LUDWIG-MAXIMILIANS-UNIVERSITÄT
TECHNISCHE UNIVERSITÄT MÜNCHEN

computational modeling in biology

Bachelorarbeit
in Bioinformatik

Mechanisms of Transcriptional Regulation of microRNA Genes

Christoph Hamm

Aufgabensteller:	Prof. Dr. Dr. Fabian Theis
Betreuer:	Steffen Sass
	Dr. Nikola Müller
Abgabedatum:	15. März 2013

Ich versichere, dass ich diese Bachelorarbeit selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

15. März 2013 _____

Christoph Hamm

English *microRNAs* (miRNAs) are small, noncoding RNA molecules which posttranscriptionally regulate the expression of protein coding genes. They specifically suppress gene expression by binding to the mRNA of the protein. This binding leads to a lower translational efficients or a faster degradation of the mRNA.

miRNAs are transcribed from DNA by polymerase II and underlie a similar transcriptional regulation as protein coding mRNAs. Therefore its expression is also regulated by transcription factors (TF). To give insight into this regulation, a network between miRNAs and the TFs regulating them could be used.

Previously constructed networks either used experimentally data, as the correlation between TFs and miRNA expression, or experimentally determined transcription factor binding sites (TFBS) to construct the network. Also conservation between species was used.

We created a pipeline to predict a TF-miRNA network based on the genome of the examined organism. Only the genomic sequence and the genomic position is needed.

A comparison to a previously constructed network, using experimentally validated data, suggest that our network has a high quality. Therefor this network could be used to give more insight into the transcriptional regulation of miRNAs.

Deutsch *microRNAs* (miRNA) sind kurze, nicht Protein kodierende RNAs die die Expression von Genen posttranskriptional regulieren. Sie binden an deren Boten-RNA (mRNA) und hemmen dadurch die Translation oder sorgen für deren vorzeitigen Abbau. Dies führt zu einer geringeren Expression des entsprechenden Proteins.

miRNAs werden, wie die Protein kodierenden mRNAs, von Polymerase II transkribiert und unterliegen daher einer ähnlichen transkriptionallen Regulation durch Transkriptionsfaktoren (TF). Um ein besseres Verständnis dieser Regulation zu erhalten, könnte das erstellen eine Netzwerks helfen, dass alle miRNAs die TFs zuordnet, von denen diese reguliert werden.

Die bis jetzt erstellten Netzwerke basierten entweder auf experimentell erzeugten Daten, wie der Korrelation zwischen TFs und der miRNA Expression oder experimentell bestimmten Transkriptionsfaktorbindestelle (TFBS), oder dem Vergleich zwischen verschiedenen Spezies.

Wir haben eine Pipeline entwickelt mit deren Hilfe sich TF-miRNA Netzwerke allein auf Basis des Genoms des untersuchten Organismus vorhersagen lassen. Es wird nur die Gensequenz und die Positionen der bekannten miRNAs darauf benötigt.

Ein Vergleich zu einem existierenden Netzwerk mithilfe von experimentellen Daten, zeigt dass unser Netzwerk eine höhere Qualität hat. Daher könnte es tiefere Einblicke in die transkriptionale Regulation von miRNA erlauben.

Contents

1 Introduction

Since their discovery in 1993 [1] microRNAs (miRNA) have revealed a new, previously unknown level of gene regulation. miRNAs are short, non-coding RNAs, that downregulate the expression of proteins coding genes through binding to their messenger-RNA (mRNA) [2].

A search in PubMed [3] for publications about miRNA since the last year (2012) returns about 4500 results (search for all articles with "microRNA" or "miRNA" in the title). Searching in the same time frame for publication about transcription factors (TF), proteins also controlling gene expression, results in only about 2500 hits (search for "transcription factor" or "tf" in the title). This shows the large efforts currently put into the understanding of this relatively young field.

TFs are DNA binding proteins that control the transcription of DNA to mRNA [4]. They can increase or decrease the rate of transcription, which influences the number of RNA present in the cell. This RNA can be mRNA, in which case the expressions of proteins is influenced, but also other, not coding RNAs are controlled by this mechanism. Therefor also the levels miRNA can be changed by TFs.

A current work of Yu et al. revealed the influence of miRNA and its transcriptional regulation in human cancer [5]. Therefore the construction of better TF-miRNA networks seems to promise further insight into cancer and other human diseases, caused by the deregulation of the cell.

Currently available networks are based either on experimental data [6, 7] or use the conservation among different species to be able to search for large regions for transcription factor binding sites [8] without getting too much false hits. As for most species, except human, only few experimental data are available, this networks cannot be used for them. The use of conservation restricts the miRNAs to the ones present in the examined species.

Here we describe a pipeline that is able to predict TF-miRNA networks based only on the genomic sequence, the genomic position of known miRNAs and information about TFs of the examined organism. We applied methods for the search of transcription factor binding sites (TFBS) on the promoter regions of the transcripts, from which the miRNA is processed. To detect these promoters we use promoter prediction tools.

2 Background

2.1 What is microRNA

microRNAs (miRNA) are small, 21-25-length, RNAs which downregulate the expression of protein coding genes [2]. This down regulation is achieved through a binding to the 3' untranslated region (3'-UTR) of mRNA of the controlled gene.

The maturing of miRNAs starts with longer RNA called primary microRNA (pre-miRNA). This pri-miRNA may contain several miRNAs. pri-miRNAs itself origin from own transcripts or introns of protein coding genes. Own transcripts are, as the protein

coding mRNAs, transcribed by polymerase II [9]. The distance between the start of this transcript and the first pre-miRNA varies between a few bp up to several kp [9, 10].

From this pri-miRNA small stem loops are cut out by Drosha, a RNA cleaving protein [11]. The results are RNAs with a length about 70 nucleotides. Dicer cuts this precursor miRNAs (pre-miRNA) to the final mature miRNAs[12]. Together with several protein this mature miRNAs form a RNA-induced silencing complex (RISC) [13].

This complex binds the 3'-UTR of the target mRNA. The binding is caused by a match of the RISCs miRNA to the mRNA [1]. The reason for the silencing could be explained through a slower translation or a degradation of the mRNA. Guo et al. showed that the main reason for mRNA suppression is the destabilisation of the mRNA decrease [14]

As the target of the miRNA is bound through base pairing too the mRNAs, this could be used to predict the targets of a miRNAs. For this an exact pairing to the seed, a seven nucleotides long part of the mature miRNA, is searched in the 3'-UTRs of all mRNAs. An example for such a method is *TargetScan* [15], which uses conservation between species and is so able to relax requirements for the base pairing to a six nucleotides long seed.

2.1.1 Transcriptional Regulation

Transcriptional regulation describes the regulation on the level of transcripts through the initiation of transcription. This could reach from a change of the transcript number to the activation or complete deactivation of the regulated transcript.

The level of transcription is influenced through transcription factors (TF) [16]. TFs are DNA binding proteins that change the activity of polymerase. Some a part of the initiator complex. This is a complex of proteins binding to the DNA which initiates the binding of polymerase II and is therefor necessary for transcription to take place [17]. Other TFs block the formation of the polymerase complex, which leads to a down regulation of the transcription [7]. Also a methylation or demethylation of a histone through the TF is possible. This changes the DNA folding and high-level chromatin folding is known for suppressing the transcription [18].

There are two types of regions were TF binding change the rate of transcription, the promoter and enhancer. The promoter is a region near the transcriptional start site (TSS) and TF binding here influence directly the binding of the polymerase [19]. The second type are enhancers. Enhancers may be several kb up- or downstream. It is able to change the rate of expression as it is physical near the promoter through the folding of the DNA [19]. Through this low physical distance to the TSS the TF binding can act on the formation of the polymerase.

In the following we only are interested in TFs binding to the promoter. Enhancers are ignored as it is difficult the find them, for their distance to the controlled gene on the genome. How a TF influence the transcription and if it is a suppressor or a activators is ignored.

2.2 Databases

2.2.1 miRBase

miRBase is a database for known miRNAs [20, 21]. Mature sequences are annotated with the genomic position of the particular pre-miRNA. The current version (19) contains 1600 precursors for human and 855 for mouse.

miRNAs need to be supported experimentally, a prediction alone is not sufficient, to get annotated. Usually only the sequence of the mature miRNA can be determined. So most of the pre-miRNA sequences stored in the database are based on hairpin structure predicted around the detected mature sequence.

miRBase also provides names to newly detected miRNAs. They can be requested after the publication of the paper supporting this miRNA is accepted.

