs.

Method	Type of method or criteria	Web site	References
Stark *et al.*	Complementarity	http://www.russell.embl.de/miRNAs	*Stark et al., 2003*
miRanda	Complementarity	http://www.microrna.org	*Enright et al., 2003, John et al., 2004*
miRanda-miRBase	Complementarity	http://microrna.sanger.ac.uk/	*Griffiths-Jones et al., 2006*
TargetScan	Complementarity	http://www.targetscan.org/	*Lewis et al., 2003*
TargetScanS	Complementarity	http://www.targetscan.org/	*Lewis et al., 2005*
DIANA microT	Thermodynamics	http://diana.pcbi.upenn.edu/	*Kiriakidou et al., 2004*
PicTar	HMM	http://pictar.bio.nyu.edu	*Grün et al., 2005*
RNAHybrid	Thermodynamics	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/	*Rehmsmeier et al., 2004*
RNA22	Pattern discovery	http://cbcsrv.watson.ibm.com/rna22.html	
Micro Inspector		http://mirna.imbb.forth.gr/microinspector/	*Rusinov et al., 2005*
Ref. 27		http://tavazoielab.princeton.edu/mirnas	*Chan et al., 2005*
HuMiTar	Complementarity		*Ruan et al., 2008*
MicroTar	Complementarity	http://tiger.dbs.nus.edu.sg/microtar/	*Thadani and Tammi, 2006*
Diana-microT	Thermodynamics	http://diana.cslab.ece.ntua.gr/microT/	*Kiriakidou et al., 2004*
RNAhybrid	Thermodynamics	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/	*Rehmsmeier et al., 2004*
PITA	Target-site accessibility	http://genie.weizmann.ac.il/pubs/mir07/index.html	*Kertesz et al., 2007*
NBmiRTar	Machine learning	http://wotan.wistar.upenn.edu/NBmiRTar/	*Yousef et al., 2007*
TargetBoost	Machine learning	https://demo1.interagon.com/demo	*Saetrom et al., 2005*

Table 1: List of miRNA target prediction programs. *Stark et al.* (2003) developed the first miRNA target prediction program. Since then, several computational methods for the miRNA-target prediction have been developed. These methods predict the target genes of a miRNA based on important features of miRNA-mRNA duplexes such as seed region complementarity and conservation, target-site accessibility, and thermodynamic stability using machine-learning based approaches.

2.1.5 Experimental validation of miRNA-target gene interactions

The experimental methods such as luciferase reporter assays, qRT-PCR and western blots are used to confirm a miRNA-mRNA target interaction (*Hsu et al.,* 2011). The Western blots and qRT-PCR methods can detect the miRNA downstream effect at the protein level and the mRNA level (*Hsu et al.,* 2011; *Kuhn et al.,* 2008). In contrast, large-scale microarray experiments including stable isotope labelling with amino acids in culture (SILAC) or pulsed SILAC (pSILAC) have been performed to study the genome-wide changes in the transcriptome or proteome given the perturbation (e.g., overexpression) of a miRNA (*Lim et al.,* 2005; *Baek et al.,* 2008; *Selbach et al.,* 2008; *Hsu et al.,* 2011).

Other high-throughput methods like degradome sequencing is also useful to examine the miRNA cleavage target sites (*Addo-Quaye et al.,* 2008; *German et al.,*

2008). Recent progress in next generation RNA-sequencing together with gene expression studies can be used to elucidate miRNA-target associations. For a detailed review see *Thomson et al.* (2011).

2.1.6 Performance of prediction methods

Over the past few years several computational programs have been developed for the prediction of miRNA-target genes. *Rajewsky et al.* (2006) and *Ruan et al.* (2008) investigate the performance of several prediction programs and conclude that methods such as PicTar and TargetScan show high accuracy and sensitivity in comparison to miRanda and RNAhybrid. *Huang et al.* (2007) rank the miRNA target prediction programs based on their performance to predict putative target genes. *Selbach et al.* (2008) show that methods based on evolutionary conservation of seed region achieves high precision levels. *Alexiou et al.* (2009) show that the prediction programs such as TargetScan and PicTar achieve precision of ~50% and sensitivities from 6 to 12% (*Dai* and *Zhous,* 2010; *Min* and *Yoon,* 2010). For a detailed review see *Sethupathy et al.* (2006), *Mazière* and *Enright* (2007), *Ritchie et al.* (2009) and *Min* and *Yoon,* (2010).

2.2 Gene set enrichment analysis

A microarray experiment typically results in a long list of differentially expressed genes (DEGs) that is the starting point to gain insights into biological mechanisms (*Gatti et al.,* 2010). Several statistical methods for the analysis of sets of DEGs have been proposed (reviewed by *Goeman et al.,* 2007, *Rivals et al.,* 2007, *Nam* and *kim,* 2008, *Ackermann* and *Strimmer,* 2009). Most methods test for the over-representation of predefined sets of genes (e.g., Gene ontology (GO), KEGG pathways) in the list of DEGs (*Hosack et al.,* 2003; *Zeeberg et al.,* 2003; *Zhang et al.,* 2004; *Martin et al.,* 2004; *Al-Shahrour et al.,* 2004, *Beissbarth et al.,* 2004; *Lee et al.,* 2005; *Pehkonen et al.,* 2005; *Khatri* and *Drăghici,* 2005; *Yi et al.,* 2006).

Pavlidis et al. (2004) use geometric mean to calculate the significance of the genes in the gene set. Gene Set Enrichment (GSE) analysis, proposed by *Mootha et al.,* (2003) and improved by *Subramanian et al.,* (2005) uses an enrichment score based on a Kolmogorov-Smirnov test statistic. GSEA has been extended (*Barry et al.,* 2005; *Huang et al.,* 2009) to cover multiclass, continuous and phenotypes, and more test statistics such as Wilcoxon and hypergeometric.

Levine et al. (2006), *Efron* and *Tibshirani,* (2006), *Nam* and *Kim,* (2008) as well as *Ackermann* and *Strimmer,* (2009) rigorously and thoroughly evaluated the performance of different tests on simulated data. Only limited supporting evidence on real data was provided as this required manually curated gold standard. Recently, GSE tests have been applied to gene sets representing miRNA or TF target genes.

2.2.1 Identification of miRNA/TF activity changes

Several authors suggest that the miRNA/TF activity changes can be inferred from the expression levels of downstream target genes. *Farh et al.* (2005) show that the miRNA activity in specific tissue can be detected by analyzing the mRNA expression measurements it regulates. They applied the Kolmogrov-Simirnov test to determine whether the miRNA target genes were expressed at lower levels than controls (*Volinia*

et al., 2010). *Sood et al.* (2006) found that the tissue-specific human miRNA has a widespread effect on the expression levels of mRNAs. They compared the expression changes between the miRNA predicted target gene set and background gene set in the same tissue applying the Wilcoxon test.

Similar approaches to explore the association between miRNAs and their target gene set were also been employed by *Arora et al.* (2008), *Cheng et al.* (2008) and *Volinia et al.* (2010). They combined the miRNA target gene predictions with expression profiles to determine the miRNA activity. *Ott et al.* (2011) analyze the miR-29 family miRNA impact on the expression levels of the downstream genes in postnatal aortic development. They show that 20 out of 30 miRNAs found to be significant applying the Wilcoxon test were also found to be down-regulated experimentally (*Ott et al.,* 2011) . Recently, *Sohler et al.,* (2005), *Liu et al.,* (2010) and *Essaghir et al.,* (2010) identified the activity of TFs by analyzing whether the TF target gene sets are enriched among a list of DEGs using a hypergeometric test.

2.2.2 Statistical hypothesis testing methods

The following state-of-the-art statistical methods have been proposed for overrepresentation analysis of gene sets derived as regulator (miRNA and TF) target genes. These tests are applied to calculate the significance of a given regulator as p-value of the observed overrepresentation of its target set among the differentially expressed genes.

Wilcoxon rank sum test

The Wilcoxon nonparametric rank-sum (WR) method (*Sood et al.,* 2006; *Gsponer et al.,* 2008; *Ott et al.,* 2011) is applied to test whether the regulator targets ($x_1, x_2 \ldots x_m$) exhibit significant rank differences in comparison to other (non-targets, $y_1, y_2 \ldots y_n$) genes. For WR test, the ranks can be derived by sorting the genes based on either their absolute or signed log fold changes (Figure 6). If the rank distributions of targets and non-targets are significantly different the null hypothesis will be rejected. Then, targets of the tested regulator exhibit greater log fold changes than non-targets and the regulator is referred to as active according to the test. The results of WR test statistic are p-values as a measure of significance of the observed change in means (see *Mann* and *Whitney,* (1947) and *Lehmann* (1975) for an overview).

The WR test statistics is calculated as:

- Merge the data and rank all observations (smallest to largest) from 1 to $m+n$.
- Calculate the test statistic (W), essentially the sum of ranks from samples m.
- Assume that the populations (i.e., m and n) have the same continuous distribution, then statistic ($Z = \frac{w-\mu}{\sigma} \sim N(0,1)$ - normal approximation) has a mean and standard deviation given by:

$$\mu = m(m + n + 1)/2$$

$$\sigma = \sqrt{\frac{mn(m+n+1)}{12}}$$

Kolmogorov-Smirnov test

Whether or not the distributions of (miRNAs and TFs) target and non-target genes are shifted with respect to each other can also be tested by another non-parametric test, the Kolmogorov-Smirnov (KS) test. The two sample KS test determines whether the two data samples (i.e., regulator targets ($x_1, x_2 \ldots x_m$) and non-target ($y_1, y_2 \ldots y_n$)) come from the same distribution. In this case, the KS statistic is calculated as:

$$D_{n,m} = sup_x |f_{1,n}(x) - f_{2,m}(x)|$$

Where $f_{1,n}$ and $f_{2,m}$ are the empirical cumulative distribution functions (cdfs) of the non-targets and regulator-target sample, respectively and $\sup_x$ is the *supremum* of the set of distances (see *Siegel,* 1956, *Boes,* 1974, *DeGroot et al.,* 1991 and *Nikiforov,* 1994 for details). The null hypothesis is rejected at given threshold level α if

$$\sqrt{\frac{nm}{n+m}} D_{n,m} > K_\alpha$$

Both WR and KS tests do not require the selection of thresholds. Both tests have not yet been applied to TF activity detection, only to predict transfecting miRNAs (*Sood et al.,* 2006; *Tu et al.,* 2009; *Volinia et al.,* 2010; *Ott et al.,* 2011).

Analysis of variance (ANOVA) test

The Analysis of variance (ANOVA) is applied to test the heterogeneity of means by analysis of set variances under the assumption that the two sampled sets (such as regulator-target and non-target genes) are normally distributed. The results for ANOVA test are p-values that are calculated using the F-statistic/distribution (see *Hoang,* 2006 and *Miller,* 1997 for details). For two samples ANOVA is equivalent to the t-test.

Hypergeometric test

For a given regulator the p-value is computed according to the hypergeometric (HG) formula:

$$p - \text{value} = \frac{\binom{m}{x}\binom{N-m}{k-x}}{\binom{N}{k}}$$

Where *N* is the number of DEGs in a given chip measurement, *m* is the number of DEGs filtered based on a given regulated gene threshold value, *k* is the number of given regulator-target genes and *x* is the number of regulator targets among the filtered DEGs (*m*).

The cumulative distribution function (cdf) refers to a sum of probabilities associated to HG test. To compute a cdf we many need to add one or more probabilities:

$$p - \text{value (less than or equal to } x) = \sum_{i=0}^{x} \frac{\binom{m}{i}\binom{N-m}{k-i}}{\binom{N}{k}}$$

$$p - \text{value (greater than } x) = 1 - \sum_{i=0}^{x} \frac{\binom{m}{i}\binom{N-m}{k-i}}{\binom{N}{k}}$$

Bootstrap sampling

The Bootstrap sampling is used to calculate the statistic for two samples (such as regulator-target and non-target genes) drawn in some way (randomly) from the original data. The results of Bootstrap test statistic are p-values as a measure of significance to the difference in means of the two samples using for instance, two-sample ANOVA (see *Efron* and *Tibshirani,* 1993 for details).

Average Fold Change

The Average Fold change (FC-score) of a regulator activity is defined as the difference of the average mean expression levels between its targets (T_{avg}) and non-targets (nT_{avg}). A positive FC score indicates that the target genes of a regulator tend to be expressed at higher levels than non-targets genes. The higher the FC score, the stronger the activation effect of a regulator on its targets (see *Cheng et al.,* 2009 for a similar approach).

$$FC = T_{avg} - nT_{avg}$$

Average gene rank

The average gene rank (FCR-score) of a regulator activity is defined as the difference of the average rank between its targets (T_{ravg}) and non-targets (nT_{ravg}). The genes ranks were derived by sorting them based on their absolute or signed fold changes (Figure 6) (see *Cheng et al.,* 2009 for a similar approach).

$$FCR = T_{ravg} - nT_{ravg} = \frac{\sum_{i=1}^{n} t_i}{t} - \frac{\sum_{i=1}^{j} t_j}{nt}$$

Where t and nt represent the number of a given regulator targets and non-targets. And t_i and t_j represent the ranks of a given regulator target (t_i) and non-target (nt_j).

2.3 Named entity recognition system

Recent technological advances have contributed to the large volume of scientific literature discussing the role of genes and proteins and interactions between them. By automatically identifying the names of biomedical entities from texts and map these to database identifiers, it becomes possible to discover new associations among those entities of interest (*Fundel et al.,* 2005). For a detailed review see *Cohen and Hersh,* 2005.

Named entity recognition (NER) refers to the task of detecting named entities (NE) in the literature (*Ananiadou et al.,* 2006). The detection of biomedical entities such as gene, protein and miRNA names in the texts is not straightforward, despite the availability of miRNA, gene or protein nomenclature published in several databases such as miRBase, HUGO, Entrez Gene (see subsequent database section). These databases do not address underlying characteristic issues of the entities in NER such as ambiguities, aliases and variations of gene, protein and miRNA names (*Seringhaus et al.,* 2008; *Griffiths-Jones,* 2008). However, the biomedical entity names show many common features such as brackets ([], {}, ()), upper case, dash, comma, hyphen, slash and digits in the scientific literature and databases as well (*Ananiadou et al.,* 2006; *Griffiths-Jones,* 2008).

This section gives a brief introduction into miRNA, gene and protein naming conventions. It then describes the approaches to NER system followed by a string matching algorithm that can be used to detect the biomedical names in the texts including performance evaluation measures.

2.3.1 miRNA, gene and protein naming conventions

The nomenclature of miRNAs and especially proteins as well as genes has evolved over time (*Fundel et al.,* 2006; *Griffiths-Jones,* 2008) and various naming conventions have been and are used in databases and in the scientific literature. For genes and proteins but also miRNAs typically several synonyms are in use. Unfortunately, synonyms often overlap with other synonyms (of other objects) or with names and abbreviations for diseases, species, experimental techniques, and even general English words (*Fundel et al.,* 2006). For instance, for the gene ADCY10 (Entrez Gene identifier 55811) more than 10 additional synonyms are known not taking into account orthographical variations, such as usage of hyphens and slashes (*Fundel et al.,* 2006; *Jensen et al.,* 2006; *Erhardt et al.,* 2006; *Ananiadou et al.,* 2006). In comparison, miRNAs naming conventions have been described early and appear to be quite simple as miRNA names are based on sequential numerical identifiers (e.g., miR-1, miR-2 ... miR-101, etc.) and a prefixed species identifier (e.g. hsa-miR-100) (*Ambros et al.,* 2003; *Griffiths-Jones,* 2004; *Griffiths-Jones et al.,* 2008).

The following conventions for miRNA naming are used:

(1) The predicted stem-loop portion of the primary transcript is named by a 3 or 4 letter species prefix and a numerical suffix (e.g. hsa-mir-100 in Homo sapiens). Whereas, the name of the excised ~22 nucleotide sequence (mature miRNA)

contains the same mir, prefix and suffix as stem-loop but with capital miR (e.g. hsa-miR-100).

(2) Orthologous miRNA sequences in different species are assigned the same names (e.g. mmu-miR-100 in Mus musculus, rno-miR-100 in Rattus norvegicus).

(3) Mature miRNA sequences can be expressed from each arm of the hairpin precursor (Figure 1). They are distinguished by additional suffixes (e.g. hsa-miR-1224-5p (5′arm) and hsa-miR-1224-3p (3′arm)). Previously, they also have been named for instance miR-142-sense (s) (5′arm) and miR-142-anti-sense (as) (3′arm). In some cases, the asterisk has been used to denote the less predominant form (e.g. hsa-miR-100*).

(4) Distinct hairpin loci in a given organism that give rise to identical mature miRNA sequences are assigned names with additional numeric suffixes (e.g. hsa-mir-101-1 and hsa-mir-101-2 indicating two genomic loci of the miRNA hsa-miR-101).

(5) Related hairpin loci that give rise to related mature miRNA sequences with only one or two base changes are assigned letter suffixes of the form (e.g. hsa-mir-10a and hsa-mir-10b are similar sequences).

Unfortunately, these conventions are not strictly followed in scientific publications. If complete names are used, e.g. hsa-miR-1224-5p, the author likely means the 5′arm predominant mature form of human miRNA-1224. On the other hand, an incomplete form e.g. miR-1224 could mean precursor or mature microRNAs, the 3′ or the 5′ variant or an unspecified variant of microRNA 1224 in some species depending on the context.

In addition, there are many naming problems: For some organisms fairly different naming conventions are used (*Griffiths-Jones et al.,* 2004; *Griffiths-Jones et al.,* 2008). For instance, in plants, miRNA names are of the form MIR472 (in Arabidopsis thaliana) and only letter suffixes are used to represent distinct hairpin loci expressing related mature miRNA sequences (*Griffiths-Jones et al.,* 2008). Viral miRNAs names are based on the gene locus from which the miRNAs derive (e.g. ebv-mir-BARTT8 is a miRNA from the BART locus of the Epstein-Barr virus (ebv)) (*Griffiths-Jones et al.,* 2008). Capitalisation of names should not always be relied on to confer information, such as mir and miR distinguishing between precursor and mature forms (*Griffiths-Jones et al.,* 2004). lin-4 and let-7 miRNAs are the apparent exceptions to the generic scheme (*Griffiths-Jones et al.,* 2008).

2.3.2 Approaches to named entity recognition

Several approaches to NER have been proposed that can be classified into four main categories.

(1) The dictionary-based approach matches database or dictionary names in the literature. The extraction performance depends on the comprehensiveness of the entries in the dictionary. *Hanisch et al.* (2003) construct the gene and protein names dictionary by merging the HUGO Gene Nomenclature Committee (HGNC) (*Bruford et al.,* 2008) and OMIM (*Amberger et al.,* 2010) databases. *Tsuruoka*

and Tsujii. (2004) curate the protein dictionary by expanding it with morphological variations of protein names.

(2) The rule-based approach constructs rules either manually or automatically to match against entities of interest in the literature. *Fukuda et al.* (1998) identify the protein names from texts by utilizing manually curated rules and patterns.

(3) The machine-learning based approach is mainly based on developing statistical models for the identification of biomedical entity names. *Collier et al.* (2000) implement the supervised learning method with hidden Markov Model (HMM) to extract biomedical domain terminologies from texts. *Zhou et al.* (2004) combine the HMM with various features such as morphological patterns, parts-of-speech, special verb trigger and name aliases to recognize biomedical entity names from the scientific literature.

(4) The NER approaches can also be combined into a hybrid approach to deal with different aspects of NER and have their own advantages and disadvantages. *Tanabe* and *Wilbur* (2002) integrate the statistical and rule-based strategies to extract gene and protein names. *Mika* and *Rost* (2004) combine the dictionary and rule based approaches for the recognition of protein names in texts.

For more detail and progress in gene and protein name detection using these approaches, refer to *Ananiadou* and *Mcnaught* (2005).

2.3.3 String matching algorithm

Aho-Corasick

The Aho-Corasick is an extension of the Knuth-Morris-Pratt algorithm developed by *Aho* and *Corasick,* 1975. It is an exact string matching algorithm that locates the occurrences of any pattern of a set $p_1, p_2 \ldots p_k$ within an input text $t_1, t_2 \ldots t_m$ of size m. It uses the string-matching automaton called the *Aho-Corasick automaton* to build tree-like deterministic finite automata (DFA) from the set of strings (e.g., dictionary keywords) and then scans the input texts t_i for all occurrences of the all set of patterns p_j (*Navarro* and *Raffinot,* 2002). This allows the algorithm to process the input text string in a single pass and perform multiple pattern searches across text as well (*Navarro* and *Raffinot,* 2002). As a result, the complexity of the algorithm is linearly proportional to the pattern plus searched text size. An important advantage of the Aho-Corasick algorithm is that once the DFA has been constructed it can be used to find occurrences of any of a finite number of keywords in an arbitrary text string without having to reconstruct it (*Aho* and *Corasick,* 1975).

2.4 Performance assessment

- **Precision:** In the field of biomedical NER system, precision is defined as the fraction of retrieved entities that are relevant.

$$precision = \frac{tp}{tp + fp}$$

- **Recall:** It is defined as proportion of the entities that are relevant to the system that are successfully retrieved.

$$recall = \frac{tp}{tp + fn}$$

Where *tp* is the number of true positive named entities (NEs) that the NER system has identified correctly and *fn* is the number of false negative NEs which were not tagged as positive instances but should have been. The *fp* is the number of false positive NEs that the system has incorrectly identified.

- **F-measure:** It is defined as a harmonic mean of the precision (or sensitivity) and recall (or specificity):

$$F - measure = \frac{precision * recall}{precision + recall}$$

- **Area under the receiver-operating characteristic:** The area under the ROC curve (AUROC, also denoted as AUC) is a summary measure that summarizes the tradeoff between true positive and false positive rate (*Prill et al.,* 2010). It is equal to the probability that a classifier will rank a positive example higher than a negative one (*Fawcett,* 2006). An AUC score of 1 represents an optimal classifier (i.e., perfect ranker) while AUC of 0.5 represents a random classifier (*Fawcett,* 2006), respectively.

3. Databases

3. Databases

Several specialized databases have been developed in the past few years which aim to collect comprehensive information on microRNAs (miRNAs), transcription factors (TFs), genes or proteins, taxonomy, protein-protein interactions (PPIs), miRNA/TF-gene interactions and gene expression repositories. In this section, we briefly explain these databases.

3.1 miRNA gene database

The miRBase serves as a centralized resource for the miRNA sequences and annotations (*Kozomara* and *Griffiths-Jones,* 2011). It contains over 16,000 mature miRNAs and 15,000 miRNA gene loci in different species. The main objective is to maintain a consistent naming scheme by which a unique name can be assigned to each miRNA (*Kozomara* and *Griffiths-Jones,* 2011; *Ambros et al.,* 2003). A web interface has been provided to submit a newly discovered miRNA sequence for naming in over 140 species. As a proof-of-concept for miRNA annotation the database has been integrated with RNA deep-sequencing results (*Kozomara* and *Griffiths-Jones,* 2011). miRBase is becoming an important tool for the miRNA information. From its inception, the number of miRNA sequences in the database has risen significantly (Figure 2).

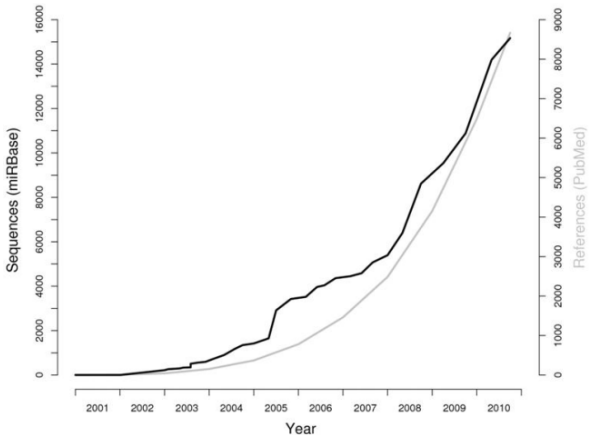

Figure 2: Growth of the miRBase miRNA sequences. The number of miRNA sequences (black) and the number of miRNA-related scientific articles in PubMed (grey). Figure taken from *Kozomara* and *Griffiths-Jones* (2011).

3.2 miRNA-gene regulatory interactions

In spite of the progress in miRNA target prediction programs, the need for a collection of experimentally validated miRNA-target gene pairs led to the development of many databases (Table 4).

TarBase is the first resource to provide manually curated miRNA-target gene interactions in different animal and plant species (*Sethupathy et al.,* 2006;

Papadopoulos et al., 2009). It documents over 1,300 miRNA-gene interactions extracted from over 200 scientific articles (Table 2). For each interaction, a brief description of validation experiment has been provided. The database is linked to other resources such as SwissProt (*Wu et al.,* 2006), HUGO (*Bruford et al.,* 2008), UCSC (*Mangan et al.,* 2008) and Ensembl (*Flicek et al.,* 2010) to extend miRNAs and their target gene information.

Organism	Number of papers	Number of entries	Microarray data	pSILAC data
Homo sapiens	110	285	328	474
Mus musculus	28	105	13	-
D. melanogaster	23	77	-	-
C. elegans	18	14	-	-
Plants	21	30	-	-
Danio rerio	1	1	-	-
Rat	2	2	-	-
Total	203	514	341	474

Table 2: List of all TarBase 5.0 entries. TarBase contains 1,333 entries describing regulatory interactions between miRNAs and their target RNA genes in different species extracted from over 200 scientific articles. Table taken from the study conducted by *Papadopoulos et al.* (2009).

Xiao et al. (2009) developed miRecords database that contains manually curated miRNA target gene interactions in different animal species from the scientific literature. They collected over 1,100 miRNA-target pairs (Table 3) and documented a brief description of experimental conditions that are used to validate those interactions. Additionally, the database has been integrated with 11 target predicted programs to provide a comprehensive list of target genes for a miRNA. *Jiang et al.* (2009) developed miR2Disease database that contains several human disease related miRNA-target genes pairs derived manually by surveying the scientific literature. *Hsu et al.* (2011) developed miRTarBase, a miRNA-target gene relation database. They manually curated over 3,500 miRNA-gene pairs from the research articles relevant to the functional studies of miRNAs. *Shahi et al.* (2006) developed Argonaute database that contains miRNA-regulated targets and their origin information derived from the scientific literature and other published databases.

Databases	miRNAs	Targets	Pairs	Low-throughput Experiments	Human miRNAs	Human Targets	pairs
TarBase	128	570	626	279	62	415	458
miRecords	301	902	1135	639	125	651	778

Table 3: Comparison between TarBase and miRecords. miRecords manually collects much more miRNA-target interactions than TarBase database. Table taken from the study conducted by *Xiao et al.* (2009).

The databases such as miRGen (*Megraw et al.,* 2007), miRGator (*Nam et al.,* 2008), miRDB (*Wang et al.,* 2008), microRNA.org (*Betel et al.,* 2008) and

miRNAMap (*Hsu et al.,* 2008) provide miRNA target information by integrating target prediction programs (such as PicTar, PITA, TargetScan and miRanda). *Nam et al.* (2009) developed MMIA that combines the predicted miRNA-target genes with miRNA-mRNA expression studies.

Databases	Web links
MD: miRBase	http://microrna.sanger.ac.uk/
MD: TarBase	http://diana.cslab.ece.ntua.gr/tarbase/
MD: miRecords	http://mirecords.umn.edu/miRecords/
MD: miR2Disease	http://www.mir2disease.org/
MD: miRTarBase	http://miRTarBase.mbc.nctu.edu.tw/
PD: Argonaute	http://www.ma.uni-heidelberg.de/apps/zmf/argonaute/
PD: miRGen	http://www.diana.pcbi.upenn.edu/miRGen.html
PD: miRGator	http://genome.ewha.ac.kr/miRGator/miRGator.html
PD: miRDB	http://mirdb.org/miRDB/
PD: microRNA.org	http://www.microrna.org/microrna/home.do
PD: miRNAMap	http://mirnamap.mbc.nctu.edu.tw/
PD: MMIA	http://129.79.244.122/~MMIA/index.html

Table 4: Databases for miRNAs and their target gene associations. To identify the target genes of miRNAs, several databases have been developed. Most databases collect miRNA-target gene association by manually surveying the published scientific articles (MD). Some databases provide the target genes of a miRNA by integrating state-of-the-art miRNA-target prediction programs (PD).

3.3 TF-gene regulatory interactions

The regulation of transcription initiation by TFs and *cis*-regulatory elements is a major regulatory step in the control of gene expression (*Hochheimer* and *Tjian*, 2003). A better understanding of the interaction between TF and *cis*-regulatory elements remains an open challenge (*Xu et al.,* 2010). Several experiments have been performed to profile the TF-DNA interactions (*Iyer et al.,* 2001; *Johnson et al.,* 2007; *Vogel et al.,* 2007).

Several repositories have been established for organizing the TFs and their target gene information. For instance, TRANSFAC (*Wingender et al., 2000*, *Matys et al.,* 2006), JASPAR (*Sandelin et al.,* 2004), TRED (*Zhao et al.,* 2005), PAZZAR (*Portales-Casamar et al.,* 2007), DBD (*Wilson et al.,* 2008) and TrSDB (*Hermoso et al.,* 2004) collect TF-target gene associations for different species. RegulonDB contains experimentally validated and manually curated TF-gene regulatory interactions in *E. coli* (*Huerta et al.*, 1998, *Gama-Castro et al.,* 2011).

Abdulrehman et al. (2011) developed YEASTRACT that contains over 48,000 TF-target associations in *S. cerevisiae*. *MacIsaac et al.* (2006) analyzed the ChIP-chip data to determine the TF binding sites in yeast. ProTF (*Bai et al.,* 2010) and RegTransBase (*Kazakov et al.,* 2007) collect regulatory interactions in prokaryotes. *MacArthur et al.* (2009) integrate over 21 TFs ChIP experiments that are performed using blastodermal cells.

The databases such as plantTFDB (*Guo et al.,* 2008; *Zhang et al.,* 2010), DPTF (*Zhu et al.,* 2007), RARTF (*Iida et al.,* 2005), PlanTAPDB (*Richardt et al.,* 2007), GRASSIUS (*Yilmaz et al.,* 2008) and PlnTFDB (*Pérez-Rodríguez et al.,* 2009) provide resources of TFs nomenclature, sequence data, genomic localization and functional annotation in plant species.

3.4 Gene and protein databases

Recently several databases have organized organism-specific genes and proteins information. The HUGO Gene Nomenclature Committee (HGNC) provides unique symbols/names to human genes maintained by *Seal et al.* (2011). They collected over 30,000 gene information including approved gene nomenclature, symbols and aliases. Several biomedical databases including Ensembl (*Flicek et al.,* 2010), Entrez-Gene (*Maglott et al.,* 2011), OMIM (*Amberger et al.,* 2009), UCSC (*Fujita et al.,* 2011) and UniProt (*Boutet et al.,* 2007) have also integrated the HGNC gene symbols. The MGD (Mouse Genome Database) combines the genetic, genomic and biological information in mouse (*Bult et al.,* 2008).

The Entrez-Gene maintains the unique identifiers assigned to genes in different organisms (*Maglott et al.,* 2011). The database has provided the detailed gene information including nomenclature, genomic location and their products (i.e., proteins). UniProtKB/Swiss-Prot provides the curated protein information for different species (*Boeckmann et al.,* 2003; *Boutet et al.,* 2007). Ensembl database maintained by *Flicek et al.,* (2010) combines the genomic information for over 35 species including human, mouse and rat.

3.5 Protein-protein interaction databases

Protein-protein interactions (PPIs) play an important role in understanding the functions of proteins and their activity in cellular processes. Several experimental approaches like the yeast two-hybrid (Y2H) system (*Chien et al.,* 1991; *Legrain* and *Selig,* 2000), X-ray crystallography or tandem affinity purification (*Rigaut et al.,* 1999; *Puig et al.,* 2001) have been performed to study the interaction between protein pairs. Several databases have collected these PPIs in different species (Table 5).

Keshava et al. (2009) developed HPRD (Human Protein Reference Database) that contains PPIs derived either from the biomedical literature or high-throughput experiments. The STRING database provides PPIs of either physical or functional association in over 600 different species (*von Mering et al.,* 2003; *Jensen et al.,* 2009; *Szklarczyk et al.,* 2011). *Szklarczyk et al.* (2011) collect the known or predicted PPIs from different sources including literature, high-throughput experiments and co-expression profiling studies. *Kerrien et al.* (2007) developed IntAct that contains manually curated binary interactions derived from the literature. *Xenarios et al.* (2002) developed the DIP (Database of Interacting Proteins) that collects experimentally validated PPIs.

3.6 Taxonomy database

The National Center for Biotechnology Information (NCBI) taxonomy database contains over 150,000 organism names/symbols and their aliases (*Sayers et al.,* 2009).

Databases	Web link
HPRD	http://www.hprd.org/
STRING	http://string-db.org/
IntAct	http://www.ebi.ac.uk/intact/main.xhtml
PINT	http://earth.liv.ac.uk/pint/Help.htm
DIP	http://dip.doe-mbi.ucla.edu
GO	http://www.geneontology.org/
KEGG	http://www.genome.jp/kegg/pathway.html
Pathguide	http://pathguide.org

Table 5: Gene and protein interaction databases. Several experimental approaches have been applied to study the gene and protein interactions. Shown are the databases that collect gene and protein-protein interaction information in different species.

3.7 Predefined gene set databases

Signal transduction pathways are often represented as cascades of proteins that regulate cellular processes like growth, survival and proliferation including gene expression (*Glaab et al.,* 2010; *Anjum* and *Blenis,* 2008). In this context, several databases have been developed to provide information about genes and proteins corresponding to specific cellular pathways and processes such as Gene Ontology (GO) (*Harris et al.,* 2004; *Berardini et al.,* 2010), Kyoto Encyclopedia of Genes and Genomes (KEGG) (*Kanehisa* and *Goto*, 2000), BioCarta (*Nishimura,* 2001), Reactome (*Joshi-Tope et al.,* 2005; *D'Eustachio*, 2011), The Molecular Signatures Database (MSigDB) (*Subramanian et al.,* 2005) and Pathguide (*Bader et al.,* 2006).

GO established by the GO consortium members, aims to maintain a controlled vocabulary of terms (ontologies) for the representation of genes, proteins and their sequences across different model organisms (*Ashburner et al.,* 2000; *Berardini et al.,* 2010). GO is clustered into three groups: biological processes contain information about processes to which a gene/protein contributes such as cellular growth or maintenance (*Ashburner et al.,* 2000). Molecular function provides the information about protein functions such as biochemical activity. Cellular component provides the information where a protein shows an activity change such as nuclear membrane, plasma membrane or Golgi apparatus (*Berardini et al.,* 2010).

The KEGG database integrates the gene and protein interactions with molecular networks and cellular pathways derived from other biological databases such as GenBank, PATHWAY and GLYCAN (*Kanehisa* and *Goto,* 2000; *Kanehisa et al.,* 2004; *Hashimoto et al.,* 2005). *Bader et al.* (2006) developed Pathguide that provides comprehensive information about biological (metabolic and signalling) pathways and gene regulatory interactions (Table 6).

Category	Number of databases
Protein-protein interactions	79
Metabolic pathways	43
Signalling pathways	41
Pathway diagrams	22
Transcription factors/gene regulatory networks	20
Protein-compound interactions	14
Genetic interaction networks	5
Protein sequence focused	12
Other	11

Table 6: Pathguide database statistics. Pathguide, a meta-database that collects information on over 150 published cellular pathways and gene/protein interaction databases. Table taken from the study conducted by *Bader et al.* (2006).

3.8 Gene expression databases

With the advent of high-throughput technologies it becomes possible to determine the mRNA levels of all genes of an organism under various experimental conditions (*Chua et al.,* 2006; *Faith et al.,* 2008; *Selbach et al.,* 2008; *Brazma,* 2009). Several databases have been established with the aim of collecting mRNA gene expression data from the published microarray studies (Table 7).

The databases Gene Expression Omnibus (GEO, *Barrett et al.,* 2007) and ArrayExpress (*Parkinson et al.,* 2006) contain gene expression repositories in different organisms. *Ringwald et al.* (2001) developed GXD (Gene Expression Database) that provides expression profiles in mouse species. *Miranda-Saavedra et al.* (2009), *Hackney et al.* (2002) and *Porter et al.* (2007) collect expression profiles related to the mouse hematopoesis system, stromal and stem cell lines. *Schulz et al.* (2009) developed FunGenES that contains expression profiles in mouse embryonic stem cells.

Databases	Web Link
GEO	http://www.ncbi.nlm.nih.gov/geo/
ArrayExpress	http://www.ebi.ac.uk/microarray-as/ae/
GXD	http://www.informatics.jax.org/
BloodExpress	http://hscl.cimr.cam.ac.uk/bloodexpress/
SCDB	http://stemcell.mssm.edu/v2/
StemBase	http://www.stembase.ca/?path=/
StroCDB	http://stromalcell.mssm.edu/
FunGenES	http://www.fungenes.org/index.html

Table 7: Databases of microarray gene expression profiles. Shown are the databases developed aim at collecting gene expression measurements in different species.

3.9 Scientific literature database

The recent biomedical advancements have witnessed significant scientific progress and discoveries. However, the acquisition of scientific knowledge contains in the literature is challenging due to its large volume and rapid growth (*Lu,* 2011). As a result, NCBI established PubMed that incorporates MEDLINE to archive peer-reviewed journals in the life sciences (*Sayers et al.,* 2009). To date, PubMed contains more than 20 million research articles published in over 4,000 scientific journals (Figure 3). Additionally, to ease the access to the published articles, several web-based tools have been developed such as RefMed (*Yu et al.,* 2010), iPubMed (*Wang et al.,* 2010), MedlineRanker (*Fontaine et al.,* 2009), MiSearch (*States et al.,* 2009), eTBLAST (*Errami et al.,* 2007) and HubMed (*Eaton,* 2006).

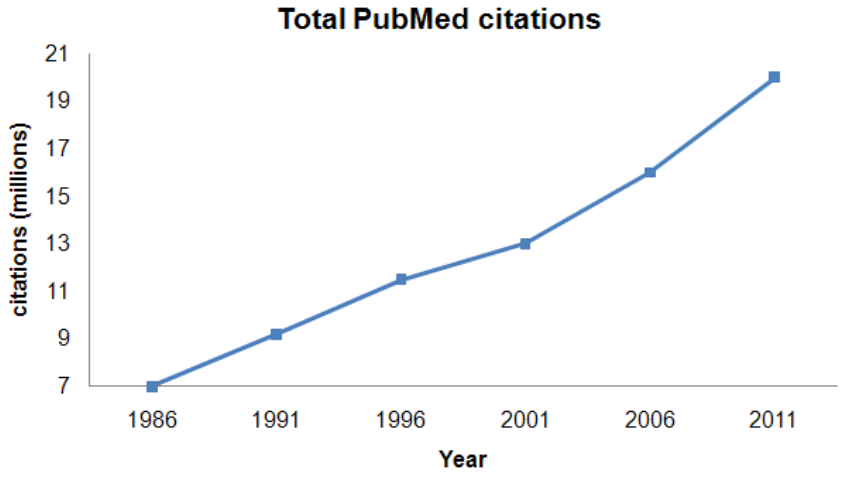

Figure 3: Growth of PubMed citations. The PubMed citations have increased significantly over the past few years. The PubMed database currently contains 20 million abstracts/papers in over 4,000 scientific journals. Figure adapted from the study conducted by *Lu* (2011).

4. Methods

To enhance the understanding of the regulation of gene expression, we address and solve three important, mutually dependent problems:

(1) To enable the proper application and assessment of enrichment methods, we compiled the current regulator target repositories as well as prediction tools and complemented them by large scale text mining (called as miRSel, *Naeem et al.,* 2010). See section: 4.1 for methods and 5.1 for results.

(2) To select the appropriate enrichment approach/method for detecting the active regulators, we conducted the first rigorous comparative assessment of 12 gene set enrichment tests (*Naeem et al.,* 2011). See section: 4.2 for methods and 5.2 for results.

(3) In order to investigate the mechanisms of gene regulation our approach MIRTFnet determines the experimental conditions where certain regulators become active and how they regulate the transcriptome via cascades of miRNAs, TFs or kinases (*Naeem et al.,* 2011). See section: 4.3 for methods and 5.3 for results.

4. Methods

4.1 miRSel: Automated extraction of associations between miRNAs and genes from the biomedical literature

The construction of a database of miRNA-gene co-occurrences via named entity recognition (NER) requires the compilation of miRNA, gene and protein name dictionaries as well as their association to database identifiers. The extraction performance depends on the completeness and uniqueness of the entries in the dictionaries. This section (based on the publication *Naeem et al.,* 2010) describes the steps required for the implementation and population of the miRSel database (Figure 4).

4.1.1 miRNA, gene, protein and taxonomy name dictionaries

The dictionaries for human, mouse and rat are compiled from several databases: HUGO Gene Nomenclature Committee (HGNC) (*Bruford et al.,* 2008), Mouse Genome Database (MGD) (*Bult et al.,* 2008), gene-centered information at NCBI (Entrez Gene) (*Maglott et al.,* 2007), Swiss-Prot Protein Database (Swiss-Prot) (*Boeckmann et al.,* 2003), miRGen (*Alexiou et al.,* 2009), miRBase (*Griffiths-Jones et al.,* 2008) and NCBI (*Sayers et al.,* 2009). The names, aliases, symbols, official names, synonyms, abbreviations, and database identifiers of taxonomy, proteins, genes and miRNAs from these databases have been merged into synonym dictionaries.

4.1.2 Extension and curation of the dictionaries

The next steps are extension and curation of the dictionaries. For proteins, we first complement the synonym lists with spelling variants, acronyms, abbreviations and long forms (e.g. *IL* ↔ *Interleukin*). Secondly, inappropriate synonyms or expressions that would lead to ambiguous or wrong identifications are identified and removed. A detailed description of the curation and the involved processing steps is given in (*Fundel et al.,* 2006).

4.1.3 Detection of miRNA in texts

In case of miRNAs we found that many miRNA names described in the literature are not yet contained in databases. Therefore, we detect miRNA names using a regular expression. This regular expression has been constructed to match all database contained synonyms and generic occurrences of miRNA names as described in the background method section on miRNA naming conventions (section 2.3.1) including species specific conventions (e.g. HUGO). The regular expression also covers frequent spelling variants mentioned in the texts (e.g. miR101b, miRNA-101b, microRNA-101b, microRNA101b, etc.) with and without species identifiers (e.g. hsa-miR-195 and miR-195). If detected miRNA synonyms are contained in public databases we map them to their database identifiers and, if possible, distinguish matches as stem loop sequence, mature sequence and gene family matches in miRSel.

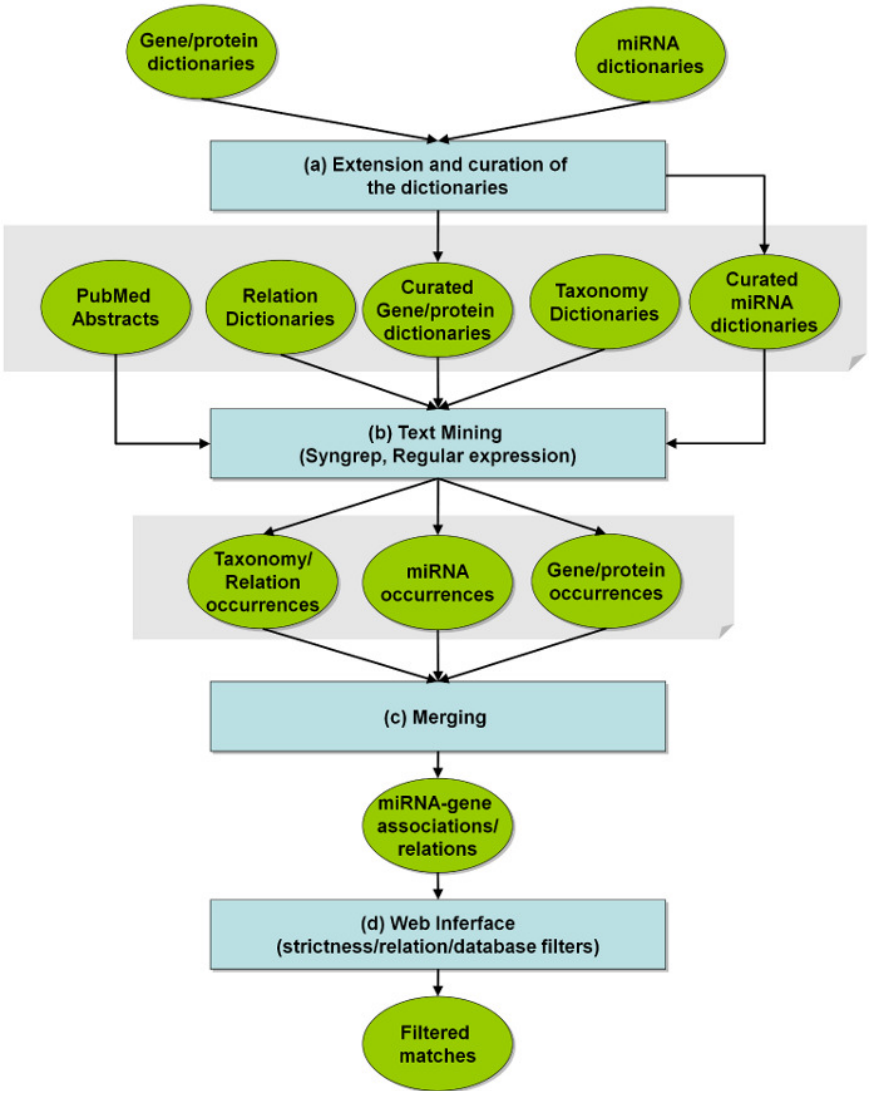

Figure 4: Workflow of the miRSel database. (a) The miRNA and gene/protein dictionaries/synonym lists for human, mouse and rat are extracted from several databases (e.g., HGNC, MGD, Entrez Gene, Swiss-Prot and miRBase) and extensively curated (see sections 4.1.1 and 4.1.2). (b) The miRNA, gene/protein, taxonomy and relation-term synonyms are then searched within PubMed abstracts by exact string matching tool syngrep (*Csaba*, 2008) (see sections 4.1.3, 4.1.4 and 4.1.5). (c) The occurrences of miRNA, gene/protein, taxonomy and relation-keyword synonyms are combined to infer miRNA-gene associations including miRNA-target relations and stored in the database. (d) Information on miRNA-gene pairs in the database can be retrieved via web interface (see section 5.1, http://services.bio.ifi.lmu.de/mirsel/). The web interface can be queried via different options, including miRNA, target, gene ontology and PubMed keyword queries. If multiple options are selected, the results are AND-combined. Several filters are provided to control recall vs. precision of the mining results (see section 5.1.2). To make sure the completeness of putative miRNA-gene associations the databases (e.g., TarBase, miRecords, and miR2Disease) have also been integrated into the miRSel database.

The database derived synonyms are summarized in Table 8. Only comparatively small number of distinct miRNA loci can actually be found in PubMed abstracts. In human, mouse and rat only 360 different miRNA loci were detected in miRNA-target pairs using the regular expression. Even fewer, only 280 different miRNA loci would have been detected based on database derived synonyms alone.

Species	miRNA			Proteins/genes	
	Mature (miR/miR*)	Stem-loop	Synonyms	Entities	Synonyms
Human	1026	162	43070	30120	473403
Mouse	767	133	32448	42130	460921
Rat	392	63	15662	39545	285483

Table 8: miRNA and gene/protein dictionaries. The identifiers and synonyms are extracted from different biological databases (such as miRBase, miRGen, HUGO, MGI, Entrez-Gene, Swiss-Prot), including manually collected miRNA identifiers for human, mouse and rat from literature. All dictionaries were processed to add frequently used synonym variants and to remove unspecific and inappropriate synonyms.

4.1.4 Detection of gene, protein and taxonomy names in texts

Protein, gene and taxonomy names/synonyms are detected in texts by string matching tool (called as syngrep, *Csaba*, 2008). syngrep uses the Aho-Corasick algorithm (*Aho* and *Corasick*, 1975, section 2.3.3) for fast matching, tolerates small synonym variations, and uses context resolution techniques to avoid and resolve ambiguities.

As mentioned above, miRNAs are matched using a regular expression. The scan of the organism specific miRNA, gene and protein synonym lists (more than 115K objects and 1.2M synonyms) against the entire PubMed (19M records, 66 Gb XML) requires about 30 minutes on a PC with 4 CPU-Cores.

The identification of named entities allows to identify miRNA-gene pairs both co-occurring in an abstract or in a single sentence. If not mentioned otherwise, we will focus on pairs extracted from single sentences as they are more reliable for extracting miRNA-gene associations. Information on relations or interactions between miRNA and genes is of interest for generating and analyzing network models of regulatory pathways. These pairs are extracted and stored in miRSel together with the PubMed abstracts and sentences where they have been found.

4.1.5 Detection of miRNA-gene target relations

We further compiled a list of 70 terms that are used to describe relations of interest between miRNA-gene pairs (Table 9). These 70 terms are indicative of five different types of relations, namely physical target, repression, co-expression, induction and cleavage. miRSel contains miRNA-gene associations of these five types which have been identified as tri-occurrences of a miRNA, a relation-term, and a gene or protein in a single sentence of PubMed abstracts. miRSel users can also retrieve abstract instead of sentence co-occurrences if recall is more important than precision.

Relation type	Synonyms
Physical targets	Target, targets, targeting, targeted.
Repression	repress, repression, down regulate, down-regulation, decreased activity
Co-expression	Deregulate, de-regulate(s), regulated, correlate(s), correlated, down modulate(s), down-modulation, deregulated expression, decreased expression.
Induction	Induction, up regulate(s), up regulation, increase(s) activity, increased activity, activation, activate(s).
Cleavage	Cleavage, cleaves, cleaved, processing shift.

Table 9: List of miRNA-gene target relations and their synonyms. Information on relation or interaction keywords between miRNAs and their target genes has been extracted by manually surveying the miRNA-related scientific literature and curated.

4.2 Rigorous assessment of gene set enrichment tests

As mentioned before, several statistical tests are available to detect the enrichment of differential expression in gene sets associated with biological processes (section 2.2). However, it is difficult to decide *a priori*, which processes will be affected in given experiments. In contrast to processes, miRNAs and TFs are amenable to direct perturbations, e.g. regulator over-expression or deletion experiments. As of second step, based on such perturbations and subsequent measurements in *E. coli*, *S. cerevisiae* and human, we assessed the ability of 12 different gene set enrichment (GSE) tests to detect the activity changes of miRNAs and TFs (*Naeem et al.,* 2011). We also analyzed the dependency of performance on the quality and comprehensiveness of the known regulator targets via an additional permutation approach. This section explains the GSE methods including datasets, testing scenarios and their pre-processing steps in the following subsections.

4.2.1 Datasets

Gene expression compendia: miRNA transfection studies

Several microarray experiments with overexpression of miRNAs have been performed to measure the global changes in the transcriptome or proteome. We collected 43 gene expression profiles of 18 different miRNA transfection studies in different human cell lines. *Selbach et al.* (2008) measured gene expression data in HeLa cells at 8h and 32h after miRNA overexpression of miR-155, miR-16 and let-7b. Expression profiles by *He et al.* (2007) include gene expression changes at 24h after miRNA overexpression of miR-34 family (i.e., miR-34a and miR-34b), in six different cell lines (e.g., HeLa, A549 H1-term and TOV21G H1-term). *Georges et al.* (2008) measured p53-inducible miRNAs, miR-192 and miR-215, at 10h and 24h after miRNA transfection in a human cell line (i.e., HCT116 Dicer -/- #2). Baek *et al.* (2006) measured the gene expression data in HeLa cells at 24h after miR-124, miR-1 and miR-181a transfection. We also use the dataset by *Grimson et al.* (2007) that measured gene expression data in HeLa cells at 12h and 24h after miRNA overexpression of miR-7, miR-9, miR-122, miR-128, miR-132, miR-133, miR-142 and miR-181a. All analyzes are based on comparing mRNA levels between transfection and control via $\log_2$ fold-changes.

Gene expression compendia: TF deletion and overexpression studies

In addition to the miRNA perturbation experiments, we also take large-scale TF perturbation (deletion (KO) or overexpression (OE)) experiments into account to investigate the influence of TFs on downstream target genes. A compendium of 907 *E. coli* microarray samples was taken from the M3D Database (*Faith et al.,* 2008). A compendium of 263 *S. cerevisiae* microarrays was obtained from the study by *Hu et al.,* 2007. *Hu et al.* systematically deleted 263 TFs in yeast, and compared each deletion strain with the wild type for genome-wide expression. We have also used the dataset by *Chua et al.,* (2006) that provides the microarray expression data resulting from the overexpression and/or deletion of 55 *S. cerevisiae* TFs. All analyzes are based on comparing gene expression levels between deletion/overexpression and control via $\log_2$ fold-changes. The microarray datasets contain basal gene levels that

can be quite different between experiments. To compensate for this, we transformed the absolute expression values into expression fold changes. Fold changes are computed by mapping each measured condition to one or more control conditions from the same experiment (*Küffner et al.,* 2011).

Gene regulatory networks: miRNA-target gene associations

Several computational algorithms have been developed to predict miRNA-target genes (see background method section). We obtained the human miRNA-target pairs predicted by PITA (*Kertesz et al.,* 2007), PICTAR (*Krek et al.,* 2005) and TargetScan (*Friedman et al.,* 2009). The PITA miRNA target predictions were compiled using a more stringent threshold (from -6 to -20) to reduce the number of false positive predictions.

In addition, several databases collect target genes of the miRNAs in different organisms (section 3.2). From miRSel (section 4.1, *Naeem et al.,* 2010) we obtained putative miRNA-gene associations and relations extracted from either biomedical abstracts by text mining (Table 12) or the curated databases (e.g., TarBase (*Papadopoulos et al.,* 2009), miRecords (*Xiao et al.,* 2009) and miR2Disease (*Jiang et al.,* 2009)).

Dataset	TFs	Targets	KO/OE TFs	Targets	Chips	References
E. coli (M3D)	167	1377	17	949	907	*Faith et al.,* 2008
S. cerevisiae (Y1)	114	1934	102	1527	263	*Hu et al.,* 2007
S. cerevisiae (Y2)	114	1934	48	1094	270	*Chua et al.,* 2006

Table 10: *E. coli* and *S. cerevisiae* expression compendia used in this study. High confidence 3425 *E. coli* interactions between 167 TFs and 1377 target genes were obtained from RegulonDB (*Gama-Castro et al.,* 2008). The Saccharomyces cerevisiae (*S. cerevisiae*) gold-standard network of 3940 interactions between 114 TFs and 1934 target genes was obtained from the study by *MacIsaac et al.* (2006). A compendium of 907 *E. coli* microarray expression profiles of 4296 genes was taken from the M3D Database (*Faith et al.,* 2008). In case of *S. cerevisiae*, two compendium of 263 (called as Y1) and 270 (called as Y2) microarray expression profiles of 5949 and 5473 genes were obtained from the study by *Hu et al.* (2007) and *Chua et al.* (2006). *Hu et al.* deleted (KO) 263 TFs in *S. cerevisiae*, and compared each deletion strain with the wild type for genome-wide expression, whereas *Chua et al.* provide microarray expression data resulting from the deletion (KO) and/or over-expression (OE) of 55 *S. cerevisiae* TFs. 102 out of 263 (Y1) and 48 out of 55 (Y2) *S. cerevisiae* TFs from the study by *Chua et al.* and *Hu et al.* were mapped to the used gold-standard regulatory interactions. 1527 out of 1934 and 1094 out of 1943 genes are targeted by Y1 and Y2 TFs dataset, respectively.

Gene regulatory networks: TF-gene regulatory interactions

E. coli TF-gene regulatory interactions (TF-GRIs) were obtained from RegulonDB (*Huerta et al.*, 1998; *Gama-Castro et al.,* 2011). RegulonDB contains experimentally validated and manually curated TF-GRIs. Recently DREAM5 uses the RegulonDB dataset to validate the predicted *E. coli* interactions (wiki.c2b2.columbia.edu/dream /index.php/D5c4). The Saccharomyces cerevisiae (*S. cerevisiae*) gold-standard TF-GRIs were obtained from *MacIsaac et al.,* 2006 who re-analyzed the *Harbison et al.,* 2004 ChIP-chip data to determine the binding locations of TFs. The *E. coli* gold-standard is considered more reliable than *S. cerevisiae* as suggested by *Narendra et al.* (2010).