Since $40 - 70\%$ of the miRNAs are not cut from own transcripts, but from introns of protein coding genes [22], known overlapping transcripts are also annotated. These transcripts are stored in the database even if they are on the opposite strand. Information about the direction and the type of the overlap are also stored in the database. The type of the overlap can be "exon", "intron", the "5'-" or the "3'-UTR".

As it is possible that several pre-miRNAs are processed from one pri-miRNA [9], miRBase contains additional information for clusters of miRNAs. A cluster is defined as set of miRNAs with a distance lower than 10 kb. The strand is ignored and therefore miRNAs on the opposite strand are also annotated to be part of the cluster.

2.2.2 Ensembl

Annotation We used *Ensembl* [23] as resource for the genome annotation. The basis for their annotation is a automatic pipeline using experimental data like cDNA and manual annotation from the *Havana*[24] project.

The "Ensembl Automatic Gene Annotation System" uses primarily cDNA data and known protein sequences to search for the chromosomal position of the corresponding gene. After the position is found a alignment is used to determine the exact positions of exons and introns. Additionally Expressed sequence tags (EST) are used. However, this kind of data is less reliable [25].

The second resource for data is the Havana project . The Havana database contains manual annotation for vertebrate genomes. All genes need evidence from ESTs, cDNA or protein sequence to be annotated.

BioMart One possibility to access the data of *Ensembl* is *BioMart* [26]. It provides an easy to use interface for data retrieval from the *Ensembl* databases, including *Ensembl Genes*, which was used here. The attributes of interest can easily be retrieved and a filter can be applied to restrict for certain entries. These can for example be a certain region or only transcripts belonging to a certain gene.

For the *R* programming language[27] the *biomaRt* [28, 29] package, provided by Bioconductor, can be used as a simple interface. Queries to the databases are returned as R *data.frames* and can directly be used for further analysis.

2.3 Existing Networks

2.3.1 TransmiR

Wang et al. published a TF-miRNA network [6] based on experimental validated interactions obtained from literature. They searched PubMed for papers containing the keywords 'microRNA' or 'miRNA' and evaluated them. As result they obtained 243 TF-miRNA interactions between 82 TFs and 100 miRNAs from 83 papers.

The current version (TransmiR v1.2, 2013 February 19) contains 745 interaction for 17 different organisms. The vast majority of the stored interactions are for human (649). Even for the important model organisms mouse and C. elegans just a few interactions are annotated (31 and 37).

As this network is literature-based, it has a high quality and should contain only few false-positive interactions.

They also annotated the type of interaction, which can be activation or suppression of the respective miRNA, as well as the PubMed ID of the paper from which the knowledge of the interaction was extracted. Additionally the network contains information about diseases associated and biological functions for each miRNA.

The drawback of this approach is the low number of interactions that were validated experimentally. So for the 1600 known miRNAs in human (miRBase release 19), only 175 are contained in the network. The most connected miRNA is regulated by 24 TFs and only 15 have more than 10 connections.

Also the number of miRNAs differ between the different TFs. The highest number of miRNAs regulated by one TF (MYC) is 45. This also shows that the network is biased. Deregulation of c-Myc is associated with human cancer and therefore a promising target in the development of new drugs [30]. A search for "myc[Title]" in PubMed (2013, February the 19) results in 8105 in hits. On the other hand "foxp3[Title]", a TF important for the development of T-cells [31], interacts with only 4 miRNA genes according to *TransmiR*. A PubMed search for this TF returns only 2147 papers. So it is likely that the more one TF is in the focus of interest, the more connections will be found and the higher the degree will be in literature based networks.

On the other hand, Wang et al. could show the conservation of one miRNA correlates with the number of associated TFs based on this network [6]. They split the miRNAs in a group of high degree and low degree, where each miRNA regulated by 3 or more TFs was considered to be of high degree. miRNAs conserved in vertebrates and more distant species were labelled as conserved, all others as non-conserved. The use of Fisher's exact test on these data returned a p-value of $p = 0.02$. This shows that conserved miRNAs have interactions with more TFs than expected by change.

On the one hand this result could indicate that the number of interactions is not only influenced by the interest in one particular miRNA or TF. On the other hand it could be an artifact from conserved miRNAs playing a more central role in cell regulation, which leads to more available information.

2.3.2 PuTmiR

A other approach for the construction of a TF-miRNA network was applied by Bandyopadhyay and Bhattacharyya [8]. They used conserved TFBS determined by the use of PWMs to find putative TF-miRNA interactions in human. Binding sites in both directions, downstream and upstream of the miRNA, are considered to play a role in the transcriptional regulation.

They retrieved the positions for the pre-miRNAs from miRBase, while miRNAs which are not on the University of California Santa Cruz (UCSC) assembly [32] were ignored. Especially miRNAs on the mitochondrial DNA and chromosomes, which are not part of the standard assembly (like HSCHR6_MHC_COX), were excluded. For the remaining miRNAs the regulatory sequences were defined as the 10 kb upstream region (USR) and the 10 kb downstream region (DSR).

In order to get TF binding in these regions, TFBS from the UCSC Table Browser were extracted by downloading data of conserved sites and conserved TFs from the TFBS conserved track of the human genome. The combination of this two data sets with the previously defined regulating regions results in two lists of TF-miRNA interactions. One for the USR and one for the DSR. As this list contains also the distance of the TFBS to the miRNA, it is possible to get only TFs in a smaller USR or DSR as the predefined 10 kb. It is also possible to use only one of the resulting lists, and therefore only TFs in the USR or DSR of the miRNA.

The conserved TFBS in the UCSC database are defined as the positions for which a common binding site is predicted in the three species human, mouse and rat. The matrices are taken from the public part of the TRANSFAC database [33]. If the score from one PWM at a certain position exceed a threshold for this species it is considered as a putative TFBS. The threshold is defined using the z-score of this matrix in this species as being 1.64 standard deviations over the mean.

The idea behind the use of only predictions conserved in all three species is that TFBS with a biological role are more likely to change only slightly between the species as cause of the evolutionary pressure. However, regions for which PWM hits are random should change between the species by mutation, as they are under no constraints.

The advantages of using TFBS prediction over searching the literature for known interactions is the higher number of detected interactions and that there is no bias for TF or miRNA being more in the focus of interest. However, the match of a PWM does not always predict a real influence on the miRNA as it could appear randomly. This is unlikely by using only conserved binding sites, but it is still possible that this function disappeared in human and the found position is just a relict of a previous existing regulation.

A regulatory network can be defined as bipartite graph with TFs on the one and miRNAs on the other side, where an edge symbolises an interaction. In theses graphs the degree of the miRNAs usually follows a power law distribution [8]. For this most of the miRNAs are regulated by only a few TFs and the number of miRNAs regulated by many TFs is low. Also they found more TFBS in the DSR as in the USR.

From the 721 extracted pre-miRNA from miRBase, TF interactions are detected for

666. The current version of miRBase contains only 855 and 466 pre-miRNAs for Mus musculus and Rattus norvegicus, where only 218 orthologous miRNAs exist in all three species (only the number of the miRNA was used for this comparison, so hsa-miR-548a-1 and hsa-miR-548i-3 would be considered as the same what also reduced the number of miRNAs in the network to 490, which is still twice as much). The conservation of a TFBS indicate that regulated element still exists in all three species, otherwise they would disappear by random mutation. Therefore this 666 miRNA should also be found in mouse and rat.

This leads to two explanations.

1. The miRNAs are conserved in all three species and some are simply not annotated in miRBase

2. The conserved TFBS found for miRNAs not conserved in all three species regulate other genes or miRNA. This seems likely because of the large size of the searched sequence. This also leads to the conclusion that some TFBS for the conserved miRNA may not regulate them.

2.3.3 ChIPBase

An other method to get information about gene regulation is the use of chromatin immunoprecipitation (ChIP) based techniques [7]. ChIP uses antibodies to detect DNA binding site of a certain protein. By using antibodies against TFs it becomes possible to get *in vivo* information about TFBS.

The goal of chromatin immunoprecipitation is to detect the binding sites of DNA binding proteins. For this purpose, snippets of DNA binding the protein of interest are purified and examined. In the case of ChIP-Seq, next generation sequencing methods are applied to get the position of the bindings.