4.2.2 Assessment of miRNA and TF activity

TF activity is regulated at the posttranscriptional level through changes in sub-cellular localization, therefore, it is challenging to measure TF activity directly (*Boorsma et al.,* 2008). To determine activity changes of miRNAs and TFs we apply several gene set enrichment approaches to test the null hypothesis (H_o) whether the expression levels of regulator downstream targets could be sampled from the background distribution of the remaining (i.e. non-target) genes (section 4.2.3). Our approach to assess gene set enrichment tests is depicted in Figure 6. Before introducing the applied enrichment approaches we describe in the following subsections how the standard of truth is derived and how sign annotations are used to treat the up- and down-regulation of target genes.

4.2.3 Standard of truth

In the proposed assessment scenario, we evaluate the ability of statistical tests to infer the identity of an experimentally perturbed (i.e. deleted or over-expressed) regulator from the expression of its target genes. Thus, the experimental annotation (identities of the perturbed regulators) represents the standard of truth. It is compiled into a label matrix that assigns a 1 or a 0 to a positive or a negative example, respectively. The number of examples and thus the size of the label matrix is |regulators|*|chips|. A positive example is given when the regulator r is perturbed (deleted or over-expressed) in the given chip c. Other measurements are considered negative examples for this regulator (Figure 6).

Since the inference of TF activities from their mRNA levels is not reliable, we exclude some TFs from the AUROC analysis based on mRNA levels. If for instance a perturbed TF does not exhibit a substantial fold-change it is unclear whether the perturbation was effective: it cannot be counted as positive in the label matrix without restrictions. The same holds for TFs that exhibit substantial fold-changes but have not been directly perturbed. Such a TF could be a direct or indirect target of a perturbed TF and cannot be regarded as true negative. If the fold-changes observed for non-perturbed (perturbed) TFs exceed (do not exceed) a predefined threshold, we will exclude it from the AUROC analysis. By varying this threshold, we can explore the performance dependency on the definition of positives and negatives.

4.2.4 Pre-processing of the data matrix

Before applying enrichment tests, the given gene expression measurements need to be pre-processed in one of two alternative ways. These are distinguished by whether or not we use *sign annotations* of interactions, i.e. the information that the TF activates (+) or inhibits (-) a given target gene.

Absolute-one-sided (H_o^{abs})

In this testing scenario, interaction signs ('+' up and '-' down regulation) are ignored. Enrichment tests are applied to absolute log fold changes, i.e. we evaluate the degree of differential expression in the target genes of regulators regardless of up- and down-regulation (Figure 5).

Signed-two-sided (H_o^{sign})

This scenario can only be applied to *E. coli* since only RegulonDB provides GRI sign annotations ('+' activation and '-' inhibition). We negate fold changes for target genes that are inhibited by the given regulator. Thus, all target genes of a regulator should either exhibit enrichment of positive fold changes (in case of increased regulator activity) or of negative fold changes (in case of decreased regulator activity). Enrichment at either tail of the distribution is then determined by two-sided tests (Figure 5).

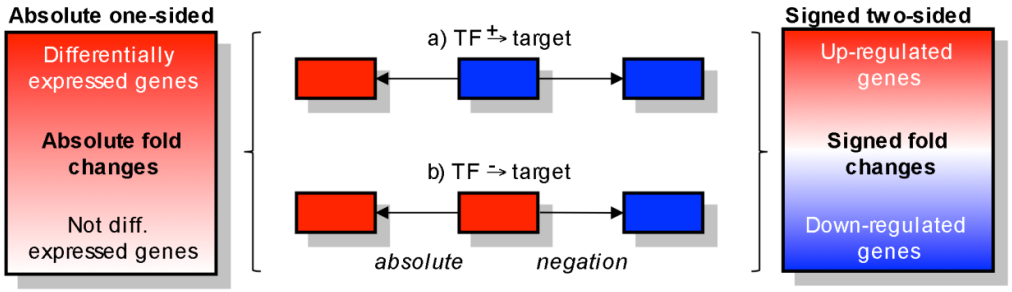

Figure 5: Pre-processing of the data matrix. Two null hypotheses, the absolute (H_0^{abs}, left) and the signed two-sided (H_0^{sign}, right) null hypotheses, can be tested after pre-processing the data matrix accordingly. For H_0^{abs}, expression profiles are transformed into absolute log fold changes. In case of *E. coli* TF-GRI where interactions are annotated as '+' for activation and '-' for inhibition, we also test H_0^{sign}. Here, we negate fold changes for target genes that are inhibited by the given regulator. Two-sided tests then detect positive or negative fold-changes corresponding to an increase or decrease, respectively, in regulator activity.

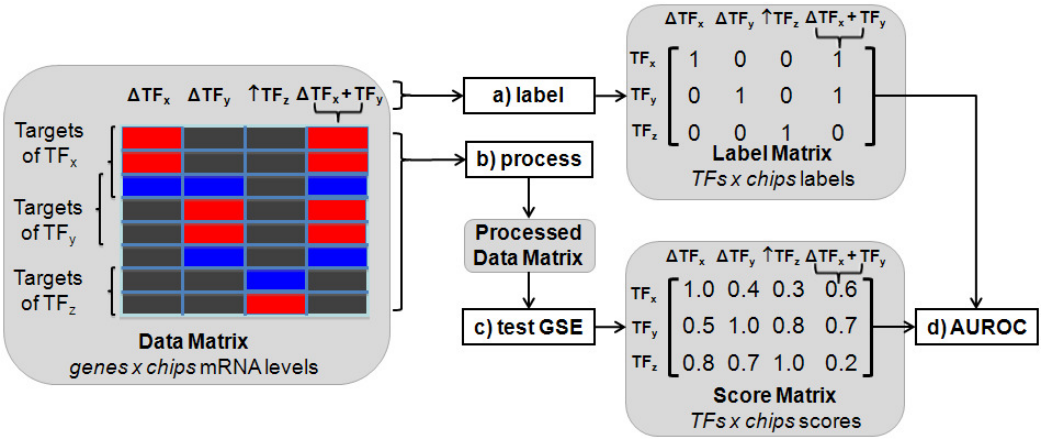

Figure 6: Overview of the miRNA/TF assessment test scenario. The data matrix consists of |*chips*| columns and |*genes*| rows where cells in the matrix represent mRNA expression levels. Chips are annotated by the experimental treatment, e.g. the perturbation (Δ=deletion or ↑=over-expression) of expression regulators, here exemplified by TFs. This annotation is compiled (a) into a label matrix that assigns a 1 (true positive) or a 0 (false positive) to a positive or a negative example, respectively (see section 4.2.3). A positive example is given when the regulator, for instance, TF_x is perturbed (e.g., deleted) in the given chip ΔTF_x. Other measurements (e.g., TF_y and TF_z) are considered negative examples for this regulator measurement. Perturbation of a regulator is expected to result in up- (red) or down-regulation (blue) of its target genes. (b) To test whether or not to use interaction signs or annotations, the data matrix is processed in one of two alternative ways (see section 4.2.4). Based on these settings, several gene set enrichment (GSE) tests are applied (c) to determine the activity of regulators based on the differential expression of their target gene set (see section 4.2.7). This results in a score matrix containing p-values or other test specific scores. For evaluation, the label matrix is compared to the score matrix to compute (d) an area under the receiver-operator characteristic (AUROC) curve (see section 4.2.5).

4.2.5 Performance assessment

Statistical tests as described in subsequent sections are applied to the processed data matrix (Figure 6). Test predictions are then evaluated against the standard of truth via the area under the receiver-operating characteristic (AUROC) (section 2.4). The AUROC compares continuous test scores (Figure 6: score matrix) against discrete regulator states (1=active, 0=inactive, compare Figure 6: label matrix). Thus, AUROC is a summary measure of the test's ability to consistently assign higher scores to active regulators and lower scores to non-active regulators based on given chip measurements. An AUROC score of 1 represents a perfect test or AUROC of 0.5 represents a random test, respectively.

4.2.6 Randomized testing

In addition to applying the tests to the data matrix, we also progressively randomized the set of regulator target genes to evaluate how much the performance of statistical methods depends on the quality of gold standards. We generate new target sets that are randomized by *x*% (where *x*=25, 50, 75…), i.e. by randomly selecting *x*% of the interactions in the gold standard and exchanging the true target gene in such an interaction by a random non-target gene. Tests are applied and evaluated as described above. An average AUROC is determined by repeating this procedure 100 times for each selected *x*.

4.2.7 Statistical hypothesis methods

We applied 12 different methods (Table 11, section 2.2) to test if the null hypothesis as described above should be rejected. In this case, distributions of regulator (miRNA and TF) targets and non-targets are significantly different.

Enrichment tests	
ANOVA	Two-sample ANOVA
WR	Wilcoxon's rank sum
KS	Kolmogorov-Smirnov
BT	Bootstrapping
CON	Consensus of all tests (see below for details)
HG-0.5	Hypergeometric, cut=0.5
HG-1.0	Hypergeometric, cut=1.0
HG-1.5	Hypergeometric, cut=1.5
FC	Average fold change
FCR	Average gene rank
MED	Median (see below for details)
FCRW	Average fold change rank weight (see below for details)

Table 11: List of statistical enrichment tests used in this study. Several statistical hypothesis testing methods have been applied to determine the activity of regulators based on their target gene sets as described in background method section. Since hypergeometric (HG) test requires a threshold parameter to select regulated genes, we applied the HG test to test the null hypothesis given regulated gene sets of different sizes compiled based on genes log fold values such as greater than 0.5, 1.0 or 1.5.

Average fold change rank weight

The average fold change rank weight (FCRW-score) of a regulator activity is defined as the difference of the combined average rank and expression levels between its targets (*T_rw*) and non targets (*NT_rw*). The ranks of genes are derived by sorting them based on their absolute or signed fold changes (Figure 6).

$$FCRW = T\_rw - NT\_rw = \frac{\sum_{i=1}^{n} w_i t_i}{\sum_{i=1}^{n} w_i} - \frac{\sum_{j=1}^{n} w_j nt_j}{\sum_{j=1}^{n} w_j}$$

Where w_i and w_j represent the rank and t_i and nt_i represent the fold changes of a target (i) and non-target (j) gene, respectively.

Median

The median (MED-score) of a regulator activity is defined as the difference of the median expression levels between its targets (*T_med*) and non-targets (*NT_med*).

$$MED = T\_med - NT\_med$$

Consensus prediction

A number of tests have been applied to a TF in a given experiment to test for over-representation of its targets among the DEGs. For each test, ranks of the regulators are determined by sorting them based on their scores. We define a consensus score (CON) to measure the regulator activity changes: the unweighted average of the ranks of a regulator determined by other statistical methods/tests as described above. This approach is called Borda count voting (*Borda,* 1781). For a given regulator *j*, the consensus score is calculated as:

$$CON = \frac{\sum_{i=1}^{n} R_{ji}}{n}$$

where n represents the number of tests applied to calculate the significance of a regulator j in a given experiment. Thus, R_{ji} represents the rank of a regulator j for a given statistical test i.

4.3 MIRTFnet: Analysis of miRNA regulated transcription factors

Patterns of gene silencing induced by miRNA are achieved by mRNA degradation or translational inhibition (section 1). Several miRNA transfection experiments have been performed to investigate the influence of miRNAs on their downstream target genes. We proposed the method MIRTFnet to identify miRNA controlled transcription factors (TFs) as active regulators.

MIRTFnet enabled the determination of active TFs in a miRNA induced expression measurements. For this purpose, we applied the selected gene set enrichment (GSE) statistical tests (section 4.2) to determine the activity of TFs and miRNAs that can change their activity in response to the transfecting miRNA. Based on the identified TFs, database (miRSel, section 4.1), computational predictions and the literature we constructed the regulatory models downstream of miRNA actions. Transfecting miRNAs are connected to active regulators via a network of miRNA-TF, miRNA-kinase-TF as well as TF-TF relationships (Figures 7 and 17).

This section (based on publication *Naeem et al.,* 2011) describes the MIRTFnet method including model of miRNA action and datasets.

4.3.1 Datasets

miRNA transfection studies

We obtained 43 gene expression profiles of 18 different miRNA (such as miR-155, miR-16, miR-34 and let-7b) transfection studies in different human cell lines such as HeLa, A549 H1-term and TOV21G H1-term (see section 4.2 for details).

miRNA-target gene associations

Human miRNA-gene associations were obtained from the databases and sequence prediction programs (Table 12, see section 4.2 for details).

TF-target gene associations

Human TF-gene regulatory relationships were predicted as described in (*Liu et al.,* 2008) using the position specific weight matrices (PWM) from the JASPAR database. We used relationships from the human genome browser at UCSC (http://genome.ucsc.edu/) (*Tu et al.,* 2009). Additionally, we collected TF-target gene associations from TRANSFAC (*Matys et al.,* 2003; *Matys et al.,* 2006) (ver. 2005), see Table 12. We refer to these TF-gene relations as JASPAR, UCSC, and TRANSFAC, respectively.

Databases	Regulators	Kind	Target genes	Pairs
DB: miRSel	486	miRNA	1969	7604
DB: TarBase	110	miRNA	837	1023
DB: miRecords	93	miRNA	614	772
DB: miR2Disease	176	miRNA	364	596
PR: PITA	640	miRNA	14065	307465
PR: PICTAR	163	miRNA	5975	44403
PR: TargetScan	249	miRNA	9446	110172
DB: UCSC	106	TF	3997	16688
DB: JASPAR	66	TF	12261	73878
DB: TRANSFAC	219	TF	304	794
DB: HPRD	462	Kinase	1800	4182

Table 12: Associations between regulators and their targets. Human miRNA/TF-gene regulatory interactions including kinase-TF associations were obtained from several databases (DB) and predictions (PR).

Protein-protein interactions

Human protein-protein interactions (PPIs) have been downloaded from the Human Protein Reference Database (http://www.hprd.org/) (*Keshava et al.,* 2009). Using PPIs, miRNA-gene associations and TF-gene relations, we compile the miRNA-kinase associations and kinase-TF including miRNA-kinase-TF physical interaction relationships, see Table 12. We compile all types of interactions into one gene network.

4.3.2 Determining active miRNAs and TFs

TFs might be activated or inhibited by modifications (e.g. phosphorylation) that cannot directly be detected by mRNA measurements. Activity changes of miRNAs and TFs can still be determined by analyzing their effects on target gene sets (section 4.2). The probability of this null hypothesis (H_o^{abs}) (p-value) can be derived by a number of statistical tests as described in section 4.2.

Here we have applied the selected GSE tests as described below. The resulting p-values are multiple testing corrected using the Benjamini and Hochberg method (*Benjamini and Yekutieli*, 2001). For corrected p-values of less than 0.05 the null hypothesis is rejected for the respective miRNAs and TFs. We refer to such regulators as active regulators in the tested experiment. Both miRNAs and TFs can be tested given lists of experimentally validated or computationally predicted targets.

TFs are also assumed to be active if they exhibit a fold change of at least two or less than 0.5 in a given expression experiment. Active miRNAs cannot be identified this way as they have not been measured on the arrays.

4.3.3 Statistical hypothesis methods

We applied the selected GSE tests (Table 13, section 2.2) to determine the activation of regulators in miRNA transfection measurements.

Enrichment tests	
WR	Wilcoxon's rank sum
KS	Kolmogorov-Smirnov
HG-1.0	Hypergeometric, cut=1.0

Table 13: List of enrichment tests used in MIRTFnet study. The Wilcoxon nonparametric rank-sum (WR) and Kolmogorov-Smirnov (KS) methods have been applied to test the null hypothesis (H_o^{abs}) as described in section 2.2. Nevertheless, both tests usually yield consistent results as found by e.g. *Gsponer et al.* (2008). MIRTFnet therefore reported TF activity changes only if they are identified by both tests. Additionally, we applied the hypergeometric (HG) test to detect active transcription factors as described in section 4.2. To enable the comparison to *Essaghir et al.* (2010) we follow their approach to regard genes as regulated if they exhibit a fold change of more than 2 or less than 0.5. Both WR and KS tests do not require such a threshold but exploits the ranks of all genes that have been measured. Note that the WR, KS and HG tests are applied to the same set of miRNA and TF target genes as obtained from databases and predictions (Table 12) and used the same procedure for multiple testing correction (*Benjamini* and *Yekutieli,* 2001). In contrast, *Essaghir et al.* (2010) augmented curated databases by their own manual literature searches.

4.3.4 Model of miRNA actions

Differentially expressed genes (DEGs) after miRNA perturbation (e.g., miRNA overexpression) suggest that miRNAs exert their regulatory effects on target as well as non-target genes (see background section). However, it is unclear which of the miRNA target genes serve as key regulators. Previous studies indicate that TFs predominate among miRNA targets (sections 1 and 2).

To analyze the direct and indirect regulatory effects of miRNA, a network model in which TFs propagate miRNA-induced regulatory signals is needed. In this network model, genes directly regulated by miRNA are at the first level and target genes of miRNA-regulated (directly or via miRNA-Kinase association) TFs or TF-regulated TFs are at the second or third level of the regulatory cascades, subject to direct or indirect regulation triggered by miRNA on TFs. This multi-layer regulatory network model explains how targets as well as non-target genes can be regulated by miRNA.

Given these observations, we construct network models of miRNA downstream actions. Here, we aim to connect the transfecting miRNA to TFs via miRNA-TF, TF-TF and kinase-TF interactions derived from databases and computational predictions (Table 12). Thus, TFs are included if they were active according to WR and KS test as described in the last section and are reachable from the transfecting miRNA by a path of known or predicted interactions.

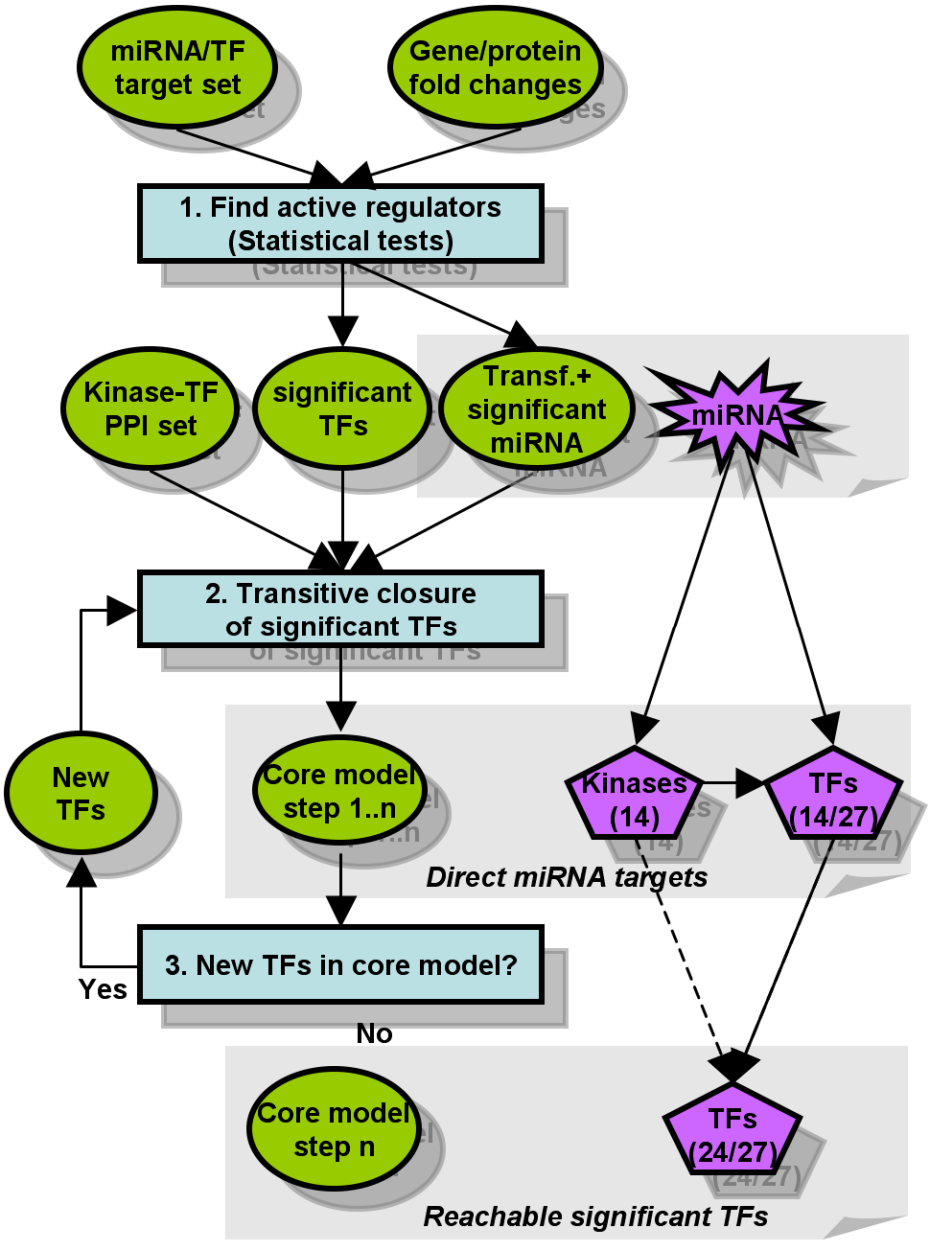

Figure 7: Modelling miRNA actions from expression measurements. Active regulators such as miRNAs and TFs are detected by their effect on the expression of downstream targets, here exemplified by the Wilcoxon test. In step 1 just the direct miRNA targets (kinases and significant TFs) are added to the model. Additional significant TFs are included if they can be connected to the model by interactions from Table 12, i.e. by repeating steps 2 and 3. The model of miR-155 transfection (8hr), for instance, includes 14 kinases and 24 out of 27 TFs detected as active by MIRTFnet. The remaining 3 TFs could not be connected by known interactions. Using these models we consider gene expression changes observed after miRNA transfection as explained if they satisfy two constrains: (1) such a gene must be targeted by an active TF, and (2) such a TF must be connectable to the transfecting miRNA by a path of known or predicted miRNA-TF, TF-TF and kinase-TF interactions.

Note that kinases are included as connectors between miRNAs and TFs in the models although the activity of kinases has not been determined in the examined studies.

Thereby, we aim to give explanations for expression changes observed after miRNA transfection. Based on these models we evaluate to what extent expression changes could potentially be explained based on the current knowledge of causal interactions.

Thus, we propose a cascade of TF activation steps (Figure 7) including the transfecting miRNA, kinases and TFs. Genes that are directly and exclusively affected by miRNAs will most likely be inhibited. This is not necessarily true for indirectly affected TFs or TF target genes (Figure 17).

5. Results

5. Results

This section describes the results for each proposed method together with databases (section 4) accordingly. First we discuss the results for miRSel database (section 5.1). Then we discuss the results for comparative analysis of 12 gene set enrichment (GSE) tests (section 5.2). Finally, we discuss the results for MIRTFnet (section 5.3).

5.1 miRSel: Automated extraction of associations between microRNAs and genes from the biomedical literature

As described above, the miRSel database has been developed to improve the coverage of miRNA-target gene associations including relations. miRSel is updated daily and can be queried using a web-based interface. This section discusses the results including the web interface, filter options and evaluation of miRSel database.

5.1.1 Web interface

miRSel provides a web interface to retrieve information on miRNA-gene pairs stored in the database (Figure 8). The interface allows to AND-combine different options to restrict query result sets.

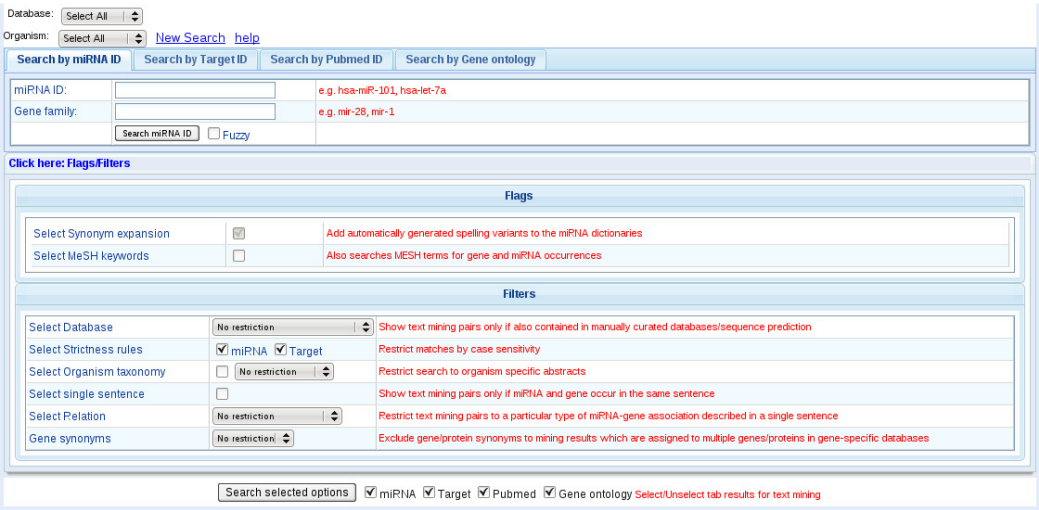

Figure 8: A web based graphical user interface for the miRSel database. miRSel can be queried via different options, including miRNA, target, gene ontology and PubMed keyword queries. If multiple options are selected, the results are AND-combined. Several filters are provided to control recall vs. precision of the mining results.

(1) Genes can be selected based on gene names, gene symbols, protein names or database identifiers.

(2) miRNAs can be selected based on miRNA identifiers and miRNA gene families (*Griffiths-Jones et al.,* 2008).

(3) A PubMed interface enables arbitrary PubMed keyword queries for searching miRSel, miRNA-gene pairs are reported only if found in PubMed abstracts matching the PubMed query.

(4) The gene ontology (GO) option restricts the reported miRNA-gene pairs to genes associated with the selected GO-terms (*Harris et al.,* 2004).

Figure 8 shows the query mask and Figure 9 schematically depicts the query procedure. As a primary query result, an annotated table of miRNA-gene pairs is presented to the user. The table shows whether the pairs are contained in one of the manually curated databases (e.g. TarBase, miR2Disease) or if they have been predicted by miRNA-target prediction algorithms. Besides the table view, miRNA-gene pairs can be analyzed graphically using the Graphviz software (*Gansner et al.,* 2000) (Figure 9). Both representations provide links to the primary database pages (e.g. miRBase, Entrez Gene) of the found entities and to the PubMed abstracts where the entity names have been found.

5.1.2 Filters

The miRSel user interface allows to query occurrences, pairs and associations of miRNAs and genes and to restrict the entries in the database using a number of filter criteria:

(1) The strictness filter enforces a strict string matching of occurrences against the dictionary entries (i.e. occurrences with special characters not in the dictionary or wrong case are removed) (default selected).

(2) The single-sentence filter reports only miRNA and gene pairs co-occurring in single sentences as opposed to pairs co-occurring in abstracts (default abstract).

(3) The relation-type filter restricts matches to a particular type of miRNA-gene association (default none is selected).

(4) The taxonomy filter aims to enforce organism specificity of the matches. Our organism specific taxonomy dictionary contains synonyms and MeSH vocabularies for all examined organisms as provided by the NCBI taxonomy database (*Sayers et al.,* 2009). We define organism specific matches as tri-occurrences of a gene name, a miRNA name and an entry of the taxonomy dictionary (default none is selected).

(5) The gene synonym filter excludes protein or gene synonyms which are assigned to multiple genes or proteins (ambiguous synonyms) in databases (default none is selected).

(6) The database filter shows the text mining pairs only if they also contained in other databases or computational predictions of miRNA gene targets (default none is selected).

The total number of associations with regard to these filters is shown in Figure 10 for human.

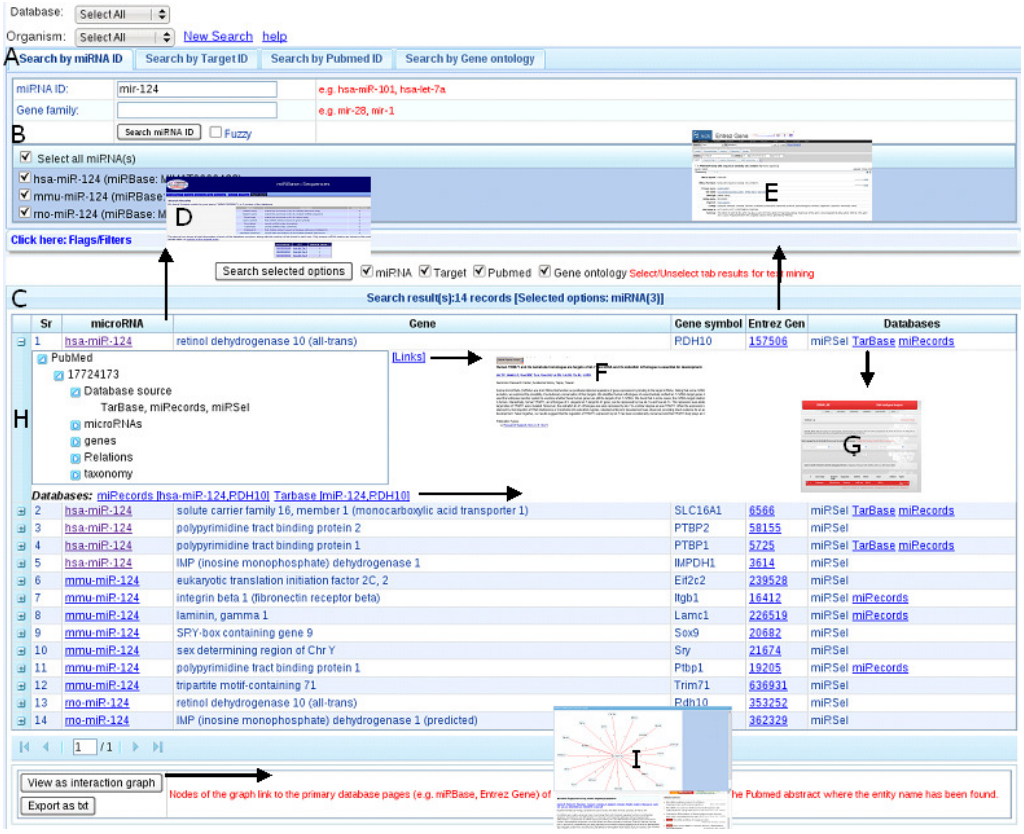

Figure 9: A schematic workflow of a miRSel search by miRNA ID. After entering a complete or partial search key (e.g. a miRNA miR-124) (A) the user can select a subset of the matching miRNAs (B). Then, the corresponding miRNA-target co-occurrences stored in the database are displayed in a tabular format (C). This table enables the navigation to miRNA or gene pages of primary databases (e.g. D=miRBase, E=Entrez Gene, PubMed abstracts that reference particular co-occurrences (F), or to the database sources from which the pair has been integrated (G). Also, details related to each miRNA-target pair e.g. all possible names for a given miRNA or protein in the literature and comparison results of other databases and sequence prediction can be displayed from the table (H). Finally, a miRNA target interaction graph (I) can be displayed that also enables the navigation to miRNA and gene pages (nodes) or PubMed abstracts (edges).

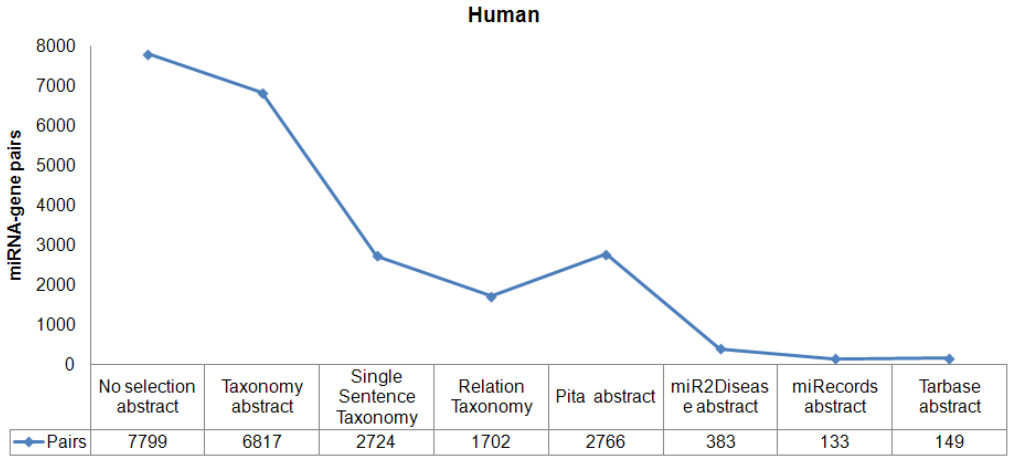

Figure 10: Number of human miRNA and gene/protein pair matches in miRSel. *No selection* all miRNA-gene co-occurrences found in the publication titles and abstracts are displayed. Counts of miRNA-target pairs in the main text refer to this first column. The organism specificity can be increased by the *taxonomy* filter that requires confirmation of the selected organism. The text-mining results can also be restricted to miRNA gene pairs found within *single sentences*. The particlular type of association in miRNA-gene pairs can be restricted *by the relation filter***.** Additional filters report pairs only if they are confirmed by target prediction algorithms (e.g. Pita) or manually curated databases (e.g. miRecords, mir2Disease, TarBase).

5.1.3 Evaluation

miRSel is based on finding occurrences of valid identifiers of genes, proteins and miRNAs in publication abstracts. Here we report on the performance of miRSel with respect to finding valid miRNA, gene and protein occurrences as well as valid miRNA-gene pairs and detailed miRNA-gene pair associations.

We estimate the reliability of the detection of miRNAs in texts. The performance of gene and protein name detection has already been evaluated in the BioCreAtIvE competition (*Fundel et al.,* 2005).

For evaluation we selected PubMed abstracts that matched our regular expression for the detection of miRNAs or contained keywords such as ‘microRNA’, ‘miRNA’ ‘mir’, ‘miR’ and ‘MIR’. Sentences containing a miRNA identifier or related keywords were additionally required to contain protein names from our synonym lists described in the implementation section. 50 PubMed abstracts were chosen randomly containing 89 sentences that met the above requirements. miRSel was compared against various manual analyzes (see below) in terms of recall (i.e. fraction of True Positive (*tp*) and all True occurrences), precision (i.e. fraction of *tp* and all predictions) and F-measure.

The evaluation of miRNA identifier occurrences is shown in Table 14 (a). Using regular expression based-matching, the detection of miRNA identifiers in texts is very reliable.

For the detection of miRNA-gene associations we manually evaluated if a gene and a miRNA have been correctly detected by miRSel and if an association between miRNA and gene is implied. As shown in Table 14 (b), many of the pairs in miRSel represent valid associations. The detection of miRNA-gene associations have been further refined by automatically resolving ambiguities to gene identifiers by using additional tissue and cell-line dictionaries (Table 14 (c)).

Besides the detection of generic miRNA-gene associations, miRSel annotates five different types of associations between miRNAs and genes (physical target, co-expression, repression, induction, and cleavage; see section 4.1 for details). Out of the 2724 single-sentence human miRNA-gene pairs in miRSel 1702 (62%) were classified into one of the five types.

From the test set described above, a subset of the sentences that contain association keywords have also been evaluated manually. If association keywords are present in sentences with miRNA-gene pairs, the precision of association detection increases slightly (Table 14, compare b and d). If a true miRNA-gene association is detected, association keywords describe the type of association correctly in 89% of the cases (Table 14, compare d and e).

Performance evaluation	abstracts	sentences	cases	recall	precision	F-measure
(a) miRNA occurrences	50	89	79	0.96	1.00	0.98
(b) miRNA-gene associations	50	89	181	0.90	0.65	0.76
(c) like b, after disambiguation	50	89	181	0.88	0.78	0.83
(d) like b, with keywords	20	29	103	0.89	0.70	0.78
(e) like b, association types	20	29	103	0.87	0.62	0.73

Table 14: Evaluation of the detection of miRNAs and miRNA-gene associations from PubMed abstracts.

5.1.4 miRSel Query results

Query Examples

The TP53 gene (Entrez Gene: 7157) encodes protein p53, which is one of the most important tumor suppressor proteins. TarBase and miRecords do not report any miRNA targeting this gene. *Xi et al.* (2006) investigate the interaction between TP53 and miRNAs in regulating gene expression using human colon cancer cell lines. They showed that TP53 and miRNAs interact in influencing posttranscriptional and translational events.

We extracted 90 different human miRNAs that co-occur with this target gene from over 80 PubMed abstracts, and some of them (e.g. hsa-let-7, hsa-mir-372, hsa-mir-181b, hsa-mir-200c, hsa-let-7g, and hsa-miR-30) are consistent with microarray-based results discussed by *Xi et al.* (2006). Therefore, miRSel together with gene expression profiling studies can be used to further explain the complex biology and miRNA functions.

Similarly, hsa-miR-21 is the most frequent miRNA in miR2Disease database, with 59 documented associations of this miRNA with diseases. miRSel contains 276 different genes co-occurring with this miRNA extracted from over 123 PubMed abstracts. 150 pairs are retrieved if miRSel results are restricted to the more reliable single-sentence pairs.

To analyze the tissue-specific gene regulation by human miRNAs based on their target genes, *Zhu et al.* (2011) integrate the miRNA-gene associations extracted from miRSel with tissue-specific including brain, heart, kidney liver, lung, skeletal muscle, pancreas, spleen and testis protein interaction networks. They found that miRNAs such as miR-155, miR-21 and miR-16 regulate more commonly expressed proteins/genes in all of the studied tissues whereas miR-15a and miR-1 regulate more tissue-specific proteins of pancreas and heart, suggesting an important role of miRNAs in the tissues. miRSel provides 288, 283, 278, 165 and 287 different target genes of miR-155, miR-21, miR-16, miR-15a and miR-1.

To study disease-related rat miRNAs involve in the pathogenesis of myocardial infarction (MI), *Zhu et al.* (2011) combine the predicted and validated (imported from miRSel) target gene associated with MI-dysregulated miRNAs with cardiac-specific protein-protein interaction network. They showed that miR-1, miR-29b, and miR-98 were key players in MI that causes significant morbidity and mortality. To date, miRSel contains 42, 14 and 2 putative target genes of these confirmed MI-related miRNAs.

miRNAs play an important role in regulating ion channel genes expression at the posttranscriptional level (*Liang et al.,* 2007). *Zhou et al.* (2011) integrate the miRNA-target genes (extracted from miRSel and miRecords) with ion channel and gap junction protein/connexin interaction network to study the pathophysiological processes of MI in rat. They found that genes such as connexin 43 (GJA1) and CACNA1C were more intensive miRNA regulation in comparison to the other protein genes. miRSel provides 5 and 4 different miRNAs targeting these GJA1 and CACNA1C genes.

miRSel Query visualization

miRSel query results can be visualized as a network of the extracted miRNA-gene pairs. Figure 11 shows an example of genes associated with two different terms from gene ontology (GO) and the respective miRNAs targeting those genes. Only small subsets of the text mining pairs have already been annotated in curated resources (bold edges). The two GO terms do not overlap with respect to the set of associated genes/proteins. Still, many miRNAs target genes from both GO terms implicating some functional relationship between the two terms.

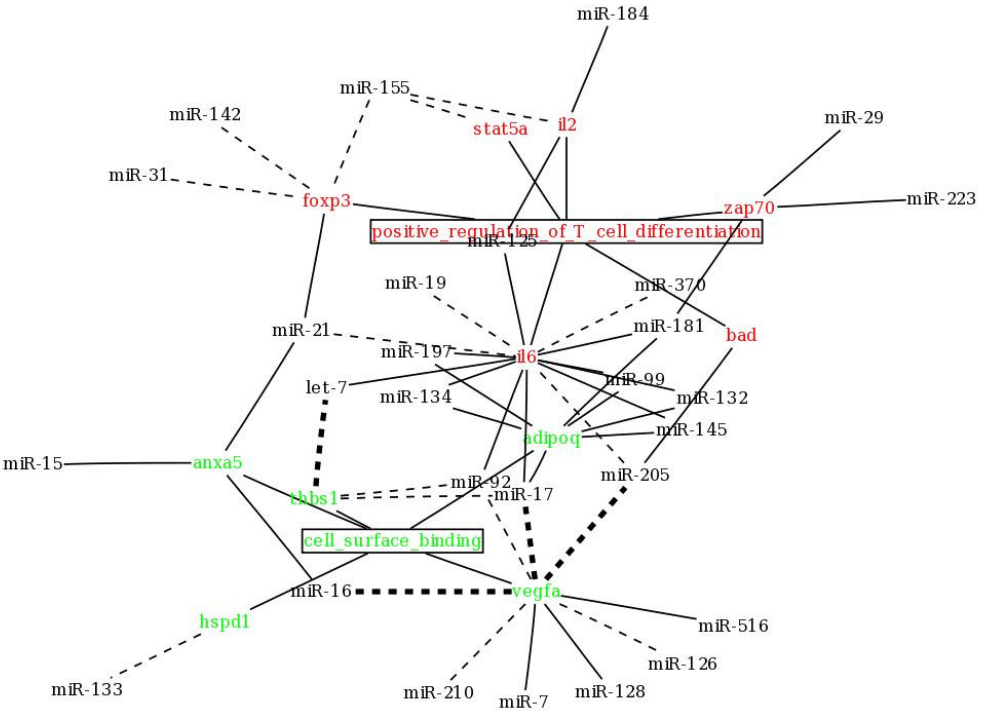

Figure 11: A network representation of miRNA-target pairs in the context of GO. Although the sets of genes from human, mouse and rat associated with the two Gene ontology (GO) terms do not overlap, they substantially overlap in the set of miRNAs. Only a fraction of the depicted relationships are available from databases such as TarBase, miR2Disease and miRecords (bold edges). Some miRNA-target relationships are annotated in miRSel with one of five association types (dashed edges). The networks for individual GO-terms can be retrieved via the web interface and combined subsequently.

5.2 Rigorous assessment of gene set enrichment tests

As of second step, we focus on assessing the ability of state-of-the-art gene set enrichment (GSE) tests to predict regulators that were perturbed (overexpressed or deleted) in given expression experiments (Figure 6). Method performance was evaluated via the area under the receiver-operator characteristic (AUROC) curve. Different gene set definitions and genome-scale real microarray compendia in different organisms have been used to evaluate the performance of each statistical method in a variety of scenarios. This section explains the results related to each step of the assessment phase.

5.2.1 Detection of TF activity without sign annotations

We first evaluated the ability of the applied enrichment tests to predict TFs that have been deleted or over-expressed. At this point, sign annotations are ignored, i.e. we test H_0^{abs}. Perturbations (e.g., TF overexpression or deletion) were only considered effective if the TFs exhibit a fold change of at least two or less than 0.5. Conversely, substantial fold-changes in non-perturbed (secondary) TFs could be due to a direct or indirect effect from the perturbed (primary TFs). Such cases are also excluded from the evaluation. In case of negative examples, we varied the fold-change cutoff to explore its influence on the performance of the enrichment tests (Figure 12). At a higher cutoff, more negative examples are included in the analysis. Although this leads to a slightly decreased performance, the selection of this parameter has only little influence on the ranking of enrichment tests. The ranking of statistical tests such as ANOVA, WR, KS and HG-0.5 test was quite consistent, whereas the ranking of CON and FC methods decreased. The HG-1.5 shows mixed results depending on the cutoff value. The methods such as MED and HG-1.0 test show more decreased performance in comparison to other tests. The resulting AUROC values at a cutoff of 0.5 are shown in Table 15.

In addition, we also combined all individual tests into a consensus. The scores in the individual score matrices (Figure 6) are transformed into ranks and averaged. Although some of the constituent tests hardly perform better than random, the consensus shows consistently good results across the applied scenarios.

5.2.2 Detection of TF activity with sign annotations

This section evaluates if test performance can be improved by exploiting the annotation provided by RegulonDB. This annotation distinguishes whether the TF activates or inhibits a given target gene. H_0^{abs} as applied in the previous section tested only for differential expression. By using H_0^{sign} instead, we additionally test whether the fold changes observed in TF targets are consistent with the given interaction sign annotations. Surprisingly, the tests performance did not improve by utilizing interaction signs (activate vs. inhibit). This might be due to incomplete sign annotations, e.g. with respect to toggle switches (*Morel et al.,* 2000) where a TF can activate or inhibit a target gene depending on the molecular context.

As shown in Table 15, neglecting the sign information improves the performance of enrichment tests without significantly changing their relative ranks. To evaluate why the interaction sign annotations did not improve the results, we compared the signs to the observed fold changes. Activating interactions conformed to our

expectation, i.e. up- (down-) regulation in a TF causes up- (down-) regulation, respectively, in their targets. In case of inhibiting effects we expected opposite fold changes in TFs and their target genes. This observed only rarely in the data (Figure 13) and thus explained the reduction in the performance of the signed tests.

5.2.3 Test performance on E. coli vs. S. cerevisiae

In addition, we also applied the enrichment tests to expression compendia in *S. cerevisiae*. The overall ranking of tests is very consistent between prokaryotic and eukaryotic datasets. The performance for *S. cerevisiae* is somewhat lower than that for *E. coli*. These results are in line with previous studies (*Narendra et al.*, 2010) that discussed the better quality of gene regulatory networks in *E. coli* as well as the simpler gene regulation in prokaryotes as possible reasons.

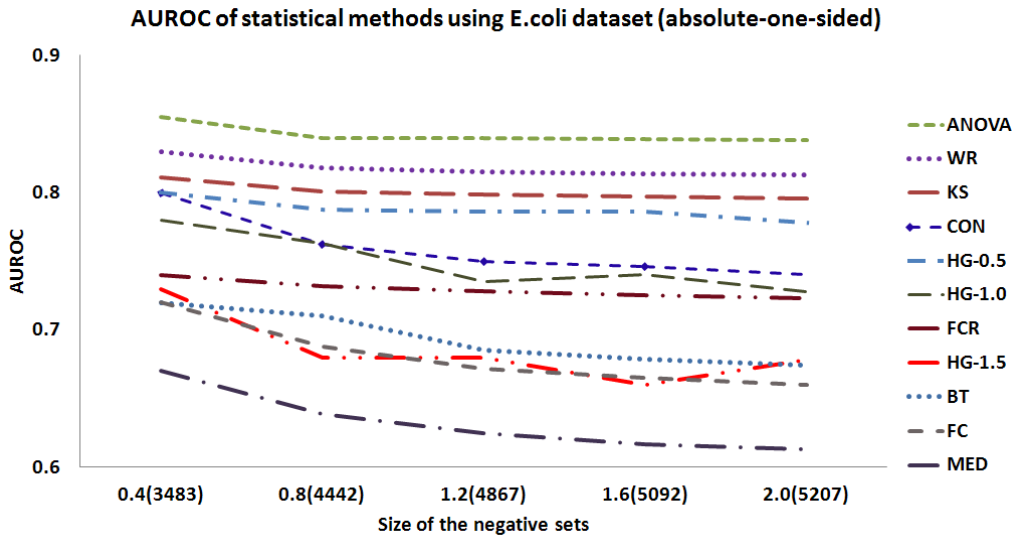

Figure 12: Dependency of AUROC on the set of negatives (*E. coli*). TFs are only considered as negatives in the AUROC analysis if they exhibit fold-changes of less than a pre-defined cutoff. The x-axis shows the sizes of different negative sets (in brackets) compiled based on different fold-change cutoffs ($|log_2$ (*fold change*)|). The size of the negative sets has only little influence on the AUROC (y-axis) or on the relative rank of the different enrichment tests. The order of curves (at 0.4) corresponds to the order of methods in the legend.

5.2.4 Detection of miRNA activity

In addition to TF-target relationships we also evaluated miRNA-target relationships. Here, miRNA transfection experiments in human cell lines were employed. A range of miRNA-target set definitions has been evaluated: databases only (ANOVA achieves an AUROC of 0.63), DBs+PICTAR+TargetScan (high precision prediction tools and ANOVA attains an AUROC of 0.83) and DBs+PITA (high recall prediction tool, ANOVA attains an AUROC of 0.84). Although the quality of computational miRNA target predictions has been discussed controversially (*Ritchie et al.* 2009),

they are required to complement the currently available manual repositories, which appear to be not sufficiently comprehensive for such an analysis. Although this setting deviates considerably from the previously discussed ones, the overall ranking of methods is again very consistent (Table 15 and Table 16). An exception is the hypergeometric test (using a differential expression threshold of 0.5) that showed the second best performance after ANOVA.

Since the expression levels of miRNAs have not been measured, all miRNAs are used to determine the performance of tests. For AUROC analysis, the positive set includes all 43 single miRNA transfection experiments (which are 43 samples based on 18 unique miRNAs) and negative set contains (18-1)*43=731 examples.

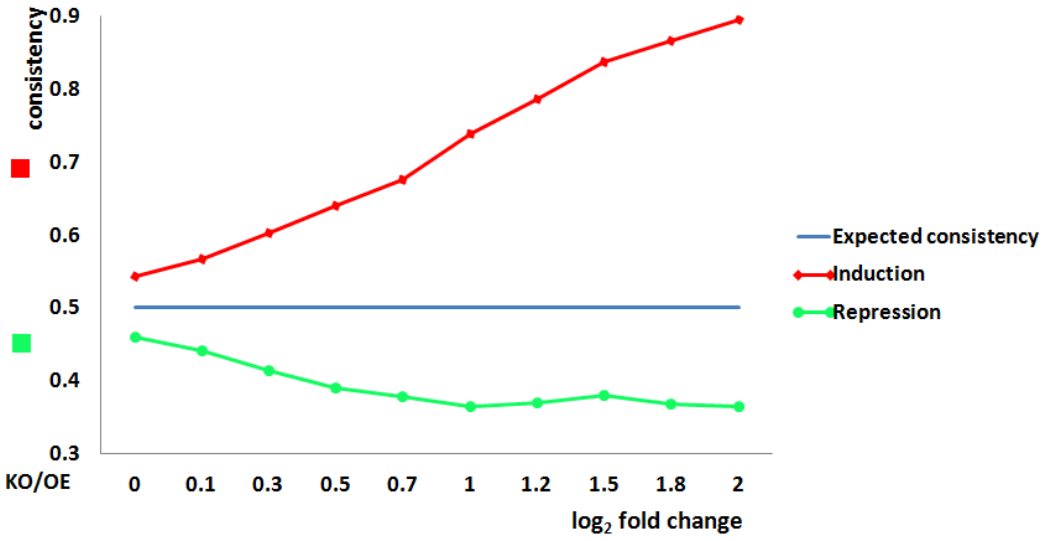

Figure 13: Consistency of sign annotations and fold changes (*E. coli*). In case of activation, fold changes of TFs and their target genes predominantly point into the same direction (red line, consistency>0.5). Consistency increases for TFs and target genes exhibiting higher fold changes. However, upregulated TFs rarely cause down-regulation (or vice versa) of targets in case of inhibiting relationships (green line, consistency<0.5). Considering only deleted or overexpressed TFs (KO/OE: dots, left side) confirms this trend.

Statistical Methods	*E. coli*		*S. cerevisiae*	
	(H_o^{abs})	(H_o^{sig})	Y1 (H_o^{abs})	Y2 (H_o^{abs})
ANOVA	0.86 (± 0.03)	0.66 (± 0.05)	0.71 (± 0.04)	0.71 (± 0.03)
WR	0.83 (± 0.05)	0.64 (± 0.05)	0.71 (± 0.03)	0.68 (± 0.03)
CON	0.80 (± 0.05)	0.60 (± 0.05)	0.73 (± 0.03)	0.67 (± 0.03)
HG-0.5	0.80 (± 0.04)	0.72 (± 0.06)	0.70 (± 0.03)	0.58 (± 0.02)
KS	0.81 (± 0.06)	0.69 (± 0.04)	0.64 (± 0.04)	0.63 (± 0.04)
HG-1.0	0.78 (± 0.04)	0.71 (± 0.04)	0.68 (± 0.04)	0.54 (± 0.06)
HG-1.5	0.73 (± 0.04)	0.67 (± 0.05)	0.72 (± 0.03)	0.56 (± 0.05)
FCR	0.74 (± 0.05)	0.53 (± 0.04)	0.71 (± 0.03)	0.68 (± 0.03)
FC	0.72 (± 0.04)	0.51 (± 0.03)	0.75 (± 0.03)	0.68 (± 0.04)
BT	0.72 (± 0.03)	0.51 (± 0.003)	0.72 (± 0.03)	0.67 (± 0.04)
MED	0.67 (± 0.05)	0.50 (± 0.03)	0.69 (± 0.03)	0.66 (± 0.03)
FCRW	0.56 (± 0.05)	0.50 (± 0.002)	0.56 (± 0.04)	0.71 (± 0.03)

Table 15: AUROC (±standard deviations) for enrichment tests across *E. coli* and *S. cerevisiae* TF expression compendia. Several statistical methods are applied to test the null hypothesis (H_o^{abs}, H_o^{sign}) given TF-target gene set derived from databases (Table 10) following different test settings as described in method section. For AUROC analysis, the positive examples include TFs that show $|log_2 (fold\ change)| >= 1$ (which are 20 for *E. coli*, 54 for *S. cerevisiae* (Y1) and 64 for *S. cerevisiae* (Y2) and the negative examples include TFs that show $|log_2 (fold\ change)| <= 0.4$ (which are 3483 for *E. coli*, 9200 for *S. cerevisiae* (Y1), and 10167 for *S. cerevisiae* (Y2).

Statistical Methods	Human miRNAs		
	(H_o^{abs}) - Databases (P1)	(H_o^{abs}) - P1 + PICTAR+ TargetScan (P2)	(H_o^{abs}) P2+PITA (-20) (P3)
ANOVA	0.63	0.83	0.84 (±0.03)
HG-0.5	0.61	0.83	0.81 (± 0.03)
CON	0.61	0.82	0.80 (± 0.01)
WR	0.60	0.80	0.77 (± 0.03)
KS	0.60	0.78	0.76 (± 0.03)
FCR	0.61	0.82	0.75 (± 0.03)
HG-1.0	0.61	0.77	0.72 (± 0.03)
MED	0.61	0.73	0.68 (± 0.03)
BT	0.62	0.62	0.66 (± 0.007)
HG-1.5	0.59	0.57	0.50 (± 0.04)
FC	0.51	0.51	0.51 (± 0.004)

Table 16: AUROC (±standard deviations) for enrichment tests for human. Statistical methods are applied to test the null hypothesis (H_o^{abs}) given human miRNA-target gene set derived from miRSel (section 4.1) and computational prediction programs (PICTAR, TargetScan and PITA) following different test settings as described in the method section. In total 50 (P1), 260 (P2) and 649 (P3) miRNAs have been evaluated in all miRNA transfection experiments. For AUROC analysis, the positive set includes those miRNAs that are used for transfection in a given

experiment contain more than 20 targets (which are 26 and 43 in case of databases and computational prediction methods).

5.2.5 Randomized testing

To determine how the test performance depends on the quality of the available gene regulatory networks, we progressively randomized the regulator target sets. The results have shown that ANOVA, CON and WR tests were found consistently perform better than the other methods, whereas FCRW and HG-1.5 were found low performing tests. The tests KS, BT, HG-1.0 and HG-0.5 were found perform in the middle level. Interestingly, the HG-0.5 in human miRNA, FC in *S. cerevisiae* (Y1) and FCRW in *S. cerevisiae* (Y2) show a higher ranking relative to other datasets. Since the *E. coli* gold-standard is considered more reliable than *S. cerevisiae* as suggested by *Narendra et al.* (2010), we observed the performance of tests is better on *E. coli* than that of *S. cerevisiae*. However, the ability of the different tests to infer the activity of regulators is surprisingly stable even if, on average, about 50% of the gene regulatory network is randomized (Figure 14).