The first step in ChIP is the in vivo crosslinking of the DNA binding proteins to the DNA by the use of *formaldehyde*. After extracting the chromatin from the cells, it is fragmented by sonication to pieces of a few hundred base pairs length. Antibodies against the examined protein are used to enrich for DNA fragments that binds these proteins. After the cross-links are removed, the DNA is isolated and can be examined [34]. One method is to sequence this DNA snippets and map the results to the genomic sequence. In this way the location of the protein binding sites can be determined.

Using antibodies against a TF, the resulting mapping shows TFBS for this TF. It has to be noted that this experiments do not reveal all TF bindings occurring in nature. As transcription depends on the cell type and the growth conditions, not all possible TFBS will be detected by one single experiment. Experiments with different conditions and cell types are therefor needed to find all existing bindings.

Using data from ChIP-Seq experiments, Yang et al. constructed a network for the transcriptional regulation of large non-coding RNAs (lncRNA) and miRNAs [35]. As we are interested in TF-miRNA interactions, in the following only the part about miRNA will be discussed and lncRNAs will be ignored.

As for PuTmiR, the starting point was the miRNA annotation of miRBase. Intronic pre-miRNAs were clustered into groups that are supposed to be transcribed together. The TSS from this cluster was supposed to be the TSS of its first pre-miRNA. For intronic pre-miRNAs, the TSS of the host gene was used.

The regulating domain was defined as the region 30 kb up- and 10 kp downstream of the TSS. This large region was chosen because the distance between the pre-miRNA can vary from a few tens of bp upstream, as 59 bp for miR-23a∼27a∼24-2 [9], and several kb upstream, as more than 30 kb for miR-34a [10].

Every TFBS that intersects with this part of the sequence was considered as TF regulating the miRNAs of the cluster or the host gene. These TFBS were extracted from 543 ChIP-Seq peak data sets, including ENCODE [36] and modENCODE [37, 38]. This data sets contained information for 252 TFs.

Data can be downloaded from http://deepbase.sysu.edu.cn/chipbase/ for six species, including human, mouse, D. melanogaster and C. elegans. The size of the regulating region can be set from 30 to 5 kb up- and from 10 to 1 kb downstream. As the number of available ChIP-Seq experiments varies between species, for example 329 for human but only 37 for C. elegans, also the number of available TF varies. For human 119 and for C. elegans 26 TFs are part of the network.

ChIP-Base is not limited by the number of experiments in the same amount as TransmiR. It also needs an experiment for every TF and cell condition, but for every miRNA it is tested if a TFBS is in its regulating region. Also the number of falsely assigned binding sites should be better, as they are validated by experiments. Also TFBS not conserved in different species are found, in contrast to PuTmiR, as only information of the investigated species is needed.

But it has similar problems as PuTmiR. Only binding sites of TFs in the neighbourhood of the miRNA are detected. No information is available if this TF changes the expression of this miRNA and what the type of regulation, suppression or activation, it is.

An other problem is the availability of data sets needed for generating the network. For every TF experimental data for different conditions and cell types should be available to construct the network. Currently it seems as only for human the number of available ChIP-Seq is high enough, through the work of ENCODE.

2.4 Promoter and TSS Prediction

2.4.1 miRStart

MiRStart [39] is a database of predicted TSS for human miRNAs. It was published by Chien et al. in 2011. The idea was to use not only the sequence in the upstream region of the miRNA, but to use also data from high-throughput experiments, in order to find the TSS. Therefore cap analysis gene expression tag (CAGE tag), TSS Seq libraries and H3K4me3 data were used.

This also shows the drawback of the approach. Large amount of additional data are needed. CAGE tag data is available for human and mouse from the FANTOM project,

but data may lack for other model organisms.

CAGE tag[40] and *TSS seq*[41] are both techniques to get the position of TSS from capped proteins. The idea is to sequence full length cDNA. Full length cDNA contains all bases of the RNA from which it was transcribed. Therefore mapping of this sequences to their genomic position reveal the real start of the transcript and so the TSS.

To isolate the cDNA in both cases the CAP of the RNA is used. As pri-miRNAs are transcribed by polymerase II they are capped, as mRNAs. Therefore this methods also reveals the TSS of pri-miRNAs.

The Functional Annotation of the Mammalian Genome 4 (FANTOM) project produced 29 million CAGE tags for 127 human RNA samples from the THP-1 cell line [42]. The goal of this project was to determine the transcription of THP-1 cells under different condition, like stimulation with phorbol myristate acetate (PMA) and the knockdown of 52 TF. Especially deepCAGE tag sequencing was used to determine the TSS under different conditions.

The TSS seq data originate from DBTSS [43]. DBTSS contains TSS verified by TSS seq. From this database MiRStart uses 316801241 sequences from eight human tissues and six cell lines.

The second type of data used were histone modifications, as they are known to influence the transcription of genes. This could give hints on the position of genes and therefore the start of their transcript. As the trimethylation of lysine 4 from the H3 histone (H3K4me3) is associated with the position of the TSS [44], data of this modification was used for the construction of this database.

The data from the histone modification was extracted from experiments performed by Barski et. al [44]. They determined positions of several kinds of histone methylation in $CD4^+$ T cells using ChIP-Seq. The method is the same as described above for the identification if TFBS. The difference is the use of antibodies against modified histones instead of TFs. The mapping of the sequence in this case shows which part of the genome is methylated in the investigated manner. For H3K4me3 the experiments returned 11.3 million tags.

To predict TSS from this data, a support vector machine (SVM) was used. In Order to train a SVM a positive and a negative test set are necessary. For the positive set 7286 protein coding genes with experimentally verified TSS were used. Only genes with just one TSS were used. The negative set was constructed using sequences 10 kb up- and downstream of this TSSs.

To test the performance of the TSS, cross-validation was applied. According to this tests, the SVM determined the TSS of the protein coding genes with a sensitivity, specificity, accuracy and precision of about 90% in each case. The SVM was then used on the 50 kb upstream region of the pre-miRNAs to predict their TSS.

2.4.2 Eponine

Eponine [45] is a prediction tool for TSS of protein coding, eukaryotic genes. It uses PWMs to model certain features near the start site like the TATA-Box or CpG-Islands. As the distance of this features can vary, a simple big PWM cannot be applied. For this

reason the distribution of the distance TSS to feature is used to check if the features are in the correct position.

The training used a mixture of a relevance vector machine and Monte Carlo sampling. Feature like TATA-Boxes or CpG-Islands were not defined a priori, but training on known promoter sequences revealed PWMs recognising this features.

2.4.3 Promoter2

Promoter2 [46] is a program to predict promoters using neuronal network (NN). A set of NNs is run over the examined sequence. Each NN takes a window of 6 nucleotides and, for all but the first one, results from the previous networks as input.

The NNs are trained using a simple genetic algorithm. The training set consisted of 200 bp length sequence, 100 promoter and 100 control. For the training the quality of the NNs was rated by their ability to distinguish between this two sets.

2.4.4 Ep3

In difference to the methods described above, *Ep3* [47] does not need any training data and therefor no set of experimentally found promoters. It is based on DNA properties as GC content, denaturation values or duplex free energy, that are assigned using tables for each di- or trinuleotide. This tables on the other hand are created using experimental data, but are universal for different species.

The prediction is based on windows, for which the average of each property is calculated. Based on the overall distribution of this values promoters a predicted for all windows exceeding a certain z-score. To get this overall distribution the whole genome sequence is needed and therefor the program is applied on the whole genome. The results are a list of all windows that were classified as putative promoters.

2.4.5 PromoterInspector

PromoterInspector [48] also predicts promoter regions. This method is based on IUPAC strings containing wild cards ("N"). Using the IUPAC strings classifiers are build, distinguishing between promoter and non-promoter regions. For constructing these, a training set is used. Three classifiers are used for which the non-promoters parts are "exon", "intron" or "3'-UTR". Only if all three assign the sequence to the class "promoter" it is considered as such.

To run the program on a large sequence a sliding window approach is used. Each window is presented to the classifiers and if the number of "promoter windows" in on region is exceeded it is considered as promoter region.

2.5 TFBS search

2.5.1 PWM

To find TFs regulating a gene, binding sites in its promoter are searched. As not all sequences a TFs binds are exactly the same, the simple search for a sequence is not possible. Also allowing several mismatches does not solve the problem. If the number of allowed mismatches is to low, too less binding sites are found. If it is to high, to many sites are predicted that do not bind the examined TF [49].

The common solution for this problem is the use of position weight matrices (PWM). The idea is to use a score for each position and possible nucleotide. To test if a sequence matches the PWM, for each position the score of the nucleotide is read from matrix. The overall score is determined by adding this values. If this score exceeds a certain threshold it is counted as match.

At a conserved site the score for the consensus nucleotide is high and for all other nucleotides are low. Therefor a non consensus nucleotide in such a position would lead to a drastic lower overall score. In a less conserved position, on the other hand, the score for all nucleotides a nearly the same. As a result a certain nucleotide at such a position has only little influence on the score of the whole match.

To search such PWM matches and calculate their score the *Biostrings package* [50] can be used. It contains function to calculate these score matrices from count matrices. Based on this PWM matches can be search on sequences and their score can be calculated.

2.5.2 Transfac

To use PWMs, as described above, count matrices are needed. Transfac [33] is a database containing matrices for TFs of different species. Its basis is a collection of DNA sites and TFs that bind to them. The sites need to be experimentally verified to be stored in the database.

Matrices for the TFs are calculated based on these sites. The matrices are stored separately. For each TF the corresponding matrices are annotated. It is possible that no matrix exists, but also different matrices could be assigned to one TF.

2.5.3 ModelInspector

The method of using PWMs to predict TF described above has the problems that most of its predicted site will not be functionally, unless the score is really high [16].

To improve the prediction, modules of TF can be used instead of single ones. The idea behind this approach is that TFs interact and therefor certain TFs appear with a higher probability in a certain distance [16].

The combination of *ModelGenerator* and *ModelInspector* [51] can be used for this type of analysis. *ModelGenerator* create models consisting of different elements in a certain order. The elements could be matrix matches, hairpins, repeats or several other DNA properties. As TFs are search here, matrix matches are from special interest. Based on

a set of the elements with their order and a set of training sequences *ModelGenerator* generates a model. This model contains information about the distances of the elements and which are necessary to identify the model.

ModelInspector is able to search this model on sequences. Using it with model for TF modules it can search promoters for TFs binding them.

2.6 miRNA targeting Signalling Pathways

Normally one miRNA has many different target mRNAs [15]. Therefor a miRNA can have more than one target in a signalling pathway. This would lead to the conclusion that it has a influence on this pathway. To find such pathway for a given set of miRNA Kowarsch et. all developed *miTALOS* [52]. As basis for the pathways the *Kyoto Encyclopedia of Genes and Genomes* (KEGG) [53] and the *National Cancer Institute Pathway Interaction Database* (NCI PID) [54] are used.

miTALOS uses different miRNA target prediction tools to identify signalling proteins which expression is probably influenced. For each pathway a *enrichment score* and a *proximate score* are calculated. The *enrichment score* is the relation between the signalling proteins hit by the miRNA to the number expected by ration of proteins targeted by this miRNA. The *proximate score* describes distance between the influenced signalling proteins. The closer they are, the more probably is a effect on this pathway.

Additionally p-values are calculated for the two properties described by the scores. Based on these p-values and scores relevant pathways are selected. In the default options a *Proximate score* over 1 and a p-value below 0.05 are used as threshold.

3 Methods

To construct the TF-miRNA network, first a list of all known pre-miRNAs with their genomic position is need. From the position we determine the host gene of intronic miRNAs and predicted the TSS for exonic miRNAs. After this step, the TSS for each miRNA is determined and based on this information the promoter region is predicted. In a final step TFBS are searched in this promoter, in order to find TFs regulating the corresponding miRNA (see Figure 1).

3.1 Extraction of miRNA positions

3.1.1 miRBase

We extracted a list of all known miRNA ids for the examined species using the *mirbase.db* package [55]. First we examined the genomic context as annotated in *miRBase*. As the received list contains overlapping genes from both strands, transcripts being on the opposite strand are omitted. They are marked by a '−' in the *contextOverlapSense* field.

For all transcripts, ignoring the *transcript_source*, we looked up the gene position from *Ensembl* using *biomaRt*. It is possible to look up all entries as transcript ids cannot be

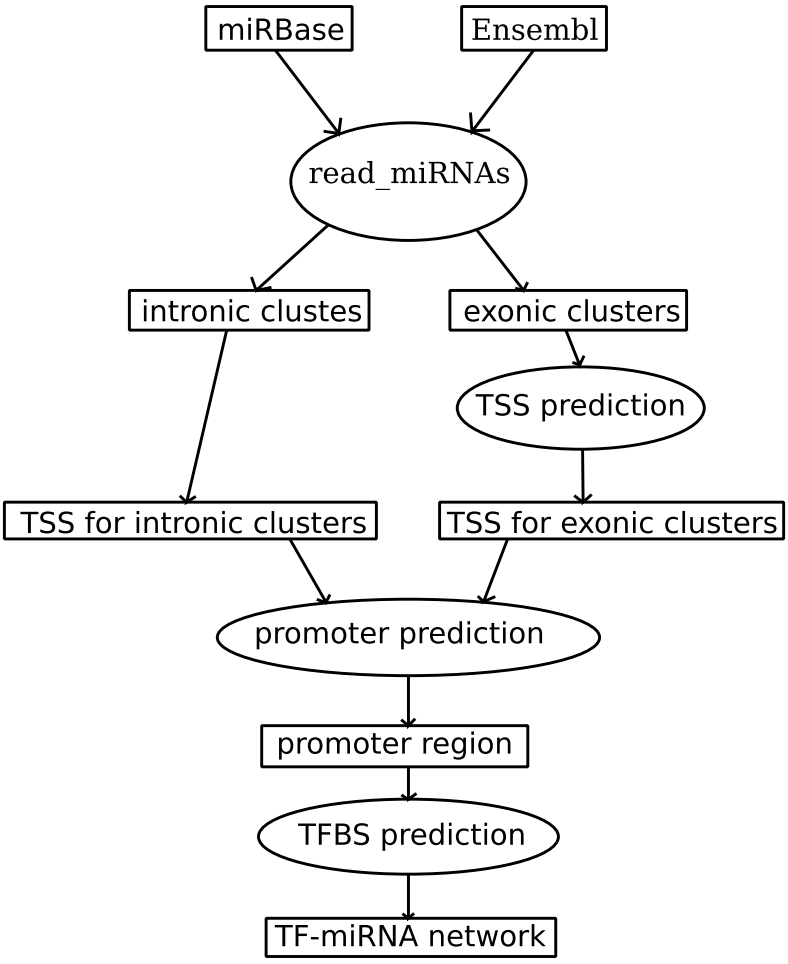


Figure 1: General flow of the pipeline. miRNAs are read from miRBase and grouped into clusters. A cluster is a set of miRNAs that is expected to be transcribed together on one pri-miRNA. A intronic cluster is part of another gene while an exonic cluster is transcribed on his one. As for the exonic ones no TSS are available, we tried to predict them. Based on the TSS of a cluster, the promoter regions is searched. On this region finally TFBS a predicted, in order to receive the TF-miRNA network.

confused between the databases. miRNAs for which a gene entry is found in *ensembl* are considered to be intronic and the start position of the gene is used as TSS.

For all transcripts having a context entry but no data is received from *BioMart*, we searched overlapping genes from ensemble. In human, it was necessary for twelve miRNAs. We found gens for eleven of them. These are also considered as intronic. pre-miRNAs being part of the same host gene are grouped into one clusters as they should be co-expressed.

The remaining pre-miRNAs are marked as exonic. As it is known that one pri-miRNA may contain several pre-miRNAs, we group pre-miRNAs into clusters according to their genomic position. The *CLUSTER* entry from *miRBase* is used for this purpose, where all pre-miRNAs being less then 10 kb apart are defined as one cluster. This size seems to be reasonable for miRNA clusters [56]. As the context entry contains miRNAs on both strands, the list has to be filtered first for miRNAs lying on the opposite strand. The TSS of each cluster is defined as its first pre-miRNA.

3.1.2 miRNA annotation from Ensembl

An other method used for receiving pre-miRNA positions is to extract data directly from *Ensembl*. For that, we receive information about the position of the gene, the *mirbase_id* and the *hgnc_symbol* for all data sets for which *with_mirbase* is true using *BioMart*.

To distinguish between intronic and exonic miRNAs, the *hgnc_symbol* was used. All entries not starting with "MIR" or without *hgnc_symbol* were considered as intronic, the remaining as exonic. We marked all miRNA having at least one intronic entry as intronic.

The TSS of each miRNA was defined by the first start of all its transcripts. We simply defined a cluster as set of miRNAs having the same TSS, which could result from lying in the same host gene or being part of a pri-miRNA annotated in *Ensembl*.