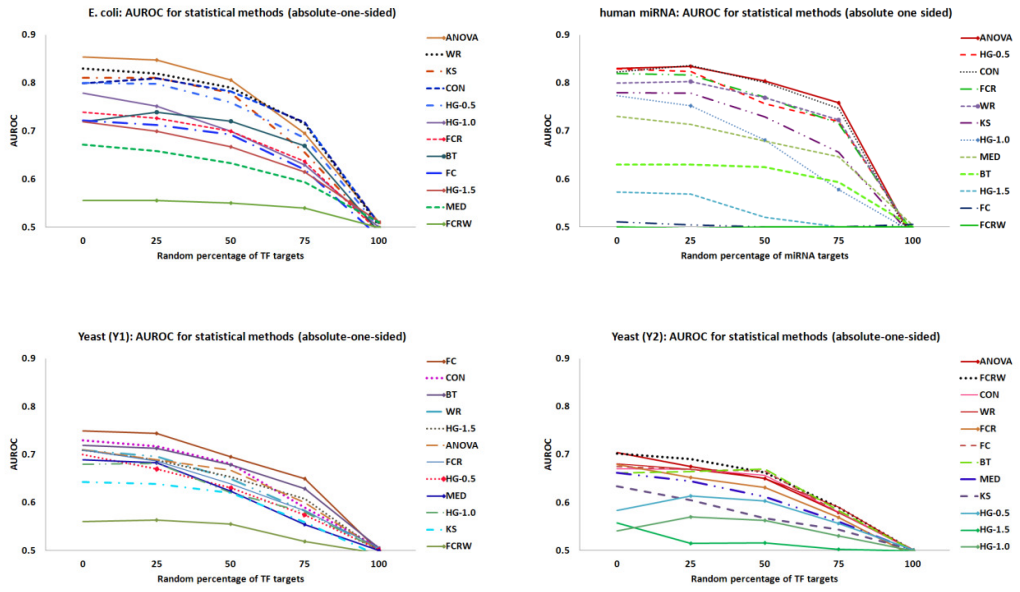

Figure 14: Progressive randomization of gene regulatory networks. The given gene regulatory relationships have been randomized in steps of 25% where 100% represents fully randomized networks. At each step, the average AUROC from 100 (partially) randomized networks is shown. The order of curves (at 0%) corresponds to the order of methods in the legend.

5.2.6 Consensus ranking of methods

We computed average ranks for the examined enrichment tests based on their performance across different partially randomized expression compendia (*E. coli*, *S. cerevisiae* and human) and different scenarios (H_0^{abs} vs. H_0^{sign}). Thereby, we derive the following ordering of methods: ANOVA > CON > WR > HG-0.5 > FCR > KS > HG-1.0 > BT > HG-1.5 > FC > MED > FCRW. ANOVA, CON (consensus) and WR (Wilcoxons rank) perform consistently well across all scenarios. While HG-0.5 (hypergeometric, with threshold 0.5), FCR (fold change rank) and KS (Kolmogorov-Smirnov) also deliver usable results but fail in individual scenarios, the remaining tests (HG-1.0, BT, HG-1.5, FC, MED, FCRW) performed below average across several scenarios (Table 17). We note that predictions by several methods are quite similar so that it is not necessary to implement all methods for a good performance of the consensus (CON). For instance, a consensus of ANOVA, WR and HG-0.5 improves the AUROC in *E. coli* from 0.80 to 0.84.

Statistical Methods	Ranking based on ≥10% permutation results Avg. of 4 test cases (*E. coli* H_o^{abs}+ 2x*S. cerevisiae* H_o^{abs}+*E. coli* H_o^{sign})	Ranking without permutation Avg. of 4 test cases (*E. coli* H_o^{abs}+ 2x*S. cerevisiae* H_o^{abs}+*E. coli* H_o^{sign})	Ranking without permutation Avg. of 3 test cases (human miRNAs: P1, P2, P3)	Ranking based on 25% permutation (human miRNAs)
ANOVA	2	1	1	1
CON	1	3	3	2
WR	3	2	4	3
HG-0.5	8	4	2	4
FCR	6	8	6	5
KS	9	5	5	6
HG-1.0	11	6	7	7
MED	10	11	8	8
BT	4	10	9	9
HG-1.5	7	7	10	10
FC	5	9	11	11
FCRW	12	12	12	12

Table 17: Ranking of statistical methods based on their performance for different test settings including permutation analysis. In total 12 different statistical methods are applied to test the null hypothesis (H_o^{abs}, H_o^{sign}) given miRNA/TF-target gene set derived from databases (miRSel, RegulonDB, TRANSFAC, JASPER and UCSC) and computational prediction programs following different test settings as described in method section. The statistical methods are ranked based on their performance across different test scenarios including *E. coli*, S. *cerevisiae* and human miRNA gene expression compendia.

5.3 MIRTFnet: Analysis of miRNA regulated transcription factors

We applied the selected GSE tests to assemble regulatory network models from miRNA transfection measurements (see section 4.3 for details). This section explains the results including TFs and miRNAs activity analysis and model of miRNA action in the following subsections.

5.3.1 Evaluation of the transfecting miRNAs

We first evaluated how well the miRNAs used for transfection (called primary miRNAs) are detected by MIRTFnet. Only for these miRNAs we can be certain that they should be recognized as active. By using miRNA targets from predictions and databases, transfecting miRNAs were recognized in 42 out of 43 miRNA transfection experiments (Table 18). In most experiments, p-values for transfecting miRNAs were well below the alpha value of 0.05 (see Supplementary Material published in *Naeem et al.* (2011)). This suggests that active regulators can be detected reliably by MIRTFnet. Here we find that the Wilcoxon (WR) test identifies 98% of the transfecting miRNAs. Only 79% and 42% of the transfecting miRNA were identified by the Kolmogorov-Smirnov (KS) and Hypergeometric (HG) tests, respectively (see Table 18). The re-detection of the transfecting miRNA from differential expression of the miRNA targets has also been described in (*Farh et al.,* 2005; *Sood et al.,* 2006; *Arora et al.,* 2008; *Cheng et al.,* 2008; *Tu et al.,* 2009; *Volinia et al.,* 2010; *Ott et al.,* 2011), where similar recall rates have been reported.

Datasets	**Transfect. / overexpressed miRNA**	**Cell line**	**Time point**	**Primary miRNA detected**		
				WR test	**KS test**	**HG test**
Selbach et al., 2008	miR-155	Hela	8	√	√	-
	miR-155	Hela	32	√	-	√
	miR-16	Hela	8	√	√	-
	miR-16	Hela	32	√	√	√
	let-7b	Hela	8	-	-	-
	let-7b	Hela	32	√	√	-
Georges et al., 2008	miR-192	HCT116	24	√	-	√
	miR-192	HCT116	10	√	-	√
	miR-215	HCT116	10	√	-	-
	miR-215	HCT116	24	√	-	√
Baek et al., 2008	miR-124	Hela	24	√	√	-
	miR-1	Hela	24	√	√	-
	miR-181a	Hela	24	√	√	-
He et al., 2005	miR-34a	A549 H-1	24	√	√	-
	miR-34b	A549 H-1	24	√	√	-
	miR-34a	HCT116	24	√	√	√
	miR-34b	HCT116	24	√	√	-
	miR-34a	TOV21G	24	√	√	-
	miR-34b	TOV21G	24	√	√	-
	miR-34a	DLD	24	√	√	√
	miR-34b	DLD	24	√	√	-

	miR-34a	HeLa	24	√	√	√
	miR-34b	HeLa	24	√	√	-
	miR-34a	A549 p53	24	√	√	-
	miR-34b	A549 p53	24	√	√	-
Grimson et al., 2007	miR-7	Hela	12	√	√	-
	miR-7	Hela	24	√	√	√
	miR-9	Hela	12	√	√	√
	miR-9	Hela	24	√	√	-
	miR-122	Hela	12	√	√	-
	miR-122	Hela	24	√	√	-
	miR-128	Hela	12	√	√	-
	miR-128	Hela	24	√	√	√
	miR-132	Hela	12	√	√	√
	miR-132	Hela	24	√	√	-
	miR-133a	Hela	12	√	√	√
	miR-133a	Hela	24	√	√	√
	miR-142-3p	Hela	12	√	-	√
	miR-142-3p	Hela	24	√	-	√
	miR-148b	Hela	12	√	√	-
	miR-148b	Hela	24	√	√	√
	miR-181a	Hela	12	√	√	-
	miR-181a	Hela	24	√	√	-

Table 18: Prediction of active miRNAs based on miRNA targets derived from databases and predictions. The transfecting miRNAs have been detected as active in 42 out of 43 miRNA transfection experiments based on Wilcoxon (WR) test. 15 out of 17 transfecting miRNAs in 34 out of 43 miRNA transfection experiments have been identified as active applying the Kolmogorov-Smirnov (KS) test. 13 out of 17 transfecting miRNAs in 18 out of 43 miRNA transfection experiments have been detected applying the hypergeometric (HG) test. 7 out of 17 transfecting miRNAs activity in 11 out of 43 miRNA transfecting experiments have been detected as active applying the WR, KS and HG tests. 4 out of 17 miRNAs detected as active in 6 out of 43 miRNA transfections applying the WR and HG tests. Overall, the WR test performs better than the KS and HG tests.

5.3.2 Area under the ROC (AUROC) analysis

In addition to recall, we also analyze the specificity of detection. Therefore, we evaluated how many other miRNAs (called secondary miRNAs) are statistically shown to be active in response to miRNA (called primary miRNA) transfection experiments. We assessed the performance of each method by the area under the receiver operating characteristic (AUROC) curve, a measure combining specificity and recall (section 2.4). Here, we considered primary miRNAs as positive examples and secondary miRNAs as negative examples. This assessment might underestimate the true performance, for instance if miRNA transfection causes activity changes in secondary miRNAs that we count as false positives. Overall, Wilcoxon test was found better than KS and HG tests.

Using the databases (i.e., miRSel, TarBase and miRecords) miRNA-gene target associations, the WR test achieves an AUROC of 0.73, KS test achieves an AUROC of 0.68 and the HG test attains an AUROC of 0.62 (first 25 experiments). We improved the target gene set by adding the PICTAR and TargetScan miRNA-gene target predictions. In this case WR, KS and HG test achieve AUROC's of 0.76, 0.75 and 0.71, respectively. We also tested the performance by complementing the testing set with PITA predictions. In case of PITA predictions compiled at stringent threshold improves the WR test AUROC to 0.81, KS test AUROC to 0.79 and HG test AUROC to 0.62. In case of a less stringent PITA prediction threshold the performance of AUROC decreases. Overall, the WR test achieves higher AUROC than KS and HG test (see Table 19 and 20 for more details).

miRNA-gene target source	**Tested secondary miRNAs**	**AUROC**		
		WR test	**KS test**	**HG test**
Databases (DB)	314	0.73	0.68	0.62
DB+ PICTAR+TargetScan(P1)	404	0.76	0.75	0.71
P1+PITA(threshold of -20)	775	0.81	0.79	0.62
P1+PITA(threshold of -11)	780	0.67	0.65	0.57
P1+PITA(threshold of -6)	780	0.67	0.66	0.57

Table 19: AUROC performance of Wilcoxon, Kolmogorov-Smirnov and hypergeometric test using 25 miRNA transfections. The performance of each method has been accessed by means of area under the ROC curve. The Wilcoxon (WR), Kolmogorov-Smirnov (KS) and Hypergeometric (HG) test has been applied to determine the significance of each miRNA (i.e., primary/transfecting and secondary miRNA) using the databases and prediction tools miRNA-gene target associations in all 25 miRNA transfections. For AUROC the primary or the transfecting miRNA in each transfecting experiment is considered as positive example (which are 25 in total) and the rest of the miRNAs are considered as negative examples. Using the databases miRNA-gene pairs the WR, KS and HG test achieve the AUROC of 0.73, 0.68 and 0.62. Combining the databases and prediction tools (e.g., PICTAR and TargetScan) miRNA-gene target associations improves the AUROC scores. Overall, the WR test achieves the best AUROC in comparison to KS and HG test.

We also use the data of Georges *et al.,* (2008) that measured the gene expression data in Hela cells at 12h and 24h after miRNA overexpression of miR-7, miR-9, miR-122, miR-128, miR-132, miR-133, miR-142 and miR-181a (18 transfections in total). We applied the WR, KS and HG test to measure the significance of all miRNAs (780 in total) in these experiments. We then measured the AUROC for combined (25+18 transfections) 43 miRNA transfection experiments. In this case combining the databases and prediction tools (e.g., PICTAR, TargetScan and PITA (using the most stringent threshold of -20)) miRNA-gene targets improves the AUROC of WR test to 0.88, KS to 0.86 and HG test to 0.65 in comparison to first 25 miRNA transfections results (Table 20). The AUROC is further increased to 0.91 if only those primary miRNAs were considered that are found statistically active by both the WR and KS test.

miRNA-gene target source	**Tested secondary miRNAs**	**AUROC**		
		WR test	**KS test**	**HG test**
Databases (DB)	314	0.63	0.59	0.58
DB+ PICTAR+TargetScan(P1)	404	0.88	0.87	0.70
P1+PITA(threshold of -20)	775	0.88	0.86	0.65
P1+PITA(threshold of -11)	780	0.71	0.69	0.58
P1+PITA(threshold of -6)	780	0.68	0.67	0.60

Table 20: AUROC performance of Wilcoxon, Kolmogorov-Smirnov and hypergeometric test using 43 miRNA transfections. The performance of each method has been accessed by means of area under the ROC curve. The Wilcoxon (WR), Kolmogorov-Smirnov (KS) and hypergeometric (HG) test has been applied to determine the significance of each miRNA (i.e., primary/transfecting and secondary miRNA) using the databases and prediction tools miRNA-gene target associations in all 43 miRNA transfections. For AUROC the primary or the transfecting miRNA in each transfecting experiment is considered as positive example (which are 43 in total) and the rest of the miRNAs are considered as negative examples. Using the databases miRNA-gene pairs the WR, KS and HG test achieve the AUROC of 0.63, 0.59 and 0.58. Combining the databases and prediction tools (e.g., PICTAR and TargetScan) miRNA-gene target associations improves the AUROC scores. In case of PITA predictions (filtered at a threshold of -11 and -6), the AUROC improves in comparison to databases but decreases in comparison to combine including PICTAR, TargetScan and databases miRNA-target gene pairs. Combining the PITA predictions (compile at stringent threshold) improve the AUROC of WR test more than the other methods including KS and HG test.

5.3.3 Detection of active transcription factors

Wilcoxon test

Active TFs (Table 21) were detected: 1) if they exhibit a fold change of at least two in a given miRNA transfection experiment or 2) via the differential expression of their direct downstream targets (obtained from JASPAR, UCSC and TRANSFAC) using statistical tests as described in method section. In the Selbach *et al.* (2008) miRNA-transfection datasets for instance, we identified more than 20 active TFs (e.g., ELK4, CREB1, E2F1 and MAFB) and 10 TF based on fold change (e.g., TP53, ZEB1, ZNF423, FOSB and FOX03) applying the Wilcoxon's test. We have also found five TFs (FOS, CREB1, ID1, ZNF423 and MYB) that are both statistically significant and differentially expressed in the miR-155 (32hr), let-7b (32hr), miR-34a and miR-34b (24hr) miRNA transfection experiments. Thus, a total of 88 TFs have been detected applying the Wilcoxon's test.

Kolmogorov-Smirnov test

In addition to WR test we also applied the Kolmogorov-Smirnov (KS) test and the hypergeometric test. The KS test identified in total 73 active TFs (Table 21). 69 out of 71 KS active TFs have also been identified by the Wilcoxon's test. In most of the cases, the TFs identified by the KS test are a subset of those identified by WR test.

Hypergeometric test

As proposed by *Sohler et al.* (2005), *Essaghir et al.* (2010) and *Liu et al.* (2010), we also applied the hypergeometric (HG) test (equivalent to Fisher's test). The HG test identified only very few active TFs (Table 21). The HG p-values were consistently higher (less significant) than the p-values derived from the WR and KS test.

Dataset	Transfec. /over expressed miRNA	Cell line	Time point	Total	Fold change (FC)	FC+ sig.	Statistically significant TFs				
							WR	KS test	Shared KS+WR	HG test	Shared HG+WR
Selbach et al., 2008	miR-155	HeLa	8	27	7	0	20	6	6	0	0
	miR-155	Hela	32	30	11	3	16	13	13	6	3
	miR-16	Hela	8	20	9	0	11	2	2	1	0
	miR-16	Hela	32	34	10	0	24	17	17	0	0
	let-7b	Hela	8	27	5	1	21	9	9	5	3
	let-7b	Hela	32	25	8	1	16	9	9	5	4
Georges et al., 2008	miR-192	HCT116	24	19	2	0	17	12	12	1	0
	miR-192	HCT116	10	25	0	0	25	12	12	0	0
	miR-215	HCT116	10	20	0	0	20	10	10	0	0
	miR-215	HCT116	24	20	1	0	19	10	10	1	0
Baek et al., 2008	miR-124	Hela	24	33	4	0	29	31	28	0	0
	miR-1	Hela	24	40	2	0	38	34	34	0	0
	miR-181a	Hela	24	22	5	0	17	24	17	0	0
He et al.,	miR-34a	A549 H-1	24	64	0	0	64	62	61	0	0

2005	miR-34b	A549 H-1	24	57	0	0	57	56	52	0	0
	miR-34a	HCT116	24	65	4	1	60	58	55	0	0
	miR-34b	HCT116	24	66	3	0	63	59	58	0	0
	miR-34a	TOV21G H1	24	71	0	0	71	59	59	0	0
	miR-34b	TOV21G H1	24	64	0	0	64	62	62	0	0
	miR-34a	DLD	24	65	1	1	63	60	59	0	0
	miR-34b	DLD	24	59	3	1	55	53	51	0	0
	miR-34a	HeLa	24	64	0	0	64	62	60	0	0
	miR-34b	HeLa	24	63	0	0	63	57	56	0	0
	miR-34a	A549 p53	24	61	0	0	61	60	58	0	0
	miR-34b	A549 p53	24	59	0	0	59	60	56	0	0
Grimson et al., 2007	miR-7	Hela	12	9	0	0	9	4	4	0	0
	miR-7	Hela	24	0	0	0	0	0	0	0	0
	miR-9	Hela	12	31	2	0	29	15	15	0	0
	miR-9	Hela	24	9	0	0	9	0	0	0	0
	miR-122	Hela	12	5	3	0	2	1	1	2	1
	miR-122	Hela	24	29	4	1	24	18	17	2	2
	miR-128	Hela	12	4	0	0	4	6	4	0	0
	miR-128	Hela	24	12	1	0	11	8	7	0	0
	miR-132	Hela	12	0	0	0	0	0	0	0	0
	miR-132	Hela	24	0	0	0	0	0	0	0	0
	miR-133a	Hela	12	27	5	1	21	14	13	4	2
	miR-133a	Hela	24	31	12	2	17	16	16	9	6
	miR-142-3p	Hela	12	3	0	0	3	2	2	0	0
	miR-142-3p	Hela	24	2	0	0	2	0	0	0	0
	miR-148b	Hela	12	8	1	0	7	6	5	0	0
	miR-148b	Hela	24	19	1	0	18	11	11	1	0
	miR-181a	Hela	12	0	0	0	0	0	0	0	0
	miR-181a	Hela	24	9	0	0	9	5	5	0	0
Total				120	52	7	88	73	71	21	12

Table 21: Prediction of active TFs based on the expression of their target genes. Overall, 120 TFs have been identified by MIRTFnet (applying the Wilcoxon (WR) test) with the used datasets. 73 and 21 TFs have been detected applying the Kolmogorov-Smirnov (KS) and hypergeometric (HG) test in 37 and 11 out of 43 miRNA transfection profiles. The WR, KS and HG test does not find any significant TF in 4, 6 and 32 out of 43 miRNA transfection profiles. In 15 out of 37 miRNA transfection profiles the detected significant TFs overlap between the WR and KS test is 100%. In all of these cases WR test has detected more active TFs than KS test.

5.3.4 Rank distribution of active TFs

We detect regulators such as miRNAs and TFs as active via the expression of their putative target genes. If the mean expression of the target genes is significantly different to the mean expression of the remaining genes we identify the corresponding regulator as active according to the applied tests. As an example for an active transcription factor we depict ELK4 in the transfection experiment of has-miR-155 at 32h (Figure 15). JASPAR predicts 1,826 putative targets of ELK4. Compared to the 16,101 remaining genes, ELK4 targets exhibit larger fold changes and thus higher ranks than expected by chance.

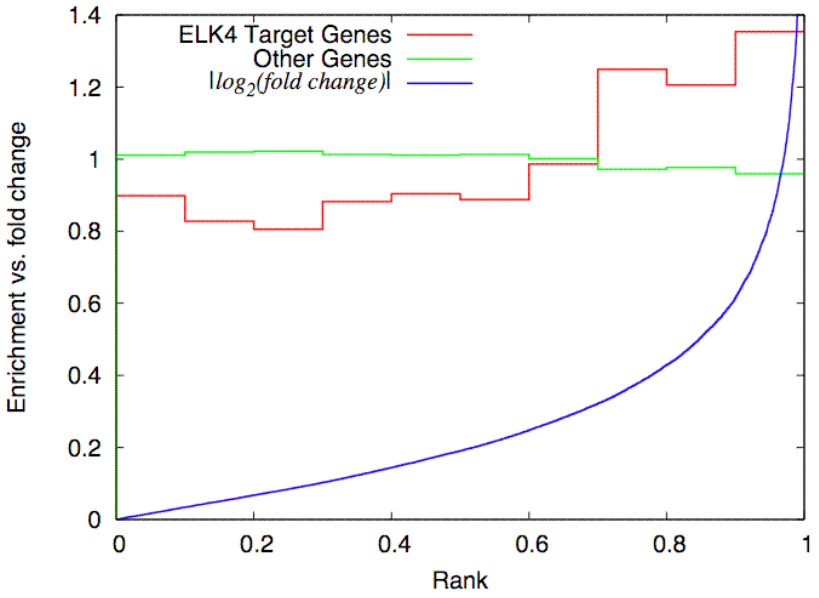

Figure 15: Rank distributions of putative ELK4 targets. The ranks are derived from the list of target genes sorted according to their fold changes (blue) in the miR-155 transfection experiment at 32h. The distributions are normalized to show the relative overrepresentation of ELK4 targets (red) vs. the remaining genes (green) in histogram bins of $|log_2(fold\ change)| > 0.4$ (corresponding to fold changes > 1.3, or < 0.75) ELK4 targets are enriched by about 50% compared to $|log_2(fold\ change)| < 0.4$. ELK4 is thus identified as an active regulator with a p-value of 2.87E-11 according to the Benjamini-Hochberg corrected Wilcoxon's test.

Interestingly, the enrichment of differentially expressed ELK4 target genes is already noticeable at moderate fold-changes (> 1.3, or < 0.75). Note that Figure 15 serves only as visualization whereas active TFs are only determined by the statistical tests described in method section. To summarize the analysis of all TFs across all miRNA transfection profiles, we show the p-value distributions as derived from WR and KS tests in Figure 16.

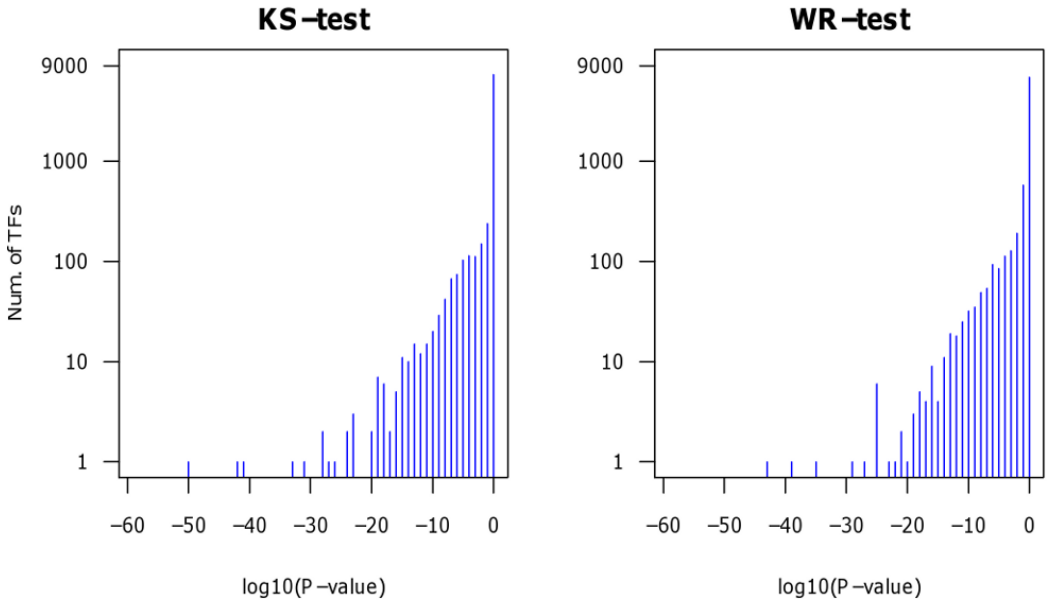

Figure 16: P-value distribution of TFs in miRNA transfection experiments. To detect active TFs, the statistical tests have been applied to 196 TFs across 43 transfection profiles (Y-axis: 196*43=8428). If a test assigns a p-value of less than 0.05 after multiple testing correction a given TF is identified as active for the given measurement. Depicted are the p-value distributions of the WR and KS tests, i.e. the number of TF that fall in a given p-value range according to the respective test.

5.3.5 Randomized Testing

We also evaluate whether TF are detected as active by chance. Here, we randomize the association of gene names and expression levels in each experiment and apply the WR and KS test as described in the methods section. We shuffle gene labels and expression levels randomly 100 times. The test did not find a single regulator as active (neither miRNA nor TF) at a corrected p-value of less than 0.05 after applying multiple testing correction using the method of Benjamini and Hochberg (*Benjamini and Yekutieli,* 2001). This was true regardless of which sub-selections of miRNA-target or TF-target data sources were used. For instance, in case of miRNA-targets we tested curated databases, databases plus low recall prediction tools (e.g., PICTAR and TargetScan) and databases plus high recall prediction tools (e.g., PICTAR, TargetScan and PITA).

5.3.6 Global expression pattern explained by active TFs

Based on protein-protein interactions, miRNA-targets and TF-targets, we constructed transfection experiment specific models that connect the transfecting miRNA via causal relationships to the TFs that were detected as active using the proposed statistical tests.

In each transfection profile, 196 TFs were tested. On average, 23 TFs were detected as active by both the WR and KS tests. Here, 21 out of 23 TFs could be connected to the transfecting miRNA based on causal relationships (compare Supplementary File S4 published in *Naeem et al.* (2011)).

We used miRNA-targets and TF-targets from curated databases as well as computational predictions (Figure 7).

We analyzed to what extent regulators (e.g., miRNAs and TFs) and their known/predicted target genes can explain the overall expression changes observed on the microarrays. Table 22 shows the gene regulation that can be explained by MIRTFnet via miRNA-TF relations. The identified TFs and their target genes thus provide a potential explanation for the majority (on average 67%, Supplementary File S1 published in *Naeem et al.* (2011) shows the exact numbers for each measurement) of the observed differential expression in the examined miRNA transfection experiments (e.g., Figure 17).

Dataset	miRNA, Time point	Cell Line	Diff. exp. genes (down/up)	miRNA down reg. targets	miRNA-regulated TF: targets (Sig. TFs)	TF-regulated TF: targets (Sign. TFs)	Regulated genes (down/up)	Percentage combined (down /up) regulated
Selbach et al., 2008	miR-155(8hr)	Hela	268(189/78)	57	97(6)	52(5)	148(109/39)	55(57/50)
	miR-155(32hr)	Hela	534(385/148)	201	283(12)	110(7)	382(303/79)	71(78/53)
	miR-16(8hr)	Hela	360(250/110)	130	108(5)	57(5)	211(168/43)	58(67/39)
	miR-16(32hr)	Hela	516(269/247)	174	277(15)	215(10)	380(235/145)	73(87/58)
	let-7b(8hr)	Hela	236(168/67)	56	102(6)	92(7)	150(112/38)	63(66/56)
	let-7b(32hr)	Hela	259(138/120)	63	117(9)	66(3)	154(99/55)	59(71/45)
Georges et al., 2008	miR-192(24hr)	HCT116	71(40/31)	33	27(6)	36(6)	55(38/17)	77(95/54)
	miR-192(10hr)	HCT116	9(5/4)	5	3(3)	3(2)	6(5/1)	66(100/25)
	miR-215(10hr)	HCT116	9(5/4)	5	3(2)	6(3)	7(5/2)	77(100/50)
	miR-215(24hr)	HCT116	105(49/56)	45	30(5)	38(4)	70(47/23)	66(95/41)
Baek et al., 2008	miR-124(24hr)	Hela	324(171/153)	65	164(18)	120(10)	203(120/83)	62(70/54)
	miR-1(24hr)	Hela	143(26/117)	8	62(15)	68(13)	82(17/65)	57(65/55)
	miR-181a(24hr)	Hela	399(205/194)	44	121(9)	158(8)	200(107/93)	50(52/47)
He et al., 2005	miR-34a(24hr)	A549	48(19/29)	12	24(28)	25(24)	33(15/18)	68(78/62)
	miR-34b(24hr)	A549	27(8/19)	3	8(15)	10(22)	12(4/8)	44(50/42)
	miR-34a(24hr)	HCT116	150(98/52)	75	93(27)	82(24)	127(92/35)	84(93/67)
	miR-34b(24hr)	HCT116	131(71/60)	30	77(24)	83(34)	96(56/40)	73(78/66)
	miR-34a(24hr)	TOV21G	11(7/4)	5	5(11)	5(10)	9(6/3)	81(85/75)
	miR-34b(24hr)	TOV21G	7(6/1)	3	1(1)	1(0)	4(4/0)	57(66/0)
	miR-34a(24hr)	DLD	126(97/29)	72	84(30)	69(25)	108(88/20)	85(90/68)
	miR-34b(24hr)	DLD	202(123/79)	46	111(19)	120(30)	145(94/51)	71(76/64)
	miR-34a(24hr)	HeLa	57(39/18)	29	23(29)	24(24)	40(31/9)	70(79/50)
	miR-34b(24hr)	HeLa	67(39/28)	19	32(22)	37(32)	42(30/12)	62(76/42)
	miR-34a(24hr)	A549	59(27/32)	19	37(30)	35(23)	45(24/21)	76(88/65)
	miR-34b(24hr)	A549	40(18/22)	4	15(19)	23(27)	26(10/16)	65(55/72)
Grimson et a., 2007	miR-7(12hr)	Hela	68(39/29)	27	5(1)	14(1)	37(32/5)	54(82/17)
	miR-7(24hr)	Hela	34(28/6)	19	0(0)	0(0)	19(19/0)	55(67/0)
	miR-9(12hr)	Hela	110(45/65)	28	42(9)	23(6)	63(35/28)	57(77/43)
	miR-9(24hr)	Hela	14(13/1)	6	0(0)	0(0)	6(6/0)	42(46/0)
	miR-122(12hr)	Hela	337(181/156)	56	0(0)	0(0)	56(54/2)	16(29/1)
	miR-122(24hr)	Hela	654(360/294)	165	295(8)	174(7)	404(255/149)	61(70/50)
	miR-128(12hr)	Hela	51(44/7)	19	4(2)	2(1)	23(22/1)	45(50/14)
	miR-128(24hr)	Hela	88(56/32)	44	27(5)	9(1)	52(47/5)	59(83/15)
	miR-132(12hr)	Hela	104(78/26)	51	0(0)	0(0)	51(51/0)	49(65/0)
	miR-132(24hr)	Hela	52(28/24)	12	0(0)	0(0)	12(12/0)	23(42/0)
	miR-133a(12hr)	Hela	77(31/46)	29	29(9)	29(6)	54(28/26)	70(90/56)
	miR-133a(24hr)	Hela	267(107/160)	82	120(10)	128(12)	186(94/92)	69(87/57)

Table 22: Percentage of differentially expressed genes explained by MIRTFnet (37 miRNA transfection experiments). In case of the miR-155 transfection at 8hr, 148/268=55% of the differentially expressed genes can be explained by the union of the 67 direct target genes of miR-155, the 97 targets of the 6 TFs directly targeted by miR-155 and the 52 targets of the 5 TFs indirectly affected by miR-155. Indirectly affected TFs are connected to miR-155 by miRNA-(kinase-)(TF-…-)TF chains (Figure 17). The 6+5 TFs were identified as active TFs by MIRTFnet (applying the Wilcoxon (WR) and Kolmogorov-Smirnov (KS) tests). We regard genes as differentially expressed if they exhibit a fold change of at least 2 or less than 0.5.

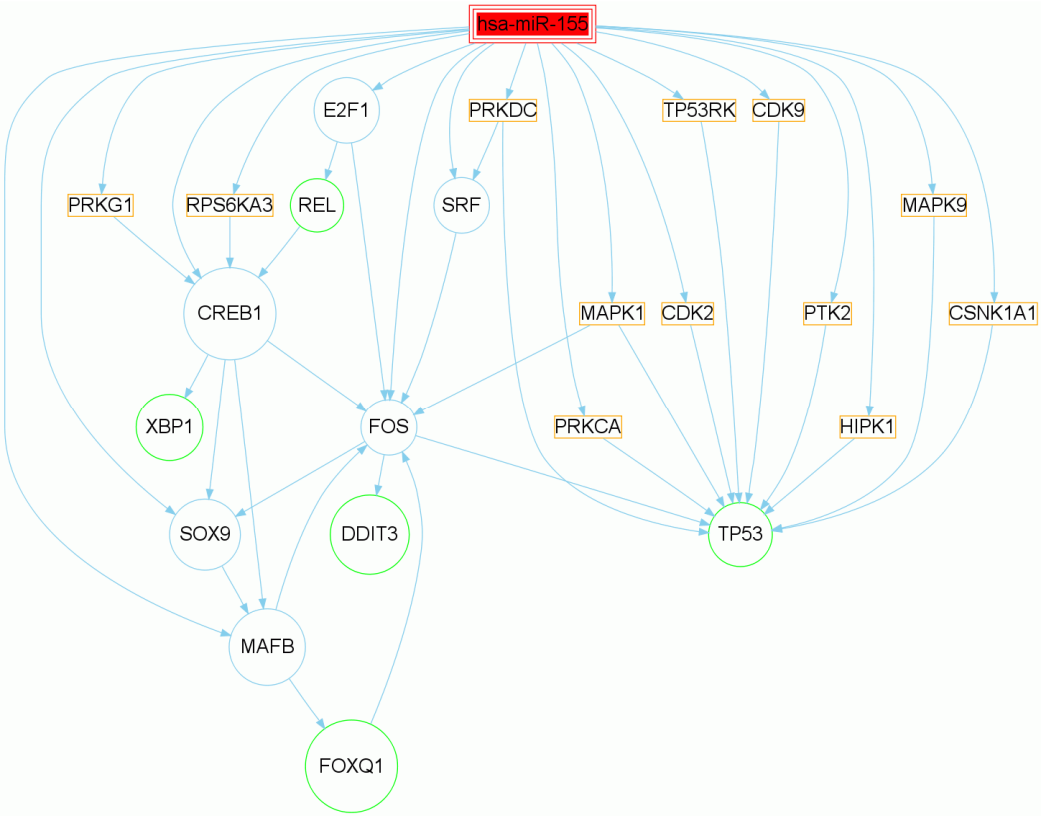

Figure 17: Model of the regulatory network induced by miR-155 (8-hr) transfection. The red box shows the transfected human (hsa) miR-155; orange box nodes indicate the miRNA-regulated kinases; blue circle nodes represent TFs that are regulated either directly by the transfected miR-155 or by miRNA-regulated kinases; the green circle nodes represent TFs that are regulated by miRNA-targeted TFs, subject to indirect regulation of miR-155 on TFs. All TFs were identified as active TFs by MIRTFnet (applying the Wilcoxon (WR) and Kolmogorov-Smirnov (KS) tests) (section 4.3). The active TFs were connected to the transfected miRNA by interactions extracted from databases or computational predictions (Table 12). Additionally, kinases were connected via miRNA-kinase-TF causal relationships i.e. they usually do not receive direct support from the expression profile. 55% of the differential expression pattern can be explained by the extracted miR-155-TF regulatory network model. Here, genes are regarded as differentially expression if they exhibit a fold change of at least 2 or less than 0.5 (Table 22).

5.3.7 miRNA-target TF associations in databases and prediction programs

Whether a connection between the transfecting miRNA and active TFs can be established depends on the current databases and sequence based prediction programs of miRNA target genes (Figures 7 and 17).

Based on these associations we aim to construct models of miRNA actions (see methods section 4.3). However, these would be very small if only databases as well as PICTAR and TargetScan are used for model construction (Figure 18). Here, only four TFs on average would be connected to the transfecting miRNA. To improve this recall, PITA miRNA-gene associations are used as well. The combined miRNA-gene associations suggest connections to about 16 active TFs for all of the examined miRNA transfection experiments.

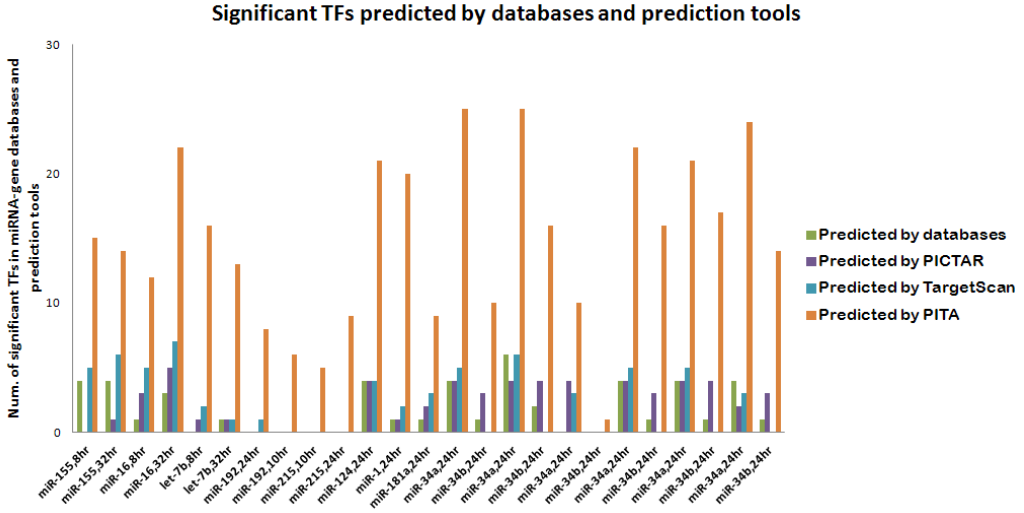

Figure 18: Significant TFs predicted by databases and/or sequence prediction programs of miRNA-target genes. TFs are detected as active by analyzing the expression levels of their downstream targets (Wilcoxon's test). Active TFs can be predicted from databases in 18 out of 25 miRNA transfection experiments (on average 3 TFs per miRNA transfection experiment). PICTAR and TargetScan prediction programs can predict on average 3 and 4 active TFs in 19 and 16 out of 25 miRNA transfection experiments, respectively. PITA can predict TFs in all 25 miRNA transfection experiments (on average 15 active TFs per transfection profiles). To improve recall, the miRNA-gene associations of databases and prediction programs are combined (on average 16 active TFs per transfection).

5.3.8 Detected TFs and their reported roles in cancer – literature mining

miRNAs play potential roles in the pathogenesis of different diseases including cancer (*Lu et al.,* 2005; *Li et al.,* 2010). Some miRNAs may be directly involved in cancer development by controlling cell differentiation and apoptosis or by targeting cancer oncogenes and/or tumor suppressors (*Jovanovic* and *Hengartner,* 2006; *Sassen et al.,* 2008; *Subramanian et al.,* 2010). All of the transfection experiments analyzed in this thesis have been described in the literature as cancer relevant.

The miR-192, miR-215 and miR-34 experiments were analyzed because these miRNAs are reportedly regulated by p53 and are thus potentially involved in cancer related processes (*He et al.,* 2007; *Georges et al.,* 2008). We also analyzed the miR-155, let-7 and miR-16 transfection experiments (*Selbach et al.,* 2008) for which interactions with p53 have been reported as well (*Suzuki et al.,* 2009; *Gironella et al.,* 2007). We thus expect to predominantly identify cancer related TFs which we will evaluate below as a proof of concept of MIRTFnet. The cancer specific involvement of many of the TFs MIRTFnet determined as active is indeed discussed in the literature.

In case of the miR-155 transfection, we detected oncogenic TFs (e.g., SPI1, MYCN, MAFB, FOS and REL) and the tumor suppressor TP53, which may suggest a tumor-induction effect. Previous reports have experimentally confirmed that SPI1 (Pu.1) reduces the transcriptional activity of the p53 tumor suppressor family (*Tschan et al.,* 2008). The deregulation of MYCN leads to cell cycle exit and terminal differentiation (*Bell et al.,* 2007; *Otto et al.,* 2009).

In the miR-16 transfection, we found target genes of oncogenic TFs (e.g., MAFB, MYB) including Cyclin D1/CCND1 and CDK6 to be differentially expressed as well. Both CCND1 and CDK6 are experimentally validated targets of miR-16 that induce cell cycle arrest (*Cimmino et al.,* 2005; *Liu et al.,* 2008).

In case of let-7b transfection, tumor suppressor TP53 and oncogenes such as E2F1, FOS and FOSB have been found active, which might hint to tumor-suppressing effects of let-7b. Recently, the let-7 family miRNAs were found to inhibit E2F family oncogenes (*Tu et al.,* 2009). The TFs (e.g., TP53, FOS and FOSB) are predicted targets of let-7b (*Kertesz et al.,* 2007). The let-7 family is described to be in many human cancers (*Boyerinas et al.,* 2010; *Barh et al.,* 2010).

Recent studies confirm that TP53 regulates apoptosis by targeting miRNAs, such as miR-34, miR-192 and miR-215 (*Subramanian et al.,* 2010; *Hermeking et al.,* 2010; *Chang et al.,* 2007; *Braun at al.,* 2008). The miR-34, miR-192 and miR-215 halt cell cycle progression by co-ordinately targeting transcripts that play critical roles in mediating cell cycle control (*Corney et al.,* 2007; *Welch et al.,* 2007; *Georges et al.,* 2008). Our results showed that miR-34 alters the activity of the MYCN, MYB, MAFB and E2FI oncogenes, all being involved in apoptosis and cell proliferation (*Wei et al.,* 2008; *Grönroos et al.,* 2004). The predicted target of miR-34, YY1 has been shown to down-regulate TP53 (*Grönroos et al.,* 2004). miR-192 and miR-215 were found to inhibit HOXA10 and several oncogenes (e.g., MYCN and MAFB). Furthermore, miR-192 and miR-215 were found to down-regulate CDC7, which might provide an additional explanation for the involvement on miRNAs in the p53 pathway to mediate cell cycle and apoptosis (*Kim et al.,* 2002).

Also in case of miR-9 transfection, tumor suppressor p53 and oncogene transcription factors such as Runx1, E2F1, MYCN and MYB have been found active. Both MYC and MYCN oncoproteins act on the mir-9-3 locus and cause activation of miR-9 expression in tumor cells (*Ma et al.,* 2010). Runx1 is an experimentally validated target of miR-9 and has been reported to act as tumor suppressor, dominant oncogene or mediator of metastasis (*Wotton et al.,* 2004; *Ben-Ami et al.,* 2009). In case of miR-122, TFs such as MAFB and SRF have been found active. SRF is an experimentally validated target of miR-122 and it regulates cell proliferation, differentiation, and cytoskeletal reorganization (*Bai et al.,* 2009).

5.3.9 miRNA-TF regulatory model upstream and downstream of TP53

The literature discussed in the previous section implies the involvement of the examined miRNAs and the identified TFs in cancer related processes. For a proof of concept of MIRTFnet, we analyze whether this common background is also reflected by a common set of TFs active across several of these experiments. Therefore, we compiled individual regulatory models (Figure 7) from all examined miRNA transfection experiments. The detailed models characterize the miRNA downstream actions in terms of kinases as well as active TFs that are mutually connected by interactions from databases or computational predictions (see Figure 17 as well as Supplementary Material on the website at http://www.bio.ifi.lmu.de/en/forschung /expression-analysis/mirtfnet). Interestingly, these models show substantial overlaps. In the following, we discuss the two intersection models constructed from the TFs and/or kinases contained in the regulatory networks (1) upstream and (2) downstream of TP53 that are contained in at least 7 of 19 individual models. By analyzing transfection experiments of sets of functionally related miRNAs we found that each set addresses a common core of transcription factors specific for that set.

The upstream miRNAs such as miR-155, miR-16 and let-7b are found to regulate TP53 (*Gironella et al.,* 2007; *Suzuki et al.,* 2009). The miRNAs such as miR-34, miR-192 and miR-215 are found to be regulated by the p53 transcription factor (*He et al.,* 2007; *Georges et al.,* 2008; *Subramanian et al.,* 2010). The upstream intersection model including miR-155, miR-16 and let-7b miRNA transfection, shows that these miRNAs regulate tumor suppressor TP53 and oncogeneic TFs (e.g., FOS, E2F1) (Figure 19). In comparion to the upstream model, in the downstream intersection model miR-34a/b, miR-192 miRNA transfection were found to regulate oncogeneic TFs (e.g., MAFB, ELK4, GATA3) (Figure 20). A minority of TFs is part of the upstream and downstream miRNA-TF models. These TFs regulate common oncogene TFs (e.g., CREB1, SPI1, etc). Thus, although the detected active TFs are all involved in cancer (further substantiated in the following section), the two regulatory models are quite distinct demonstrating that specific results are obtained from analyzing different sets of miRNAs.

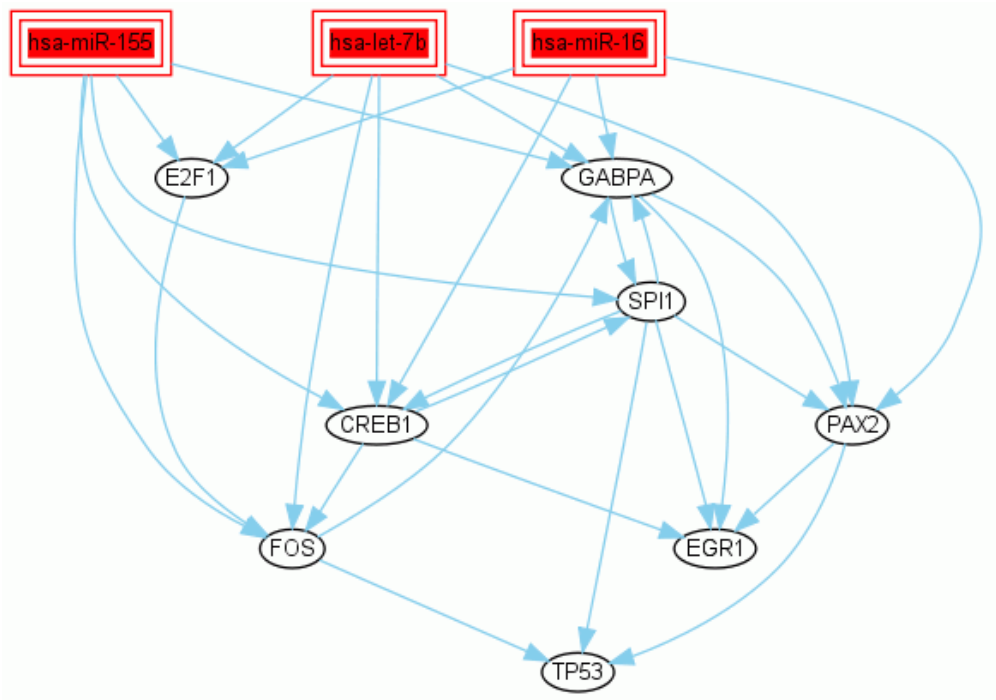

Figure 19: Upstream of TP53: intersection of miR-155, miR-16 and let-7b models. Individual regulatory models of the miR-155, miR-16 and let-7b transfection experiments are compiled by MIRTFnet and intersected (based on common set of active TFs). These models show substantial overlaps, regulating directly or indirectly oncogene TFs (such as TP53, FOS, CREB1). The interaction of these miRNAs with p53 have also been reported in the literature (*Gironella et al.,* 2007; *Suzuki et al.,* 2009).

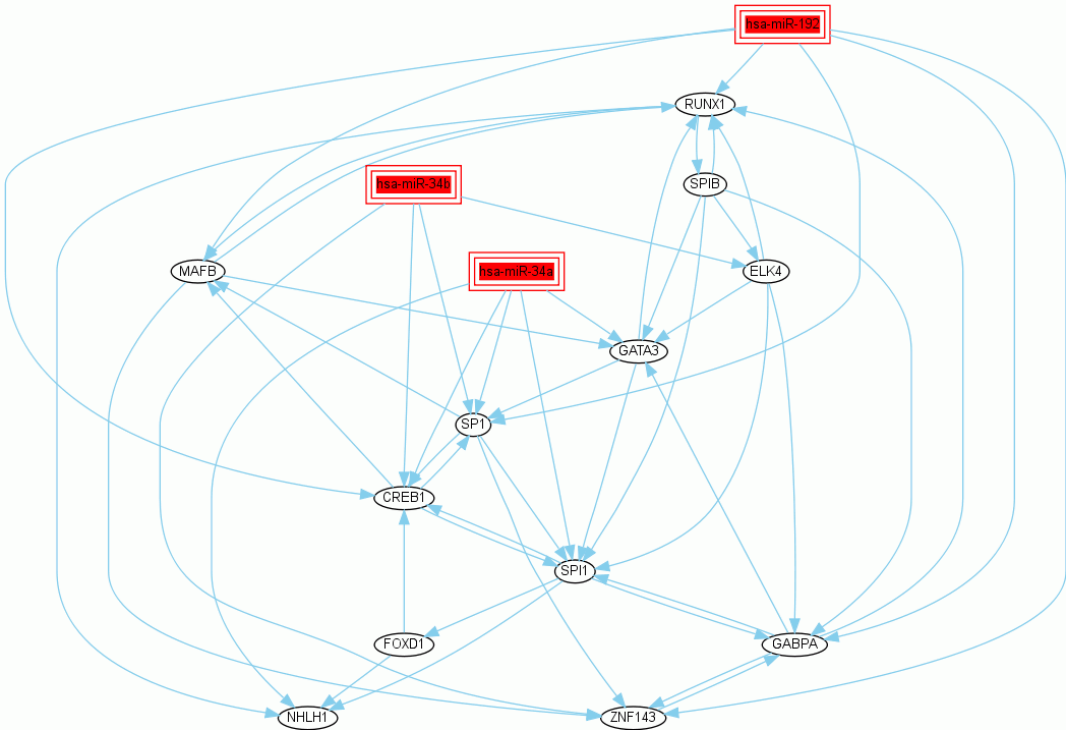

Figure 20: Downstream of TP53: intersection of miR-34a, miR-34b and miR-192 models. As in Figure 19, individual models of miR-192 and miR-34 microRNAs have been intersected. The shown microRNAs are reportedly regulated by TP53 and are thus potentially involved in cancer related processes (*Chang et al.,* 2007; *Hermeking et al.,* 2010).

5.3.10 Pathway and Gene Ontology analyses of the regulatory model

Here, we disregarded 6 examined datasets to avoid a bias towards miR-34. The intersection model contains 21 TFs and 34 kinases. We first analyzed the contained kinases. Kinases were included because of miRNA-kinase-active TF links, i.e. they usually do not receive direct support from the expression measurements. According to a pathway analysis using DAVID (*Huang et al.,* 2009), these kinases are associated with several KEGG signalling pathways including the MAPK, cancer, cell cycle and apoptosis pathways. 17 out of 34 kinases are part of the KEGG MAPK signalling pathway (e.g. MAPK9, MAPK8, CHUK, NLK and MAPK14). The MAPK signalling pathway is immediately connected to the p53 signalling pathway. 12 kinases are also part of the KEGG cancer signalling pathway (e.g. PTK2, MAPK3 and SKP2).

Notably, most of the active TFs detected by our approach are well known for their involvement in cancer. According to the DAVID analyzes in pathway databases such as KEGG or BioCarta, only cancer related pathways were detected with statistical significance. These included the KEGG pathways ‘prostate cancer’, ‘pancreatic

cancer', 'apoptosis' and 'pathways in cancer', which account for 10 of the TFs identified as active (i.e., ELK4, NFKB1, TP53, FOS, SPI1, CREB1, RELA, REL, E2F1 and ARNT). According to enrichment analysis of GO terms (DAVID), the TFs in the intersection model are associated with over 100 categories including cell differentiation.

For instance, CREB1 as well as the NFκB TF complex (NFKB1, RELA, REL) trigger cell survival and cell proliferation processes. Four additional TFs are oncogenes (REL, ELK4, MYB and MAFB). Another two TFs (PAX5 and SP1) are involved in cell differentiation, which also is a cancer associated process. For the remaining TF YY1 associations with cancer through p53 regulation have been reported in the literature (*Grönroos et al.,* 2004). The relationships between 19 of the 21 TFs as derived from the STRING database (*Jensen et al.,* 2009) are depicted in the Figure 21.

The details on the examined miRNAs, kinases and TFs as well as their interactions are available as supplementary tables published in *Naeem et al.* (2011). In addition to the above definition of a core model, the supplementary material thus enables analyzes on arbitrary combinations of the individual models (see Supplementary Material at http://www.bio.ifi.lmu.de/en/forschung/expression-analysis/mirtfnet).

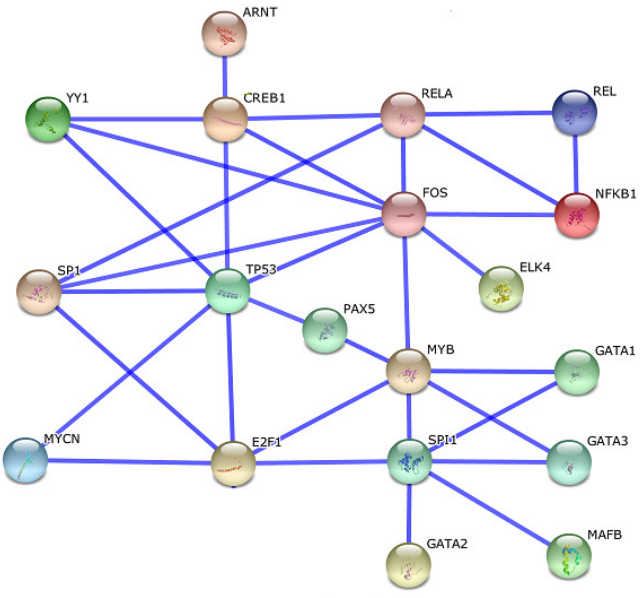

Figure 21: TF mediators of miRNA-triggered regulation active in at least 7 out of 19 miRNA-transfection experiments. The edges show association between TFs derived from the STRING database (http://string-db.org/) at a significance threshold of 0.9. The activity of TFs were determined applying MIRTFnet in miRNA transfection measurements (section 4.3).

6. Conclusion and discussion

6. Conclusion and discussion

This thesis aims to enhance the understanding of gene regulation controlled by microRNAs (miRNAs) and transcription factors (TFs). In order to examine the mechanisms of gene regulation we evaluated the experimental conditions where regulators (miRNAs and TFs) become active. For this purpose, we assessed the predictive ability of the enrichment tests to determine the activity of regulators based on their set of differentially expressed target genes. In turn, we collected regulators target gene sets comprehensively.

6.1 Contributions of this thesis

Our first contribution, miRSel database, increased the number of miRNA-gene associations by at least three-fold as compared to e.g. TarBase, a state-of-the-art resource for miRNA-gene relationships (*Naeem et al.,* 2010). Our second contribution, rigours assessment of 12 gene set enrichment (GSE) approaches/methods provide a guide for the selection of existing tests as well as a basis for the development and assessment of novel tests (*Naeem et al.,* 2011). Based on miRSel, our final contribution MIRTFnet detects active regulators (TFs) very reliably and explains a large part of the observed expression changes via models rooted at perturbed miRNAs (*Naeem et al.,* 2011). We discovered that a range of different miRNAs eventually induce activity changes in a common core of TFs involved in cancer related processes such as proliferation or apoptosis.