3.2 Transcription Start Site search

As the distance of the real TSS to the pre-miRNA can vary in wide range, from a few bases up the several kb [9, 10], it seems reasonable to search for the TSS. As a database is only available for human, but networks should also be created for other species, the TSS could only be predicted. As the TSS of intronic miRNAs are assumed to be the same as for the host gene, which is known, this prediction is only needed for exonic miRNAs.

3.2.1 Eponine

For the prediction of the TSS we used *Eponine*. For each cluster we extracted the 100 kb up- and downstream DNA sequence. We applied the Eponine program on these sequences. Eponine returns for each prediction a score between 0 and 1, where 1 indicates

a completely certain prediction. Only results with a score higher than 0.95 were used. As Eponine predicts start sites on both strand, results for the opposite strand are filtered.

Eponine predicts about 47 TSS per examined sequence. Because the use of this high number of regulating sequences for each miRNA to predict TFBS would lead too a high number of false-positives, some TSS need to be selected. For this purpose, we calculated the density of all prediction for this sequence. All maxima of the density before the cluster reaching the half of the highest peak are used as TSS. If no predictions are available, the we use the start of the cluster as TSS.

3.3 Promoter search

To search for TFBS, a region which should be examined is needed. We used promoter predictions tools to search promoters, which we later used to search for this TFBS.

3.3.1 Promoter2

The first tool we tested was Promotor 2, using the public available *Promoter 2.0 Prediction Server*. As only 50 sequences can be submitted at once, we only tested this number.

For this, we selected 50 cluster. It was only mind that the cluster contain the miRNAs 23a∼27a∼24-2, as their TSS was determined experimentally. For this 50 cluster the 50 kb upstream and 500 bp downstream region was saved as a *fasta-file* and uploaded to the server.

We converted the resulting position of promoter 2 to their genomic position. All results were saved with their score.

3.3.2 Ep3

The second promoter predictor used, was *Ep3*. Ep3 needs to be run on the whole genome. The regions in which promoters are searched are the same as for *Promotor2*. From the *gff-file* created by *Ep3* we extracted all results overlapping this region. As *Ep3* predicts TSS independently of their direction no filtering for the strand is necessary.

3.3.3 PromotorInspector

PromotorInspector is a promoter prediction tool that is part of the *Genomatix Software Suite*. We used the region 9000 bp up- and 500 bp downstream of the TSSs. In order to retrieve the sequences, the positions are save as *BED-file*, uploaded to Genomatix and converted to DNA sequence.

We run *PromoterInspector* on these sequences. Afterwards we downloaded the result as *fasta-file*. From the description lines of this files the positions and the strand of the predictions is extracted. We used only results on the positive strand. As the prediction was run on extracted sequences, the positions have to be converted back to the genomic position.

For all sequences without a prediction, we used the region 1000 bp up- and 500 bp downstream as promoter. This length was selected for Cheng et al. showed that TF binding is enriched in this region for C. elegans [57].

3.4 TFBS search

To retrieve the TFs regulating the miRNAs we searched the previous defined promoters for TFBS. We saved these regions as *bed-file* and uploaded them to *Genomatix*. This regions were converted to DNA sequences and downloaded as *fasta-file*

3.4.1 PWM from TRANSFAC

To predict TFBS based on PWMs, we first needed count matrices. We extracted these matrices of the examined species from *TRANSFAC*. The *PWM* function of the *Biostrings* package does not work on all these matrices, as sum of entries for each position varies for most of the matrices. Also some matrices (e.g. "M01152") obviously do not contain counts as their entries are no integers.

First we converted the matrix a stochastic matrix, in which each entry represents the probability for this position of containing the corresponding base. The entries for each position sum up to 1. Afterwards, this matrix is normalised using the *unitScale* function from *Biostrings*. After this step the lowest possible match score of this matrix is 0 and the highest 1.

Using this PWMs, we searched for each promoter sequence matches with a score higher than 0.8. All matches are saved with their score. Using information from *Transfac*, we mapped the PWMs to their corresponding TFs. In the same step we saved files for different score thresholds between 0.8 and 1. After the mapping of the clusters to their miRNA, the TF-miRNA network for the different thresholds are complete.

3.4.2 ModelInspector

As second method to get TFBS we used *ModelInspector*. As the number of results is limited to 20000, we split the *bed-file* to parts containing maximal 300 entries. Each *bed-file* is converted to a DNA sequence and used as input for *ModelInspector*. Additionally, the sequences are downloaded as *fasta-files*.

We saved the results as *tsv-file*. These files do not contain the name of the sequence as specified in the *bed-file*, but a name for each sequence of the from 'Region_1'. The association of this region number to the cluster number is read from the identifier lines of the corresponding fasta file and the region numbers are exchanged by the cluster numbers.

The result of *ModelInspector* does not contain single TFs, but models of two or more TF families. We extracted these families from the model names by simply splitting it at the '_' signs, ignoring the last part as it only contains a number. As by the use of PWM from *Transfac* described above, in a last step we converted the clusters to their miRNAs to get a TFfam-miRNA network.

4 Result

We used our pipeline to predict a TF-miRNA network for human. We used this organism as the most data is available for it. *miRStart* only list TSS for human miRNAs and *ChIPBase* and *TransmiR* have most of its data for human. Based on our prediction and the available data we evaluated the different methods and created our pipeline. Of course the pipeline can be used for all organisms for which a genomic sequence, the genomic position for miRNAs and information about TFs is available.

4.1 TSS search

From the 1600 miRNA annotated in *miRBase* for human, 658 are considered as exonic. As no TSSs are available for these miRNAs from *miRBase* other sources are needed.

4.1.1 Ensembl annotation

Using *biomaRt* to retrieve data from *Ensembl* returned 2236 transcripts for 1426 miRNAs for human. So information on 174 miRNAs annotated in *miRBase* is missing.

Comparison of the position of this transcript to the position of the pre-miRNA as annotated in *miRBase* shows that for 815 they are not the same. As this transcript belongs to 186 different miRNAs, only for this number additional information can be get from *Ensembl*.

From this 815 transcripts, only 62 overlap the respective pre-miRNA, leaving 753 ones for which it is a part of the transcript. This only partial overlap can be caused either by a transcript of an other gene overlapping the pre-miRNA or by differences in the annotated data (see Figure 2).

After removing all transcripts belonging to other genes 101 one are left. The type of the gene is determined by the *HGNC Symbol*, where all symbols not starting with "MIR" are removed. This 101 give additional information on 59 miRNAs.

As this corresponds to only about 10% of all exonic miRNAs, it is not used and we tried to predict the TSS from the upstream sequence of the respective miRNA cluster.

4.1.2 Comparison miRStart and Eponine prediction

As we could not retrieve the information about the TSS of the miRNAs from *Ensembl* we needed to predict them. For this prediction *Eponine* was used.

Eponine returned 26344 TSS for the examined regions around the exonic miRNAs of human. For the 570 exonic clusters, TSS with the required score are found for 557, leaving 13 clusters without prediction. The average number of TSS for a cluster is therefore 47.3.

For 529 cluster peaks are found by the method as described above. In total, 605 TSS are detected in this way. More then one TSS was determined for 66 clusters, were the highest number of TSS per cluster is 4.

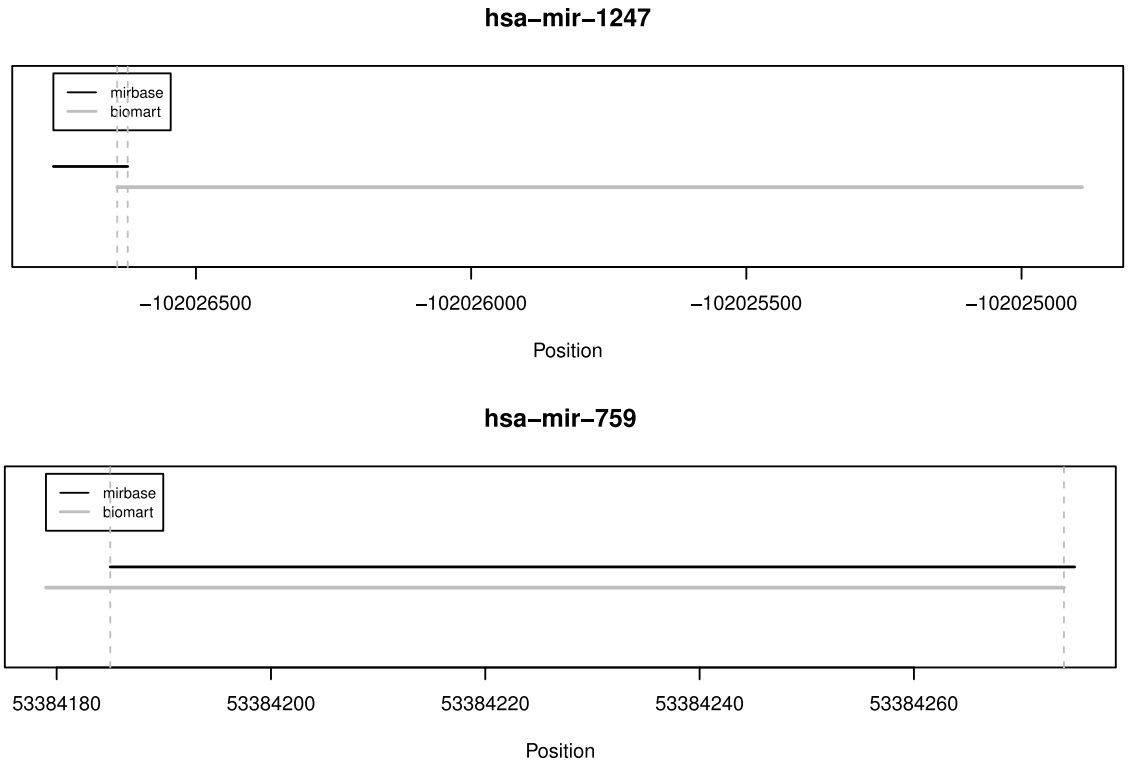


Figure 2: Different kinds of pre-miRNA only partial overlapping transcripts from Ensembl. *hsa-mir-1247* is overlapping DIO3OS but only partial, but is annotated in *Ensembl* for this miRNA. In the case of *hsa-mir-759* the transcript is the miRNA, as the HGNC_Symbol "MIR759" indicates, but the position differs between the two databases.

The distribution of the *Eponine* prediction over the used region is heterogeneous (see Figure 3). Some clear peaks in front of the cluster are possible, but also homogeneous distributions over a wide region of more then 100 kb can be found.

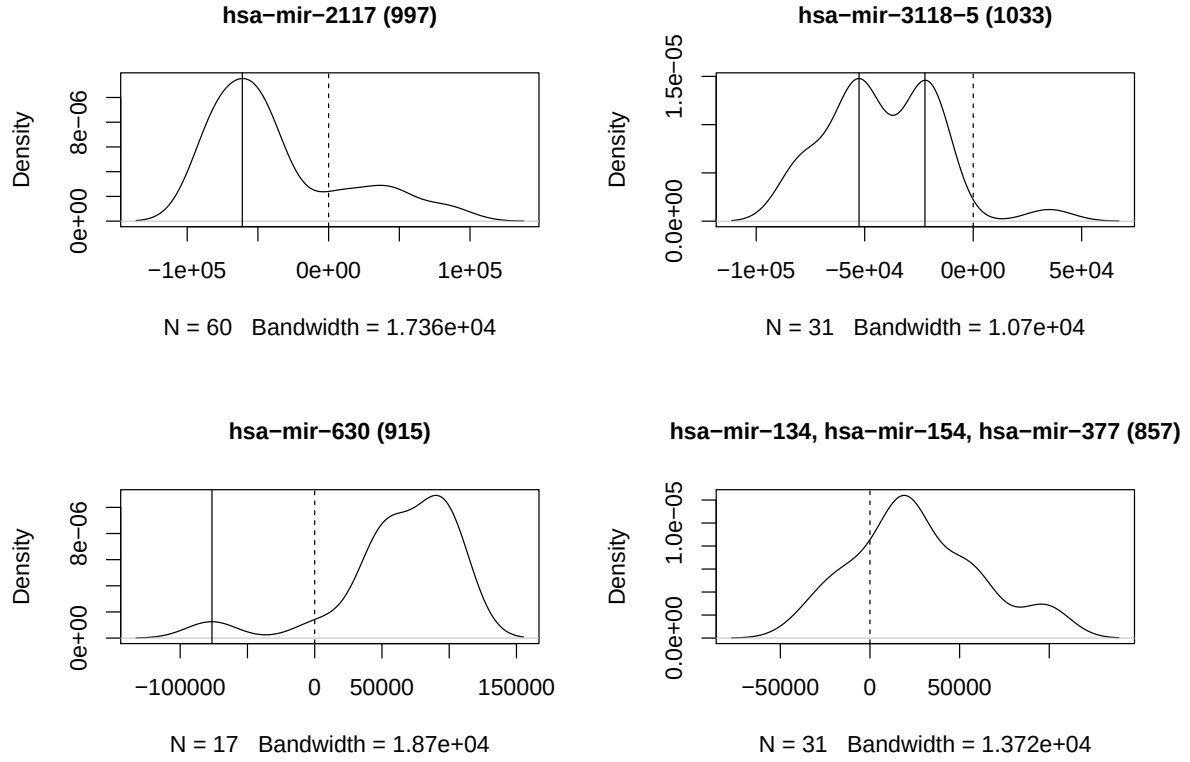


Figure 3: Examples of different results from *Eponine*. Plots show the density of all predicted TSS exceeding as score of 0.95. The number on the top represents the cluster for which the TSS should be predicted. The position is given relative to the first pre-miRNA of the cluster. Solid lines indicate all TSS used, selected as described in the *Methods* section. *977* shows a prediction with a single result and a clear peak before the first pre-miRNA. As *1033* shows, similar results are also possible with more than one maximum. As the threshold of 0.5 from the highest peak only considers the ones before the cluster, it is possible that they are much smaller then maxima behind the first pre-miRNA, as shown by *915*. In the case of *857*, all peaks are located behind the start of the cluster and therefore no TSS is predicted.

The distribution of the distance between thee TSS and the pre-miRNA annotated in *miRStart* shows no common distance. Even for more TSS are in the first 10 kb as on average, the overall distribution looks nearly homogeneous (see Figure 4a). This is not expected as it leads to a high number of pre-miRNA with a large pri-miRNA in front of them.

As the number also does not drop towards the end of the 50 kb region examined

by *miRStart*, distances beyond this border should be expected to exist. So TSS are expected that can not be annotated in *miRStart*.

The distribution of distances for the *Eponine* data differs from the distribution of *miRStart*. Also the density is nearly homogeneous for TSS and drops only slightly with higher distance, it differs from that of *miRStart*. The number of TSS in the first 10 kb is visible higher than the average. This cannot be explained by miRNAs for which no prediction was available, as they were excluded from the examination.

In order to test the predictions of Eponine, we compared the relative positions of the TSS to the pre-miRNA to the relative positions from *miRStart*. As the TSS of the intronic miRNA is determined by their host gene TSS, they are assumed to be the same for the to data sets and we only examined exonic miRNAs.

miRNAs without an prediction were excluded from the test leaving 616 of the 720 exonic miRNAs. If more than one predicted TSS was available, we used the one closest to the miRStart data. This seems justified as more then one TSS is possible for genes and therefor for pri-miRNA, too.

The direction comparison of the two distance sets shows no correlation between them (see figure 4b). The *Pearson correlation* of only 0.11 supports the conclusion drawn from the plot. Therefore we did not use the data from the *Eponine* prediction in the further steps.

4.2 Comparison of miRStart to Promoter prediction

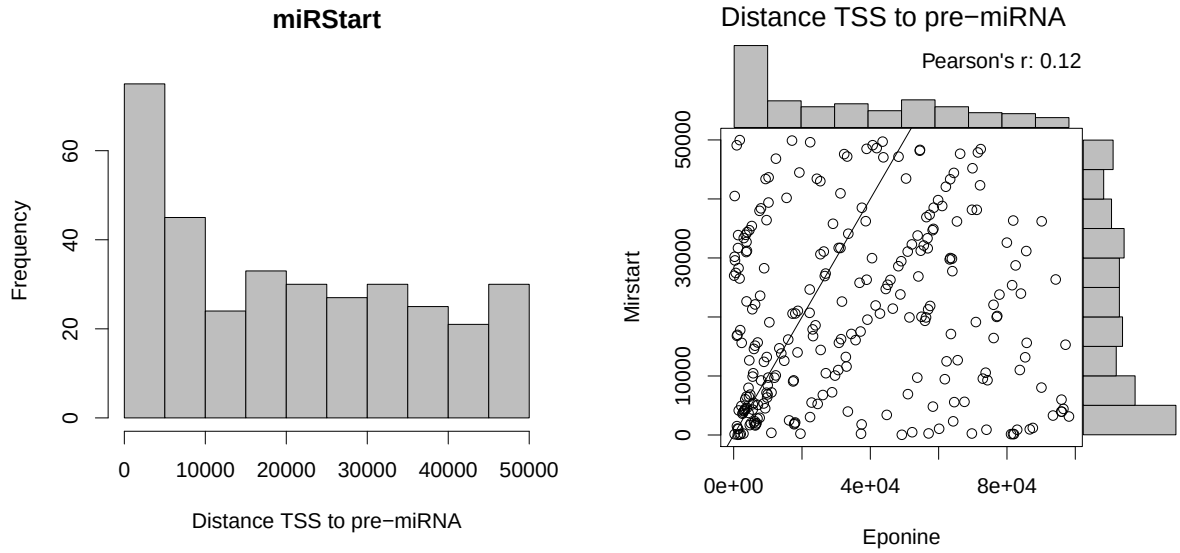
As both, the annotated data and the predictions, can not be used, we simply use the start of the first pre-miRNA as TSS for exonic miRNAs. From this position we searched the promoter. As promoter prediction is similar to the prediction of the TSS, the disadvantage of the missing real TSS should be marginal [19].