6.1.1 miRSel: Automated extraction of associations between microRNAs and genes from the biomedical literature

For the first step, we improved the coverage of miRNA and TF target gene sets through integrated analysis of several resources. Although, only few experimentally confirmed miRNA targets are available in databases. Many of the miRNA targets stored in databases were derived from large scale experiments that are considered not very reliable. The TarBase, a resource for miRNA-target relationships, for instance, contains only 1,134 (in human 1,031, mouse 101 and rat 2) such pairs (TarBase version 5). Moreover, only a fraction of the content in current databases has been derived by manual curation of experimentally validated targets. Instead, the major part of the content stems from the supplemental material of few research articles describing large scale experiments. *Ritchie et al.* (2009) proposed to exclude such studies for lack of a sufficient experimental validation. Only 262 out of 1,135 miRNA-target pairs remain after excluding just two such studies from TarBase.

In contrast to manual curation we proposed a simple, automated approach for biological name identification (named entity recognition, NER) that collects many potential targets for miRNAs not contained in current databases (published as miRSel, *Naeem et al.,* 2010). We found that text mining of miRNA, gene or protein names results in good recall and precision for miRNA-gene associations detected in single sentences. We thereby extracted many pairs from human (2,724 pairs), mouse (1,183 pairs) and rat (274 pairs) abstracts as well as 452 pairs from abstracts discussing other organisms. This represents an about 10-fold increase with respect to TarBase if

miRNA target pairs derived from large scale experiments are excluded (a threefold increase as compared to the whole TarBase). miRSel also characterized many miRNA target pairs with one of the five different association types. Here, 1,702 in human (62% of single sentence pairs), 813 in mouse (69% of single sentence pairs), and 219 (79% of single sentence pairs) in rat have been thus annotated in miRSel. Such an annotation is also available from public databases, but only for very few pairs, e.g. 199 pairs in TarBase. miRSel can also provide 7,799 pairs that co-occur in abstracts for human, 3,644 for mouse and 505 for rat, which are expected to be less reliable compared to pairs derived from single sentences.

To keep the miRSel database up-to-date, newly available PubMed abstracts are included daily. A full refresh of the synonym list generation and, subsequently, the scan of the entire PubMed is performed monthly to ensure the validity of all identifiers.

We also provide a web interface for querying miRSel via miRNA names, gene or protein names and via restricting the results using gene ontology terms or PubMed queries. We provide additional filter options, for instance to ensure the taxonomy context of matches.

6.1.2 Rigorous assessment of gene set enrichment tests

As a second step, we detected the activity of regulators based on our improved sets of regulator target genes. For the analysis of such gene sets, several gene set enrichment (GSE) tests have been established. These GSE tests originally proposed to detect an overrepresentation of differentially expressed genes in pre-defined gene sets that correspond to biological processes. However, a dependable standard-of-truth is not available since it is difficult to decide *a priori*, which biological processes will be affected on the mRNA level. This has previously prevented the objective selection and evaluation of enrichment tests. Instead, we derived gene sets from the targets of gene expression regulators including TFs and miRNAs whose experimental perturbation (e.g., TF overexpression or deletion) directly offers the required standard-of-truth. In this setting, we evaluated the ability of 12 different statistical tests to distinguish regulator perturbations from random fluctuations in the data (*Naeem et al.,* 2011). For method comparison, we focused on the most frequently used enrichment tests (*Subramanian et al.,* 2005; *Rivals et al.,* 2007; *Huang et al.,* 2008).

The detection of regulator activities is difficult: simple tests based on the rank difference between regulator targets and non-targets are not appropriate. We observed that ANOVA and Wilcoxons (WR) test consistently outperform other frequently used tests such as the Kolmogorov-Smirnov (KS) test. The hypergeometric (HG) test yields mixed results depending on the threshold parameter and the respective setting (TF vs. miRNA). Although the performance of the used tests was quite diverse (AUC between 0.5 and 0.85 for *E. coli* data), an unweighted consensus integrating all of the examined approaches consistently showed very good results.

Surprisingly, test performance did not improve by utilizing interaction signs (activation vs. inhibition). Here, we tested whether the fold changes observed in TF targets are consistent with the given interaction sign annotations. Fold changes and signs are clearly consistent in case of activation but not in case of inhibition. This might indicate either serious problems in the annotation of inhibiting relationships or

fundamentally different types of consequences from activation versus inhibition. While activator expression changes reliably cause target expression changes of the same sign, we did not find any similarly coherent relationship between the fold changes of repressors and their targets. According to *Herrgård et al.* (2003), this low correlation is due to the fact that either inhibitors or their targets exhibit low expression levels that cannot be profiled reliably.

To ensure the broad applicability of our results, we employed a variety of settings. In terms of microarray data, we used TF perturbations in *E. coli* (one expression compendium) and *S. cerevisiae* (two compendia) to compare results between a prokaryote and a eukaryote model organism. We also analyzed a third setting, the transfection of human cell lines with miRNAs. Performance on *S. cerevisiae* and human is lower than that for *E. coli*, which might be due to the lower quality of the available gold standards of TF and miRNA target networks as well as the more complex regulation in eukaryotes (*Hu et al.,* 2007; *Michoel et al,* 2009; *Narendra et al.,* 2010). The performance ranking of the tests is very consistent between each of the examined scenarios, with methods such as ANOVA or WR test always performing substantially better than the other methods.

Via an additional permutation approach, we analyzed how enrichment tests depend on the quality and comprehensiveness of the known regulator-target relationships. Most methods show only a moderate decrease in performance even after randomizing 50% of the gene regulatory network. We therefore conclude that the gene set definitions derived from the known gene regulatory interactions are sufficient to enable the comparative assessment of enrichment tests as well as the detection of regulator activities in real mRNA expression compendia.

6.1.3 MIRTFnet: Analysis of miRNA regulated transcription factors

Finally, we applied the selected GSE tests to assemble regulators cascades from expression profiles where cancer related miRNA have been over-expressed (published as MIRTFnet, *Naeem et al.,* 2011). In the examined experiments, we find that the WR test detected the transfecting miRNAs more reliably than the KS and HG tests (recall: WR=42/43=98%, KS=34/43=79%, HG=18/43=42% and AUC: WR=90%, KS=86%, HG=65%). The AUC improves to 91% if only those TFs are considered active that are detected by both WR and KS. Therefore, MIRTFnet reported TFs as active regulators if they are identified by both WR and KS tests.

The miRNAs used in the overexpression examined in this thesis were predominantly selected by the authors of the corresponding studies because of their reported involvement in cancer. In case of the detection of active TFs, we thus expected MIRTFnet to predominantly propose cancer related TFs. We could clearly confirm this expectation, and thereby ensure the reliability of our active TF predictions, as the involvement in cancer is indeed known for almost all of our detected TFs.

Starting from the over-expressing miRNA, we constructed putative models based on known or predicted regulator (i.e. miRNA, TF and kinase) target relationships. For each examined overexpression experiment, most of the detected TFs could be connected directly or indirectly to the transfecting miRNA. Indirect connections in our models included miRNA-kinase-TF and miRNA-TF-TF relationships. Our models provide potential explanations for the majority of the observed expression

changes as all known TFs were tested by MIRTFnet. These models also contained relationships to unregulated genes. This is not surprising as many genes might be regulated in a synergistic fashion, i.e. require different regulators being active at the same time. Relationships to unregulated genes might also be caused by incorrect target predictions.

An additional unexpected result stems from intersecting the proposed regulatory models constructed for the individual miRNAs. We detected several active TFs across many different overexpression studies. This could potentially suggest common regulatory mechanisms downstream of cancer relevant miRNAs or of the respective TFs (i.e. p53). At the same time, the responses of TFs to different subsets of miRNAs can be quite distinct depending on whether these miRNA act either upstream or downstream of p53 (Figures 19 and 20).

Our results further reinforce the growing awareness that these small non-coding RNAs have an intrinsic function in gene regulatory networks including TFs related to key cellular contexts such as cell proliferation and apoptosis.

6.2 Perspectives for future research

In this thesis we described novel methods together with new databases for determining the activity changes of regulators and interaction between different regulators using high-throughput data. These methods give rise to several areas for future research.

To improve the coverage of current repositories, we proposed to use text mining of publication abstracts for extracting miRNA-gene associations including miRNA-target relations (miRSel, sections 4.1 and 5.1) in human, mouse and rat. We provide a web interface for querying miRSel via miRNA names, gene or protein names and via restricting the results using gene ontology terms or PubMed queries. For future development the miRSel can be extended to include the other species and full free text to further improve the coverage of regulator target gene sets. The database can be integrated with different heterogeneous data resources (section 3) to provide a comprehensive data set to assess miRNA targeting features in different species that will be useful for the validation, development of computational target prediction programs and deciphering diverse biological functions of miRNAs and their regulation in various diseases.

Secondly, to provide a guide for the selection of existing GSE tests and basis for the development and assessment of novel tests, we performed the rigorous comparative assessment of 12 GSE tests for analysing gene sets derived as miRNA and TF target genes in *E.coli*, *S. cerevisiae* and human. For future work, gene set definition can be extended to several other species and state-of-the-art other statistical methods as well as techniques for identifying the experimentally perturbed regulator activity can be investigated. Additionally, the presented work can be further analyzed together with other databases (e.g. RegulonDB and YEASTRACT) to improve the annotation of interaction signs (sections 4.2 and 5.2, activation vs. inhibition).

Finally, we mined the miRNA and TF activation in miRNA-induced gene expression measurements applying MIRTFnet that sheds light on the scope of the extended regulatory effects downstream of miRNAs (sections 4.3 and 5.3). This study reveals the following areas for future work.

1) Computational identification of the activity of kinases is important to understanding the regulation of gene expression and dynamic cellular mechanisms. In MIRTFnet, we included the kinases as connectors between miRNAs and TFs in the network models although the activity of kinases has not been determined. The MIRTFnet can be extended to determine the activity of the kinases if their downstream targets including TFs and their target genes were differentially expressed.
2) The extracted network models (section 5.3) can be used to further increase the reliability of miRNA-target predictions and to fully understand miRNA extended regulatory effects and functions as well.

Bibliography

Abdulrehman D, Monteiro PT, Teixeira MC, Mira NP, Lourenço AB, dos Santos SC, Cabrito TR, Francisco AP, Madeira SC, Aires RS, Oliveira AL, Sá-Correia I, Freitas AT. (2011). YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic Acids Res*, **39**(Database issue):D136-40.

Ackermann M, Strimmer K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10:**47.

Addo-Quaye C, Eshoo TW, Bartel DP, Axtell MJ. (2008). Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. *Curr Biol*, **18**(10):758-62.

Aho, VA, Corasick MJ. (1975). Efficient string matching: An aid to bibliographic search. *Communications of the ACM*, **18**: 333–340.

Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG. (2009). Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, **25**(23):3049-55.

Alexiou P, Vergoulis T, Gleditzsch M, Prekas G, Dalamagas T, Megraw M, Grosse I, Sellis T, Hatzigeorgiou AG. (2009). miRGen 2.0: a database of microRNA genomic information and regulation. *Nucleic Acids Res*, 1-5.

Al-Shahrour F, Díaz-Uriarte R, Dopazo J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics,* **20**(4):578-80.

Amberger J, Bocchini CA, Scott AF, Hamosh A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res*, **37**(Database issue): D793-6.

Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, Matzke M, Ruvkun G, Tuschl T. (2003). A uniform system for microRNA annotation. *RNA*, **9**(3):277-279.

Ananiadou S, Kell DB, Tsujii J. (2006). Text mining and its potential applications in systems biology. *Trends Biotechnol*, **24**:571-579.

Ananiadou S, Mcnaught J. (2005). Text Mining for Biology And Biomedicine. *Artech House*, 1-286.

Anjum R, Blenis J. (2008). The RSK family of kinases: emerging roles in cellular signalling. *Nat Rev Mol Cell Biol*, **9**(10):747-58.

Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, Kerssemakers J, Leroy C, Menden M, Michaut M, Montecchi-Palazzi L, Neuhauser SN, Orchard S, Perreau V, Roechert B, van Eijk K, Hermjakob H. (2010). The IntAct molecular interaction database in 2010. *Nucleic Acids Res*, **38**(Database issue):D525-31.

Arora A, Simpson DA. (2008). Individual mRNA expression profiles reveal the effects of specific microRNAs. *Genome Biol*, **9**(5):R82.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. (2000). The Gene Ontology Consortium. *Nat Genet*, **25**(1):25-9.

Bader GD, Cary MP, Sander C. (2006). Pathguide: a pathway resource list. *Nucleic Acids Res*, **34**(Database issue):D504-6.

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

Bai J, Wang J, Xue F, Li J, Bu L, Hu J, Xu G, Bao Q, Zhao G, Ding X, Yan J, Wu J. (2010). proTF: a comprehensive data and phylogenomics resource for prokaryotic transcription factors. *Bioinformatics,* **26**(19):2493-5.

Bai S, Nasser MW, Wang B, Hsu SH, Datta J, Kutay H, Yadav A, Nuovo G, Kumar P, Ghoshal K. (2009). MicroRNA-122 inhibits tumorigenic properties of hepatocellular carcinoma cells and sensitizes these cells to sorafenib. *J Biol Chem*, **284**(46):32015-27.

Barh D, Malhotra R, Ravi B, Sindhurani P (2010) Microrna let-7: an emerging next-generation cancer therapeutic. *Curr Oncol*, **17**:70-80.

Barry WT, Nobel AB, Wright FA. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics,* **21**:1943-1949.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R. (2007). NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res*, **35**(Database issue):D760-5.

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

Beissbarth T, Speed TP. (2004). GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics,* **20**:1464-1465.

Bell E, Lunec J, Tweddle DA. (2007). Cell cycle regulation targets of MYCN identified by gene expression microarrays. *Cell Cycle*, **6**:1249-1256.

Ben-Ami O, Pencovich N, Lotem J, Levanon D, Groner Y. (2009). A regulatory interplay between miR-27a and Runx1 during megakaryopoiesis. *Proc Natl Acad Sci USA*, **106**(1):238-43.

Benjamini Y, Yekutieli D. (2001). The Control of the false discovery rate in multiple testing under dependency. *Ann Statist*, **29**:1165-1188.

Betel D, Wilson M, Gabow A, Marks DS, Sander C. (2008). The microRNA.org resource: targets and expression. *Nucleic Acids Res*, **36**(Database issue):D149-53.

Berardini TZ, Li D, Huala E, Bridges S, Burgess S, McCarthy F, Carbon S, Lewis SE, Mungall CJ, Abdulla A, Wood V, Feltrin E, Valle G, Chisholm RL, Fey P, Gaudet P, Kibbe W, Basu S, Bushmanova Y, Eilbeck K, Siegele DA, McIntosh B, Renfro D, Zweifel A, Hu JC, Ashburner M, Tweedie S, Alam-Faruque Y, Apweiler R, Auchinchloss A, Bairoch A, Barrell D, Binns D, Blatter MC, Bougueleret L, Boutet E, Breuza L, Bridge A, Browne P, Chan WM, Coudert E, Daugherty L, Dimmer E, Eberhardt R, Estreicher A, Famiglietti L, Ferro-Rojas S, Feuermann M, Foulger R, Gruaz-Gumowski N, Hinz U, Huntley R, Jimenez S, Jungo F, Keller G, Laiho K, Legge D, Lemercier P, Lieberherr D, Magrane M, O'Donovan C, Pedruzzi I, Poux S, Rivoire C, Roechert B, Sawford T, Schneider M, Stanley E, Stutz A, Sundaram S, Tognolli M, Xenarios I, Harris MA, Deegan JI, Ireland A, Lomax J, Jaiswal P, Chibucos M, Giglio MG, Wortman J, Hannick L, Madupu R, Botstein D, Dolinski K, Livstone MS, Oughtred R, Blake JA, Bult C, Diehl AD, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Sitnikov D, Collmer C, Torto-Alalibo T, Laulederkind S, Shimoyama M, Twigger S, D'Eustachio P, Matthews L, Balakrishnan R, Binkley G, Cherry JM, Christie KR, Costanzo MC, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Hong EL, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Weng S, Wong ED, Aslett M, Chan J, Kishore R, Sternberg P, Van Auke K, Khodiyar VK, Lovering RC, Talmud PJ, Howe D, Westerfield M. (2010). The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res*, **38**(Database issue):D331-5.

Bewick V, Cheek L, Ball J. (2004). Statistics review 13: receiver operating characteristic curves. *Crit Care,* **8**(6):508-12.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, **31**:365-70.

Boes DC, Graybill FA, Mood AM. (1974). Introduction to the Theory of Statistics, 3rd ed. New York: McGraw-Hill.

Borda J. (1781). Memoire sur les elections au scrutin. Histoire de l'Academie des Sciences, Paris.

Boorsma A, Lu XJ, Zakrzewska A, Klis FM, Bussemaker HJ. (2008). Inferring condition-specific modulation of transcription factor activity in yeast through regulon-based analysis of genomewide expression. *PLoS One,* **3**(9):e3112

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. (2007). UniProtKB/Swiss-Prot. *Methods Mol Biol*, **406**:89-112.

Boyerinas B, Park SM, Hau A, Murmann AE, Peter ME. (2010). The role of let-7 in cell differentiation and cancer. *Endocr Relat Cancer*, **17**:F19-36.

Braun CJ, Zhang X, Savelyeva I, Wolff S, Moll UM, et al. (2008). p53-Responsive MicroRNAs 192 and 215 are capable of inducing cell cycle arrest. *Cancer Res*, **68**: 10094-104.

Brazma A. (2009). Minimum Information About a Microarray Experiment (MIAME)--successes, failures, challenges. *Scientific World Journal*, **9**:420-3.

Brodersen P, Voinnet O. (2009). Revisiting the principles of microRNA target recognition and mode of action. *Nat Rev Mol Cell Biol.*, **10**(2):141-148.

Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E. (2008). The HGNC Database in 2008: a resource for the human genome, *Nucleic Acids Res*, **36**: D445-448.

Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA; Mouse Genome Database Group. (2008). The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res*, **36**:724-728.

Calin GA, Cimmino A, Fabbri M, Ferracin M, Wojcik SE, Shimizu M, Taccioli C, Zanesi N, Garzon R, Aqeilan RI, Alder H, Volinia S, Rassenti L, Liu X, Liu CG, Kipps TJ, Negrini M, Croce CM. (2008). MiR-15a and miR-16-1 cluster functions in human leukemia. *Proc Natl Acad Sci USA*, **105**(13):5166-71.

Chan CS, Elemento O, Tavazoie S. (2005). Revealing posttranscriptional regulatory elements through network-level conservation. *PLoS Comput Biol,* **1**(7):e69.

Chang TC, Wentzel EA, Kent OA, Ramachandran K, Mullendore M, et al. (2007). Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis. *Mol Cell*, **26**:745-752.

Chen K, Rajewsky N. (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet*, **8**(2):93-103.

Cheng C, Fu X, Alves P, Gerstein M. (2009). mRNA expression profiles show differential regulatory effects of microRNAs between estrogen receptor-positive and estrogen receptor-negative breast cancer. *Genome Biol,* **10**(9):R90.

Cheng C, Li LM (2008). Inferring microRNA activities by combining gene expression with microRNA target prediction. *PLoS One*, **3**:e1989.

Chi SW, Zang JB, Mele A, Darnell RB. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature,* **460**(7254):479-86.

Chien CT, Bartel PL, Sternglanz R, Fields S. (1991). The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc Natl Acad Sci USA*, **88**(21):9578-82.

Chua G, Morris QD, Sopko R, Robinson MD, Ryan O, Chan ET, Frey BJ, Andrews BJ, Boone C, Hughes TR.Source. (2006). Identifying transcription factor functions and targets by phenotypic activation. *Proc Natl Acad Sci USA,* **103**(32):12045-50.

Magic/lCimmino A, Calin GA, Fabbri M, Iorio MV, Ferracin M, et al. (2005). miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc Natl Acad Sci USA* **102**:13944-13949.

Cohen AM, Hersh WR. (2005). A survey of current work in biomedical text mining. *Brief Bioinform*, **6**(1):57-71.

Collier N, Obata C, Tsujii JI. (2000). Extracting the names of genes and gene products with a hidden Markov model. *In Proceedings of the 18th conference on Computational linguistics* -Volume 1.

Corney DC, Flesken-Nikitin A, Godwin AK, Wang W, Nikitin AY. (2007). MicroRNA-34b and MicroRNA-34c are targets of p53 and cooperate in control of cell proliferation and adhesion-independent growth. *Cancer Res*, **67**:8433-8.

Csaba, G. (2008). syngrep – Fast synonym-based named entity recognition, Personal communication, LMU, Munich.

Nishimura D. (2001). BioCarta. *Biotech Software & Internet Report*. **2**(3): 117-120.

Dai Y, Zhou X. (2010). Computational methods for the identification of microRNA targets. *Dovepress Journal*, **2**:29-39.

DeGroot MH. (1991). Ch. 9 in Probability and Statistics, 3rd ed. MA: Addison-Wesley.

D'Eustachio P. (2011). Reactome knowledgebase of human biological pathways and processes. *Methods Mol Biol*, **694**:49-61.

Eaton AD. (2006). HubMed: a web-based biomedical literature search interface. *Nucleic Acids Res*, **34**(Web Server issue):W745-7.

Efron B, Tibshirani RJ. (1993). An Introduction to the Bootstrap. *Chapman & Hall*.

Efron B, Tibshirani R. (2006). On testing the significance of sets of genes. Ann. Appl. Stat., 107-129.

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

Erhardt RA, Schneider R, Blaschke C. (2006). Status of text-mining techniques applied to biomedical text. *Drug Discovery Today*, **11**:315-325.

Errami M, Wren JD, Hicks JM, Garner HR. (2007). eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. *Nucleic Acids Res*, **35**(Web Server issue):W12-5.

Esquela-Kerscher A, Slack FJ. (2006). Oncomirs -microRNAs with a role in cancer. *Nat Rev Cancer*, **6**:259-269.

Essaghir A, Toffalini F, Knoops L, Kallin A, van Helden J, et al. (2010). Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. *Nucleic Acids Res* **38**:e120.

Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS. (2008). Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res,* **36**(Database issue):D866-70.

Farh KK, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP. (2005). The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, **10**(5755):1817-21.

Fawcett T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27, 861–874.

Feller W. (1968). The Hypergeometric Series. An Introduction to Probability Theory and Its Applications. *Wiley*, 41-45.

Flicek P, Aken BL, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Gräf S, Haider S, Hammond M, Howe K, Jenkinson A, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Koscielny G, Kulesha E, Lawson D, Longden I, Massingham T, McLaren W, Megy K, Overduin B, Pritchard B, Rios D, Ruffier M, Schuster M, Slater G, Smedley D, Spudich G, Tang YA, Trevanion S, Vilella A, Vogel J, White S, Wilder SP, Zadissa A, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Herrero J, Hubbard TJ, Parker A, Proctor G, Smith J, Searle SM. (2010). Ensembl's 10th year. *Nucleic Acids Res*, **38**(Database issue):D557-62

Fontaine JF, Barbosa-Silva A, Schaefer M, Huska MR, Muro EM, Andrade-Navarro MA. (2009). MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Res*, **37**(Web Server issue):W141-6.

Friedman RC, Farh KK, Burge CB, Bartel DP. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, **19**:92-105.

Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ. (2011). The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*, **39** (Database issue):D876-82.

Fukuda K, Tamura A, Tsunoda T, Takagi T. (1998). Toward information extraction: identifying protein names from biological papers. *Pac Symp Biocomput*, 707-18.

Fundel K, Güttler D, Zimmer R, Apostolakis J. (2005). A simple approach for protein name identification: prospects and limits. *BMC Bioinformatics*, **6**(Suppl 1):S15.

Fundel K, Zimmer R. (2006). Gene and protein nomenclature in public databases. *BMC Bioinformatics*, 7:372.

Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muñiz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, García-Sotelo JS, López-Fuentes A, Porrón-Sotelo L, Alquicira-Hernández S, Medina-Rivera A, Martínez-Flores I, Alquicira-Hernández K, Martínez-Adame R, Bonavides-Martínez C, Miranda-Ríos J, Huerta AM, Mendoza-Vargas A, Collado-Torres L, Taboada B, Vega-Alvarado L, Olvera M, Olvera L, Grande R, Morett E, Collado-Vides J. (2011). RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res,* **39**(Database issue):D98-105.

Gansner ER, North SC. (2000). An open graph visualization system and its applications to software engineering. *Software: Practice and Experience*, 1203-1233.

Gatti DM, Barry WT, Nobel AB, Rusyn I, Wright FA. (2010). Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics,* 11:574.

Georges SA, Biery MC, Kim SY, Schelter JM, Guo J, et al. (2008). Coordinated regulation of cell cycle transcripts by p53-Inducible microRNAs, miR-192 and miR-215. *Cancer Res*, **68**:10105-12.

German MA, Pillay M, Jeong DH, Hetawal A, Luo S, Janardhanan P, Kannan V, Rymarquis LA, Nobuta K, German R, De Paoli E, Lu C, Schroth G, Meyers BC, Green PJ. (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat Biotechnol*, **26**(8):941-6.

Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, Enright AJ, Schier AF. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science,* 312(5770):75-9.

Gironella M, Seux M, Xie MJ, Cano C, Tomasini R, et al. (2007). Tumor protein 53-induced nuclear protein 1 expression is repressed by miR-155, and its restoration inhibits pancreatic tumor development. *Proc Natl Acad Sci USA*, **104**:16170-5.

Glaab E, Baudot A, Krasnogor N, Valencia A. (2010). Extending pathways and processes using molecular interaction networks to analyse cancer genome data. *BMC Bioinformatics*, 11:597.

Goeman JJ, Bühlmann P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics,* **23**(8): 980-7.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Res*, **36**:D154-8

Griffiths-Jones S. (2004). The microRNA Registry. *Nucleic Acids Res*, **32**:109-111.

Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, et al. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, **27**:91-105.

Grönroos E, Terentiev AA, Punga T, Ericsson J. (2004). YY1 inhibits the activation of the p53 tumor suppressor in response to genotoxic stress. *Proc Natl Acad Sci USA*, **101**:12165-70.

Grün D, Wang YL, Langenberger D, Gunsalus KC, Rajewsky N. (2005). microRNA target predictions across seven Drosophila species and comparison to mammalian targets. *PLoS Comput Biol*, **1**(1):e13.

Gsponer J, Futschik ME, Teichmann SA, Babu MM. (2008). Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science* **322**(5906):1365-8.

Guo AY, Chen X, Gao G, Zhang H, Zhu QH, Liu XC, Zhong YF, Gu X, He K, Luo J. (2008). PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Res*, **36**(Database issue):D966-9.

Hackney JA, Charbord P, Brunk BP, Stoeckert CJ, Lemischka IR, Moore KA. (2002). A molecular profile of a hematopoietic stem cell niche. *Proc Natl Acad Sci USA*, **99**(20):13061-6.

Hammell M et al. (2008). mirWIP: microRNA Target Prediction Based on miRNP Enriched Transcripts. Nat Methods, **5**(9): 813-819.

Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J. (2005). ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, **6**(Suppl 1):S14.

Hanisch D, Fluck J, Mevissen HT, Zimmer R. (2003). Playing biology's name game: identifying protein names in scientific text. *Pac Symp Biocomput*, 403-14.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature,* **431**(7004):99-104.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R. (2004). Gene Ontology Consortium: The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, **32**:D258-61.

Hashimoto K, Goto S, Kawano S, Aoki-Kinoshita KF, Ueda N, Hamajima M, Kawasaki T, Kanehisa M. (2005). KEGG as a glycome informatics resource. *Glycobiology*, **16**(5):63R-70R.

He L, He X, Lim LP, de Stanchina E, Xuan Z, et al. (2007). A microRNA component of the p53 tumour suppressor network. *Nature,* **447**:1130-1134.

Hermeking H. (2010). The miR-34 family in cancer and apoptosis. *Cell Death Differ,* **17**:193-199.

Hermoso A, Aguilar D, Aviles FX, Querol E. (2004). TrSDB: a proteome database of transcription factors. *Nucleic Acids Res*, **32**(Database issue):D171-3.

Herrgård MJ, Covert MW, Palsson BØ. (2003). Reconciling gene expression data with known genome-scale regulatory network structures. Genome Res., **13**(11):2423-34.