Looking through the results of the different promoter predictions tools shows that results from *Ep3* and *PromoterInspector* are comparable. Most of the *PI* results were in regions were the score of *Ep3* was also high. Additionally their results are often near the TSS as it is the case for *miRStart* (see Figure 5).

The similarity between *Ep3* and *PromoterInspector* is also supported by a comparison between the *Ep3* score of region overlapping the *PromoterInspector* result to the ones not overlapping (see Figure 6). Comparing the distribution of this two sets using the *Wilcox test* results in a p-value smaller than $2.2 \cdot 10^{-26}$. Therefore this two distributions are shown to be different. The *Ep3* score for *PromoterInspector* regions is significantly higher for the rest of the examined sequence.

Promoter2 shows for each input sequence several predictions with an high score. As the use of all of them would increase the region searched for TFBS and selecting from many possible regions has a high risk of using the wrong, we used *PromoterInspector*. Its results are similar to the ones of *Ep3*, but it returns a region instead of scores.

PromoterInspector returns results for 714 clusters. For 179 clusters more than one promoter was predicted. As it was applied for 1384 sequences, it returns a result for half of them. Therefore its application seems to be useful and we use its results.



(a) (b)

Figure 4: (a) Distribution of the distance between the TSS as determined by miRStart and the position of the pre-miRNA as extracted from mirbase. It shows clearly that near the pre-miRNA the TSS are more frequent. But also the number near 50 kb does not drop towards zero and it should be expected that TSS behind 50 kb would be found if the searched region would be enlarged. (b) Comparison of the distance between the TSS and the pre-miRNA as annotated in miRStart and predicted by Eponine. The histograms on the top and the right side show that the predictions of Eponine have a smaller distance to the pre-miRNA. The black line shows were the points should be if the predictions were the same as the miRStart data.

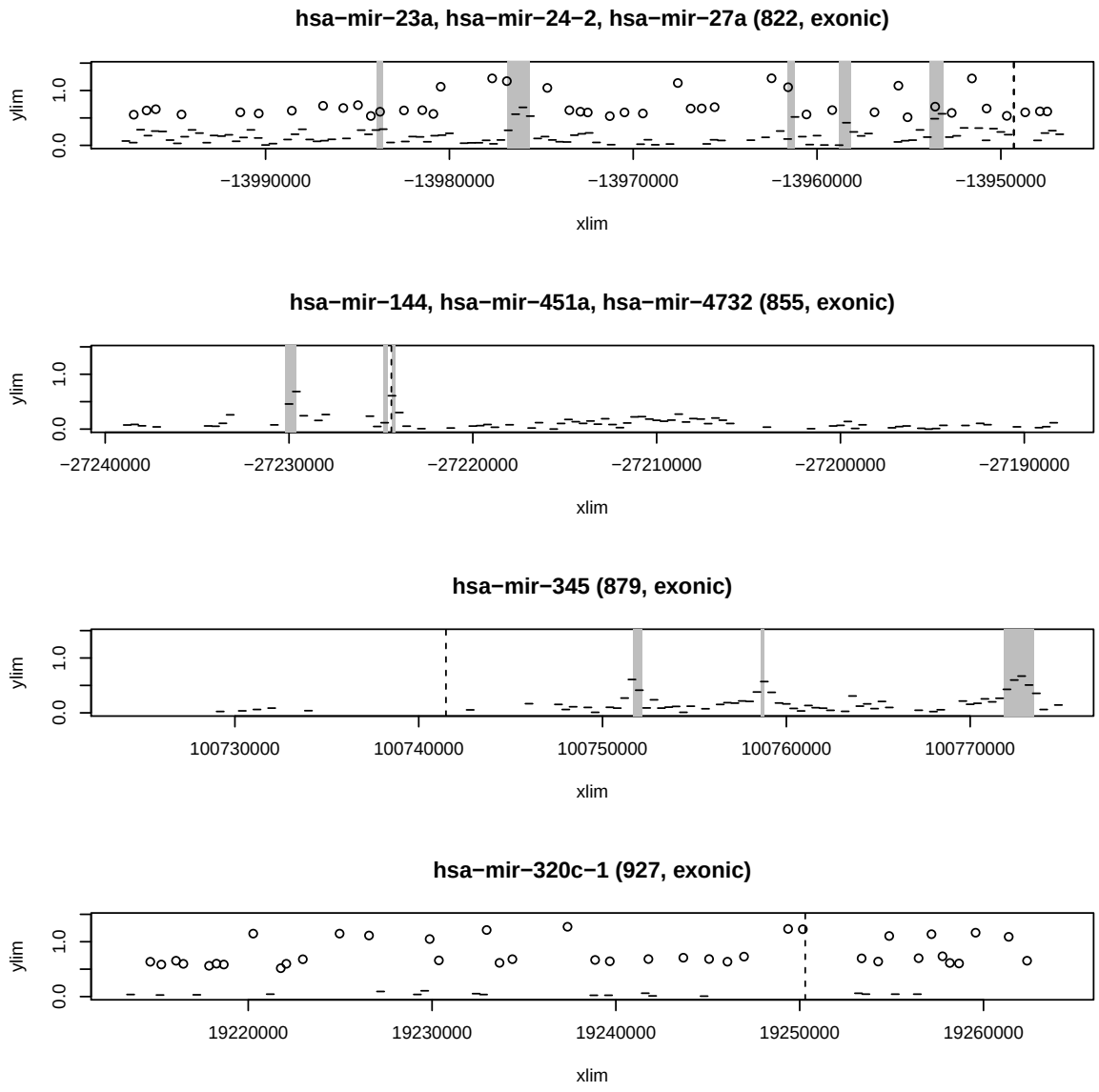


Figure 5: Examples for the comparison of the different promoter prediction tools. As the promoter is expected to be near the TSS the one as annotated in miRStart is shown as dashed line. The circles show results from *Promoter2*, the lines from *Ep3* and the grey boxes from *PromoterInspector*. For *822*, where the TSS is known to be more downstream as annotated in *miRStart*, the results are even further apart from the pre-miRNA. The predictions of *Ep3* and *PI* seem to be similar and often near the TSS. *927* is a example for *Promoter2* being near the TSS but the both other tools do not return a result.

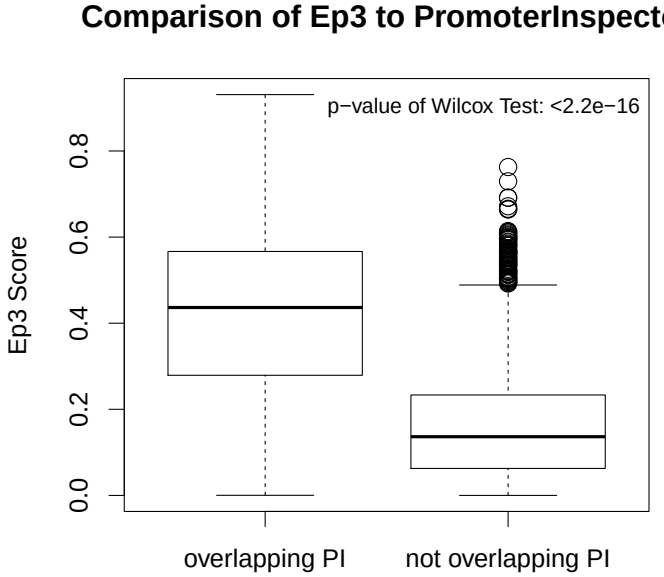


Figure 6: Comparison between Ep3 scores of regions overlapping the predictions of PromoterInspector and regions not overlapping. The median for overlapping regions is 0.44 and for not overlapping regions is 0.14. The p-value of the *Wilcox Test* also indicates that this two sets have a different distribution.