Hoang P. (2006). Springer Handbook of Engineering Statistics. *XLIV,* **1120** p. 384.

Hobert O. (2008). Gene regulation by transcription factors and microRNAs. *Science*, **319**:1785-1786.

Hochheimer A, Tjian R. (2003). Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. *Genes Dev*, **17**(11):1309-20.

Hosack DA, Dennis G Jr, Sherman BT, Lane HC, Lempicki RA. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biol,* **4**(10):R70.

Hsu SD, Chu CH, Tsou AP, Chen SJ, Chen HC, Hsu PW, Wong YH, Chen YH, Chen GH, Huang HD. (2008). miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. *Nucleic Acids Res*, **36**:165-169.

Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, Tsai WT, Chen GZ, Lee CJ, Chiu CM, Chien CH, Wu MC, Huang CY, Tsou AP, Huang HD. (2011). miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res,* **39**(Database issue):D163-9.

Hu H. (2010). An efficient algorithm to identify coordinately activated transcription factors. *Genomics,* **95**(3):143-50.

Hu Z, Killion PJ, Iyer VR. (2007). Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet,* **39**(5):683-7.

Huang DW, Sherman BT, Lempicki RA. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc*, **4**:44-57.

Huang JC, Babak T, Corson TW, Chua G, Khan S, et al. (2007). Using expression profiling data to identify human microRNA targets. *Nat Methods*, **4**:1045-1049.

Huang da W, Sherman BT, Lempicki RA. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res,* **37**(1):1-13.

Huerta AM, Salgado H, Thieffry D, Collado-Vides J. (1998). RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res,* **26**(1):55-9.

Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, Konagaya A, Shinozaki K. RARTF: database and tools for complete sets of Arabidopsis transcription factors. *DNA Res*, **12**(4):247-56.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**(6819):533-8.

Jensen LJ Saric J, Bork P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*, **7**:119-129.

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, **37**:D412-6.

Jeyaseelan K, Lim KY, Armugam A. (2008). MicroRNA expression in the blood and brain of rats subjected to transient focal ischemia by middle cerebral artery occlusion. Stroke, **39**(3):959-66.

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

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. (2004). Human MicroRNA targets. *PLoS Biol*, **2**(11):e363.

Johnson DS, Mortazavi A, Myers RM, Wold B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science,* **316**(5830):1497-502.

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*, **33**(Database issue):D428-32.

Jovanovic M, Hengartner MO. (2006). miRNAs and apoptosis: RNAs to die for. *Oncogene*, **25**(46):6176-87.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res*, **32**(Database issue): D277-80.

Kanehisa M, Goto S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**(1):27-30.

Kazakov AE, Cipriano MJ, Novichkov PS, Minovitsky S, Vinogradov DV, Arkin A, Mironov AA, Gelfand MS, Dubchak I. (2007). RegTransBase--a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Res,* **35**(Database issue):D407-12.

Kel' AE, Kolchanov NA, Kel' OV, Romashchenko AG, Anan'ko EA, Ignat'eva EV, Merkulova TI, Podkolodnaia OA, Stepanenko IL, Kochetov AV, Kolpakov FA, Podkolodnyĭ NL, Naumochkin AA. (1997). TRRD: a database of transcription regulatory regions in eukaryotic genes. *Mol Biol (Mosk)*, **31**(4):626-36.

Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H. (2007). IntAct--open source resource for molecular interaction data. *Nucleic Acids Res*, **35**(Database issue): D561-5.

Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. (2007). The role of site accessibility in microRNA target recognition. *Nat Genet*, **39**(10):1278-84.

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009). Human Protein Reference Database - 2009 Update. *Nucleic Acids Res* **37**:D767-72.

Khatri P, Drăghici S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics,* **21**:3587-3595.

Kim JM, Nakao K, Nakamura K, Saito I, Katsuki M, et al. (2002). Inactivation of Cdc7 kinase in mouse ES cells results in S-phase arrest and p53-dependent cell death. *EMBO J*, **21**:2168-2179.

Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, Hatzigeorgiou A. (2004). A combined computational-experimental approach predicts human microRNA targets. *Genes Dev*, **18**:1165-78.

Kozomara A, Griffiths-Jones S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, **39**(Database issue):D152-7.

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

Küffner R et al. (2011). Inferring Gene Regulatory Networks by ANOVA, submitted to ISMB2011.

Kuhn DE, Martin MM, Feldman DS, Terry AV Jr, Nuovo GJ, Elton TS. (2008). Experimental validation of miRNA targets. *Methods*, **44**(1):47-54.

Kuss AW, Chen W. (2008). MicroRNAs in brain function and disease. *Curr Neurol Neurosci Rep*, **8**(3):190-7.

Lall S, Grün D, Krek A, Chen K, Wang YL, Dewey CN, Sood P, Colombo T, Bray N, Macmenamin P, Kao HL, Gunsalus KC, Pachter L, Piano F, Rajewsky N. (2006). A genome-wide map of conserved microRNA targets in C. elegans. *Curr Biol*, **16**(5):460-71.

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

Lee HK, Braynen W, Keshav K, Pavlidis P. (2005). ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics,* 6:269.

Legrain P, Selig L. (2000). Genome-wide protein interaction maps using two-hybrid systems. *FEBS Lett*, **480**(1):32-6.

Lehmann EL. (1975). Nonparametric Statistical Methods Based on Ranks. McGraw-Hill.

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

Levine DM, et al. (2006). Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. Genome Biol., **7**(10):R93.

Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. (2003). Prediction of mammalian microRNA targets. *Cell*, **115**:787-798.

Li M, Li J, Ding X, He M, Cheng SY (2010). microRNA and Cancer. *AAPS J*, **12**: 309-317.

Li M, Marin-Muller C, Bharadwaj U, Chow KH, Yao Q, Chen C. (2009). MicroRNAs: control and loss of control in human physiology and disease. *World J Surg*, **33**(4):667-684.

Liang H, Li WH. (2007). MicroRNA regulation of human protein protein interaction network. *RNA,* **13**(9):1402-8.

Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, **433**(7027):769-73.

Liu GE, Weirauch MT, Van Tassell CP, Li RW, Sonstegard TS, et al. (2008). Identification of conserved regulatory elements in mammalian promoter regions: a case study using the PCK1 promoter. *Genomics Proteomics Bioinformatics*, **6**:129-143.

Liu J. (2008). Control of protein synthesis and mRNA degradation by microRNAs. *Curr Opin Cell Biol*, **20**:214-221.

Liu Q, Fu H, Sun F, Zhang H, Tie Y, et al. (2008). miR-16 family induces cell cycle arrest by regulating multiple cell cycle genes. *Nucleic Acids Res*, **36**:5391-404.

Liu Q, Tan Y, Huang T, Ding G, Tu Z, et al. (2010). TF-centered downstream gene set enrichment analysis: Inference of causal regulators by integrating TF-DNA interactions and protein post-translational modifications information. *BMC Bioinformatics*, **11**:S5.

Lodish H, Berk A, Matsudaira P, Kaiser CA, Krieger M, Scott MP, Zipurksy SL, Darnell J (2004). Molecular Cell Biology.

Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, et al. (2005). MicroRNA expression profiles classify human cancers. *Nature*, **435**:834-838.

Lu Z. (2011). PubMed and beyond: a survey of web tools for searching biomedical literature. *Database (Oxford),* **11**: baq036.

Ma L, Young J, Prabhala H, Pan E, Mestdagh P, et al. (2010). miR-9, a MYC/MYCN-activated microRNA, regulates E-cadherin and cancer metastasis. *Nat Cell Biol*, **12**(3):247-56.

MacArthur S, Li XY, Li J, Brown JB, Chu HC, Zeng L, Grondona BP, Hechmer A, Simirenko L, Keränen SV, Knowles DW, Stapleton M, Bickel P, Biggin MD, Eisen MB. (2009). Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol*, **10**(7):R80.

Maglott D, Ostell J, Pruitt KD, Tatusova T. (2011). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res,* **39**(Database issue):D52-7.

Maglott D, Ostell J, Pruitt KD, Tatusova T. (2007). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, **35**(Database issue):D26-31.

Mangan ME, Williams JM, Lathe SM, Karolchik D, Lathe WC. (2008). UCSC genome browser: deep support for molecular biomedical research. *Biotechnol Annu Rev*, **14**:63-108.

Mann HB, Whitney DR. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics*, **18**(1):50-60.

Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, Vergoulis T, Koziris N, Sellis T, Tsanakas P, Hatzigeorgiou AG. (2009). DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res*, **37**(Web Server issue):W273-6.

Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Münch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, **31**(1):374-8.

MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E. (2006). An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**:113.

Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B. (2004). GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol,* **5**(12):R101.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, **34**:D108-10.

Mazière P, Enright AJ. (2007). Prediction of microRNA targets. *Drug Discov Today*, **12**(11-12):452-8.

Megraw M, Sethupathy P, Corda B, Hatzigeorgiou AG. (2007). miRGen: a database for the study of animal microRNA genomic organization and function. *Nucleic Acids Res,* **35**(Database issue):D149-55.

Michoel T, De Smet R, Joshi A, Van de Peer Y, Marchal K. (2009). Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Syst. Biol,* 3:49.

Mika S, Rost B. (2004). Protein names precisely peeled off free text. *Bioinformatics*, 20 Suppl 1: i241-7.

Miller RG. (1997). Beyond ANOVA: Basics of Applied Statistics. Boca Raton, FL: Chapman & Hall.

Min H, Yoon S. (2010). Got target? Computational methods for microRNA target prediction and their extension. *Exp Mol Med*, **42**(4):233-44.

Miranda-Saavedra D, De S, Trotter MW, Teichmann SA, Göttgens B. (2009). BloodExpress: a database of gene expression in mouse haematopoiesis. *Nucleic Acids Res*, **37**(Database issue):D873-9.

Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, Lim B, Rigoutsos I. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**(6):1203-17.

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet,* **34**(3):267-73.

Naeem H, Küffner R, Zimmer R. (2011). MIRTFnet: Analysis of miRNA regulated transcription factors. *PLoS One*, **6**(8):e22519.

Naeem H, Zimmer R, Tavakkolkhah P, Küffner R. (2011). Rigorous assessment of gene set enrichment tests (submitted to *Bioinformatics*, revision).

Naeem H, Küffner R, Csaba G, Zimmer R. (2010). miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature. *BMC Bioinformatics*, **11**:1-8.

Nam D, Kim SY. (2008). Gene-set approach for expression pattern analysis. *Brief Bioinform,* **9**(3):189-97.

Nam S, Kim B, Shin S, Lee S. (2008). miRGator: an integrated system for functional annotation of microRNAs. *Nucleic Acids Res,* **36**:159-164.

Narendra V, Lytkin NI, Aliferis CF, Statnikov A. (2010). A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks. *Genomics,* **97**(1):7-18.

Navarro G, Raffinot M. (2002). Flexible Pattern Matching in Strings -- Practical on-line search algorithms for texts and biological sequences. Cambridge University Press.

Nestler EJ, Hyman S.E, Malenka R. (2004). Molecular Neuropharmacology, *McGraw-Hill Professional*, 1-498.

Nikiforov, A.M. (1994). Algorithm AS 288: Exact Smirnov two-sample tests for arbitrary distributions. *Applied Statistics,* **43**: 265-284.

Orom UA, Lund AH. (2009). Experimental identification of microRNA targets. *Gene*, 1-5.

Ott CE, Grünhagen J, Jäger M, Horbelt D, Schwill S, et al. (2011). MicroRNAs differentially expressed in postnatal aortic development downregulate elastin via 3' UTR and coding-sequence binding sites. *PLoS One*, **6**(1):e16250.

Otto T, Horn S, Brockmann M, Eilers U, Schüttrumpf L, et al. (2009). Stabilization of N-Myc is a critical function of Aurora A in human neuroblastoma. *Cancer Cell*, **15**:67-78.

Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, Hatzigeorgiou AG. (2009). The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res*, **37**:D155-8.

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A. (2007). ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*, **35**(Database issue):D747-50.

Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E. (2004). Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res,* **29**(6):1213-22.

Pehkonen P, Wong G, Törönen P. (2005). Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinformatics,* 6:162.

Pérez-Rodríguez P, Riaño-Pachón DM, Corrêa LG, Rensing SA, Kersten B, Mueller-Roeber B. (2010). PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Res*, **38**(Database issue):D822-7.

Portales-Casamar E, Kirov S, Lim J, Lithwick S, Swanson MI, Ticoll A, Snoddy J, Wasserman WW. (2007). PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation. *Genome Biol*, **8**(10):R207.

Porter CJ, Palidwor GA, Sandie R, Krzyzanowski PM, Muro EM, Perez-Iratxeta C, Andrade-Navarro MA. (2007). StemBase: a resource for the analysis of stem cell gene expression data. *Methods Mol Biol*, **407**:137-48.

Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G. (2010). Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS One,* **5**(2): e9202.

Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Séraphin B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods*, **24**(3):218-29.

Ragan C, Cloonan N, Grimmond SM, Zuker M, Ragan MA. (2009). Transcriptome-wide prediction of miRNA targets in human and mouse using FASTH. *PLoS One*, **4**(5):e5745.

Rajewsky N. (2006). microRNA target predictions in animals. *Nat Genet*, **38** Suppl:S8-13.

Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. (2004). Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**(10):1507-17.

Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, et al. (2002). Prediction of plant microRNA targets. *Cell*, **110**:513-520.

Richardt S, Lang D, Reski R, Frank W, Rensing SA. (2007). PlanTAPDB, a phylogeny-based resource of plant transcription-associated proteins. *Plant Physiol,* **143**(4):1452-66.

Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Séraphin B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, **17**(10):1030-2.

Ringwald M, Eppig JT, Begley DA, Corradi JP, McCright IJ, Hayamizu TF, Hill DP, Kadin JA, Richardson JE. (2001). The Mouse Gene Expression Database (GXD). *Nucleic Acids Res*, **29**(1):98-101.

Ritchie W, Flamant S, Rasko JE. (2009). Predicting microRNA targets and functions: traps for the unwary. *Nat Methods*, **6**(6):397-8.

Rivals I, Personnaz L, Taing L, Potier MC. (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**(4):401-7.

Robins H, Li Y, Padgett RW. (2005). Incorporating structure to predict microRNA targets. *Proc Natl Acad Sci USA*, **102**(11):4006-9.

Ruan J, Chen H, Kurgan L, Chen K, Kang C, Pu P. (2008). HuMiTar: a sequence-based method for prediction of human microRNA targets. *Algorithms Mol Biol,* 3:16.

Rusinov V, Baev V, Minkov IN, Tabler M. (2005). MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. *Nucleic Acids Res,* **33**(Web Server issue):W696-700.

Saetrom O, Snøve O Jr, Saetrom P. (2005). Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. *RNA*, 11(7):995-1003.

Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, **32**(Database issue):D91-4.

Sassen S, Miska EA, Caldas C. (2008). MicroRNA: implications for cancer. *Virchows Arch*, **452**:1-10.

Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J. (2009). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **37**:5-15.

Schulz H, Kolde R, Adler P, Aksoy I, Anastassiadis K, Bader M, Billon N, Boeuf H, Bourillot PY, Buchholz F, Dani C, Doss MX, Forrester L, Gitton M, Henrique D, Hescheler J, Himmelbauer H, Hübner N, Karantzali E, Kretsovali A, Lubitz S, Pradier L, Rai M, Reimand J, Rolletschek A, Sachinidis A, Savatier P, Stewart F, Storm MP, Trouillas M, Vilo J, Welham MJ, Winkler J, Wobus AM, Hatzopoulos AK. (2009). Functional Genomics in Embryonic Stem Cells Consortium. The FunGenES database: a genomics resource for mouse embryonic stem cell differentiation. *PLoS One*, **4**(9):e6804.

Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA. (2011). genenames.org: the HGNC resources in 2011. *Nucleic Acids Res*, **39**(Database issue):D514-9.

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

Seringhaus MR, Cayting PD, Gerstein MB. (2008). Uncovering trends in gene naming. *Genome Biol*, 9(1):401.

Sethupathy P, Corda B, Hatzigeorgiou AG: TarBase. (2006). A comprehensive database of experimentally supported animal microRNA targets. *RNA*, **12**:192-197.

Sethupathy P, Megraw M, Hatzigeorgiou AG. (2006). A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat Methods*, **3**(11):881-6.

Shahi P, Loukianiouk S, Bohne-Lang A, Kenzelmann M, Küffer S, Maertens S, Eils R, Gröne HJ, Gretz N, Brors B. (2006). Argonaute--a database for gene regulation by mammalian microRNAs. *Nucleic Acids Res*, **34**(Database issue):D115-8.

Siegel, S (1956). Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill.

Skipper M. (2008). Gene regulation: A tiny missing link for regulatory networks. *Nature Reviews Genetics*, **11**: 813-813.

Sohler F, Zimmer R (2005). Identifying active transcription factors and kinases from expression data using pathway queries. *Bioinformatics* **21**:115-122.

Sood P, Krek A, Zavolan M, Macino G, Rajewsky N (2006). Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc Natl Acad Sci USA*, **103**(8):2746-51.

Spiegel MR. (1992). Theory and Problems of Probability and Statistics. McGraw-Hill, pp. 113-114.

Stark A, Brennecke J, Russell RB, Cohen SM. (2003). Identification of Drosophila MicroRNA targets. *PLoS Biol*, **1**(3):E60.

States DJ, Ade AS, Wright ZC, Bookvich AV, Athey BD. (2009). MiSearch adaptive pubMed search tool. *Bioinformatics*, **25**(7):974-6.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, **102**:15545-15550.

Subramanian S, Steer CJ. (2010). MicroRNAs as gatekeepers of apoptosis. *J Cell Physiol*, **223**:289-298.

Suzuki HI, Yamagata K, Sugimoto K, Iwamoto T, Kato S, Miyazono K. (2009). Modulation of microRNA processing by p53. *Nature*, **460**:529-533.

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. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*, **39**(Database issue):D561-8.

Tanabe L, Wilbur WJ. (2002). Tagging gene and protein names in biomedical text. *Bioinformatics*, **18**(8):1124-32.

Thadani R, Tammi MT. (2006). MicroTar: predicting microRNA targets from RNA duplexes. *BMC Bioinformatics*, 7 Suppl 5:S20.

Thomson DW, Bracken CP, Goodall GJ. (2011) Experimental strategies for microRNA target identification. *Nucleic Acids Res*, **39**(16):6845-53.

Tschan MP, Reddy VA, Ress A, Arvidsson G, Fey MF, Torbett BE. (2008). PU.1 binding to the p53 family of tumor suppressors impairs their transcriptional activity. *Oncogene,* **27**:3489-93.

Tsuruoka Y, Tsujii J. (2004). Improving the performance of dictionary-based approaches in protein name recognition. *J Biomed Inform*, **37**(6):461-70.

Tu K, Yu H, Hua YJ, Li YY, Liu L, et al. (2009) Combinatorial network of primary and secondary microRNA-driven regulatory mechanisms. *Nucleic Acids Res*, **37**:5969-5980.

Vogel MJ, Peric-Hupkes D, van Steensel B. (2007). Detection of in vivo protein-DNA interactions using DamID in mammalian cells. *Nat Protoc*, **2**(6):1467-78.

Volinia S, Visone R, Galasso M, Rossi E, Croce CM. (2010). Identification of microRNA activity by Targets' Reverse EXpression. *Bioinformatics*, **26**(1):91-7.

von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res,* **31**(1):258-61.

Wang J, Cetindil I, Ji S, Li C, Xie X, Li G, Feng J. (2010). Interactive and fuzzy search: a dynamic way to explore MEDLINE. *Bioinformatics*, **26**(18):2321-7.

Wang X: miRDB. (2008). a microRNA target prediction and functional annotation database with a wiki interface. *RNA*, **14**:1012-1017.

Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, Zhang T, Shahab A, Yong HC, Fu Y, Weng Z, Liu J, Zhao XD, Chew JL, Lee YL, Kuznetsov VA, Sung WK, Miller LD, Lim B, Liu ET, Yu Q, Ng HH, Ruan Y. (2006). A global map of p53 transcription-factor binding sites in the human genome. *Cell*, **124**(1):207-19.

Wei JS, Song YK, Durinck S, Chen QR, Cheuk AT, et al. (2008). The MYCN oncogene is a direct target of miR-34a. *Oncogene*, **27**:5204-13.

Welch C, Chen Y, Stallings RL. (2007). MicroRNA-34a functions as a potential tumor suppressor by inducing apoptosis in neuroblastoma cells. *Oncogene*, **26**:5017-5022.

Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA. (2008). DBD--taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res*, **36**(Database issue):D88-92.

Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prüss M, Reuter I, Schacherer F. (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*, **28**(1):316-9.

Wotton SF, Blyth K, Kilbey A, Jenkins A, Terry A, Bernardin-Fried F, Friedman AD, Baxter EW, Neil JC, Cameron ER. (2004). RUNX1 transformation of

primary embryonic fibroblasts is revealed in the absence of p53. *Oncogene*, **23**(32):5476-86.

Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B. (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*, **34**(Database issue):D187-91.

Xenarios I, Salwínski L, Duan XJ, Higney P, Kim SM, Eisenberg D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, **30**(1):303-5.

Xi Y, Shalgi R, Fodstad O, Pilpel Y, Ju J. (2006). Differentially regulated micro-RNAs and actively translated messenger RNA transcripts by tumor suppressor p53 in colon cancer. *Clin Cancer Res*, **12**(7 Pt 1):2014-24.

Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. (2009). miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res*, **37**:D105-10.

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature,* **434**(7031):338-45.

Xu H, Schaniel C, Lemischka IR, Ma'ayan A. (2010). Toward a complete in silico, multi-layered embryonic stem cell regulatory network. *Wiley Interdiscip Rev Syst Biol Med*. **2**(6):708-33.

Yi M, Horton JD, Cohen JC, Hobbs HH, Stephens RM. (2006). Wholepathwayscope: a comprehensive pathway-based analysis tool for high-throughput data. *BMC Bioinformatics,* **7**:30.

Yilmaz A, Nishiyama MY Jr, Fuentes BG, Souza GM, Janies D, Gray J, Grotewold E. (2008). GRASSIUS: a platform for comparative regulatory genomics across the grasses. *Plant Physiol*, **149**(1):171-80.

Yu H, Kim T, Oh J, Ko I, Kim S, Han WS. (2010). Enabling multi-level relevance feedback on PubMed by integrating rank learning into DBMS. *BMC Bioinformatics*, 11 Suppl **2**:S6.

Yousef M, Jung S, Kossenkov AV, Showe LC, Showe MK. (2007). Naïve Bayes for microRNA target predictions--machine learning for microRNA targets. *Bioinformatics*, **23**(22):2987-92.

Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol,* **4**(4):R28.

Zhang B, Schmoyer D, Kirov S, Snoddy J. (2004). GO Tree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics,* 5:16.

Zhang H, Jin J, Tang L, Zhao Y, Gu X, Gao G, Luo J. (2010). PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Res*, **39**(Database issue):D1114-7.

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

Zhao F, Xuan Z, Liu L, Zhang MQ. (2005). TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. *Nucleic Acids Res*, **33**(Database issue):D103-7.

Zhao Y, Samal E, Srivastava D. (2005). Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature,* **436**(7048):214-20.

Zhou G, Zhang J, Su J, Shen D, Tan C. (2004). Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, **20**(7):1178-90.

Zhou R, Hang P, Zhu W, Su Z, Liang H, Du Z. (2011). Whole genome network analysis of ion channels and connexins in myocardial infarction. *Cell Physiol Biochem,* **27**(3-4):299-304.

Zhu QH, Guo AY, Gao G, Zhong YF, Xu M, Huang M, Luo J. (2007). DPTF: a database of poplar transcription factors. *Bioinformatics*, **23**(10):1307-8.

Zhu W, Yang L, Du Z. (2011). MicroRNA Regulation and Tissue-Specific Protein Interaction Network. *PLoS One,* **6**(9):e25394.

Zhu W, Yang L, Shan H, Zhang Y, Zhou R, Su Z, Du Z. (2011). MicroRNA expression analysis: clinical advantage of propranolol reveals key microRNAs in myocardial infarction. *PLoS One,* **6**(2):e14736.

Zuker M, Stiegler P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res,* **9**(1):133-48.

Acknowledgments

This thesis would not complete without expressing my sincere gratitude to those who motivated and encouraged me during my Ph.D. studies.

First of all, I am really grateful to Prof. Dr. Ralf Zimmer for giving me an opportunity to work under his supervision, for allowing me to pursue my own ideas, for his guidance, motivation and helpful discussions.

I would also like to thank Dr. Robert Küffner for his supervision, exchange of ideas and guideline for implementing them. I thank him for his time, effort and patience devoted to my work.

It is also important to mention Prof. Dr. Volker Heun for his continuous support during my work and for being my second referee. I am also thankful to all my colleagues especially Gergely Csaba and Tobias Petri for their help during this time. I will always remember Franziska Schneider for her welcoming smile and her help in handling administrative work.

Moreover, I would like to thank Prof. Dr. Hans-Werner Mewes for being part of my dissertation committee and to be my second examiner. I would also like to thank Prof. Dr. Rolf Hennicker and the other thesis committee members.

I also want to acknowledge the Deutscher Akademischer Austausch Dienst (DAAD) in collaboration with Higher Education commission of Pakistan and the Deutsche Forschungsgemeinschaft (DFG) - International Research Training Group "Regulation and Evolution of Cellular Systems" for the financial support during this work.

Finally, I would like to thank my parents and sister for always supporting me, my friends for their encouragement and suggestions.

Curriculum Vitae

PERSONAL DATA

First Name	Haroon
Family Name	Naeem
Date of birth	15.05.1981
Place of birth	Lahore, Pakistan

EDUCATION

Ludwig Maximilians Universität (LMU), Germany — 2007 - 2012
PhD - Bioinformatics research group — (PhD defense Feb 10, 2012)

Lahore University of Management Sciences (LUMS), Pakistan — 2003 - 2005
Masters of Science in Computer Science

University of Lahore (UOL), Pakistan — 1999 - 2003
Bachelors of Computer Science (Honors)

PUBLICATIONS

Naeem H, Küffner R, Zimmer R. (2011) MIRTFnet: Analysis of miRNA regulated transcription factors. *PLoS ONE* 6(8): e22519.

Naeem H, Küffner R, Csaba G, Zimmer R. (2010) miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature. *BMC Bioinformatics*, 11:1-8.

Naeem H, Zimmer R, Tavakkolkhah P, Küffner R. (2011) Rigorous assessment of gene set enrichment tests (submitted to *Bioinformatics*, revision).

CONFERENCES/WORKSHOPS

- GCB - German Conference on Bioinformatics (Freising, Germany, Sep. 2011).
- RECOMB Satellite Conference on Regulatory Genomics (New York, USA, Nov. 2010).
- ECCB 9th European Conference on Computational Biology (Ghent, Belgium, Sep. 2010).
- NGFN Plus/Transfer program of medical Genome research (Berlin, Germany, Nov. 2009).
- Next Generation Sequencing (NGS) (FU Berlin, Germany, Nov. 2009).
- Epigenetic Regulation (FU Berlin, Germany, Nov. 2009).

PRESENTATIONS

- Rigorous assessment of gene set enrichment tests paper oral presentation at the GCB 2011.
- MIRTFnet: miRNA-TF regulatory network paper poster presented at the RECOMB 2010.

- miRSel: miRNA-gene association paper poster presented at the ECCB 9th conference.

PROFESSIONAL EXPERIENCE

PhD researcher — Sep. 2007- to date
Ludwig-Maximilians-Universität, Germany

Software Engineer — Jun. 2005 - Jul. 2007
The Resource Group (Pakistan - trgworld.com)

Teaching Assistant — Sep. 2004 - Feb. 2005
Lahore University of Management Sciences

Technical Writer (Intern) — Jul. 2004 - Aug. 2004
State Bank of Pakistan

HONOURS AND AWARDS

- PhD DAAD (Deutscher Akademischer Austausch Dienst) - HEC Scholarship, 2007-2011, LMU, Germany.
- BS Computer Science Scholarship, 1999-2003, UOL, Pakistan.
- Shield of Excellence for outstanding academic performance, 2001-2002, UOL, Pakistan.