4.3 Evaluation of the Network

4.3.1 miRNA degree

The degree distribution of the miRNAs shows two maxima for the network created using the *Transfac* data (Figure 7b). For the networks based on *ModelInspector* these two clear peaks do no exist (Figure 7a).

This can be explained by the different size of the used promoter regions. For the clusters with a prediction from *PromoterInspector* this region has an average size of 508 bp, but a region of 1500 bp is used for the ones without prediction, which should lead to a higher number of detected TFBS.

Splitting the data for these two kind of miRNAs into two sets, one for miRNAs with promoter prediction and one for those without, and examining the their miRNA degree distribution indicates, that the degree of the ones without prediction is higher as for the ones with prediction (Figure 7c and 7d). This confirms the explanation above.

Also for the *ModelInspector* data, the two peaks are clearly visible after the size of the promoter is distinguished. As the distance is not as big as for the *Transfac* based network and also the overlap, between the two types of promoters, is higher, the size seems to have a small influence on the *ModelInspector* predictions.

4.3.2 Comparison to other Networks

For the evaluation of the so constructed networks, they were compared to the data from *ChIPBase*. In an first step all TFs and miRNAs are identified that are contained in all examined networks. For the networks based on *Transfac* data we used the one with the lowest threshold, as this contains the most miRNAs and TFs. As together with this network also *PuTmiR* should be compared to *ChIPBase*, it is also considered by the determination of the used TFs and miRNAs.

For the evaluation all possible TF-miRNA interactions were used to create a contingency table. The two examined classifications were if they were discovered by our prediction and if they are stored in *ChIPBase* . *Fisher's exact test* was used to examine this contingency table.

As all examined networks contain TF-miRNA interactions, their information is expected to be similar. Therefore a miRNA and a TF which have an interaction in one network should have a higher probability to have an interaction in the other network. Because of that, the two classification should be dependent. As *Fisher's exact test* checks for the independents of the two examined variables, it should return a small *p-value*.

As *ModelInspector* does not return TFs, but TF families, before the comparison the TFs of the other network have to be conversed to the same TF families as used by *Genomatix*. For this the mapping is extracted from the *MatBase* site describing all vertebrates TF families. Then the data from *ChIPBase* is transferred from a TF-miRNA to a TFfam-miRNA network.

Comparing the *ModelInspector* data to *ChIPBase* reveals that the classification of both are dependent (Fisher's exactest p-value : $3.18 * 10^{-29}$ (see contingency table 1a). Therefore we consider the network created with *ModelInspector* as useful, as it is

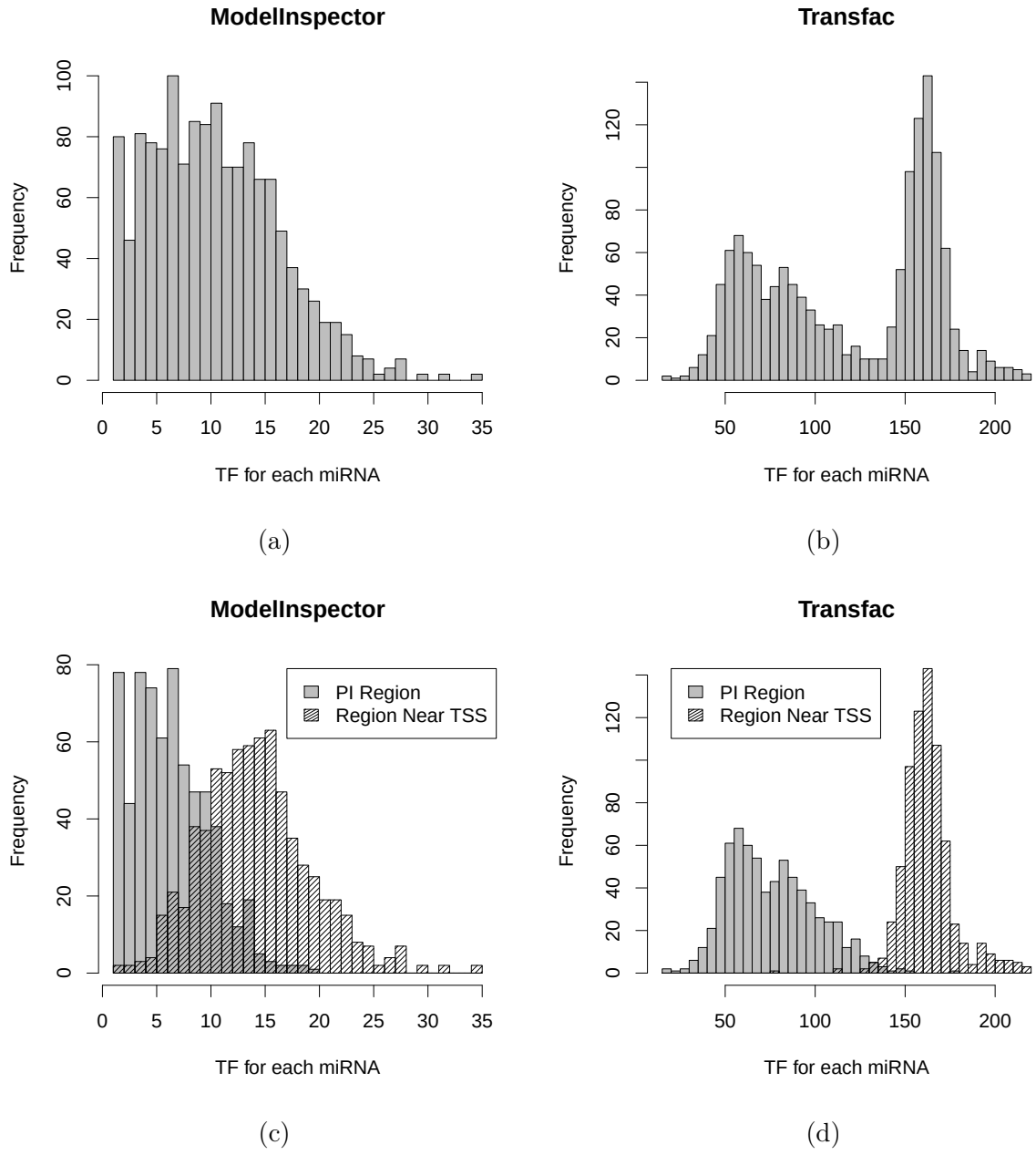


Figure 7: Distribution of the number of TF per miRNA. *Transfac* marks data for which simple PWM from Transfac were used. *ModelInspector* indicates plot, for which this method was used and the number is the number of TF families as defined by *Genomatix*. For the bottom two plots, the grey bars show the numbers for miRNAs their promoter was predicted by *PromoterInspector*. The hatched boxes are for miRNAs where no promoter could be predicted and simply the regions surrounding the TSS was used. It is clearly visible that the two peaks of for the *Transfac* diagrams are a result of the different length of the promoters predicted and the 1.5 kb region used around the TSS in the case of no predictions available.

comparable to a network created with experimentally data. A comparison to *TransmiR* was not possible, for no TF family is in both networks.

Comparison of the *TRANSFAC* data against *ChIPBase* for different thresholds reveals a optimal threshold of 0.94 (Figure 8). For the following analysis this value is used. As the p-value for this network is much smaller as the ones for *PuTmiR*, this network is more similar to the *ChIPBase* network.

A closer look on the contingency tables shows that for both, *TRANSFAC* and *PuTmiR*, only using upstream TFBSs, (Table 1b and d), the number of experimentally not confirmed predictions is nearly the same (about 1000 bp). However, the number of experimentally confirmed predictions for the *TRANSFAC* based network (538) is about three fold higher than for the *PuTmiR* upstream data (172). Even using all data from *PuTmiR*, it has less experimentally confirmed interactions (316) as the network we created using *TRANSFAC* (Table 1e). Additionally the number of not confirmed interaction rises to about 2000, what is about twice as much as before.

A test between the network predicted using *TRANSFAC* data and *TransmiR* (table 1c) also shows a similarity between this to networks (p-value of Fisher's exact test: 0.00012).

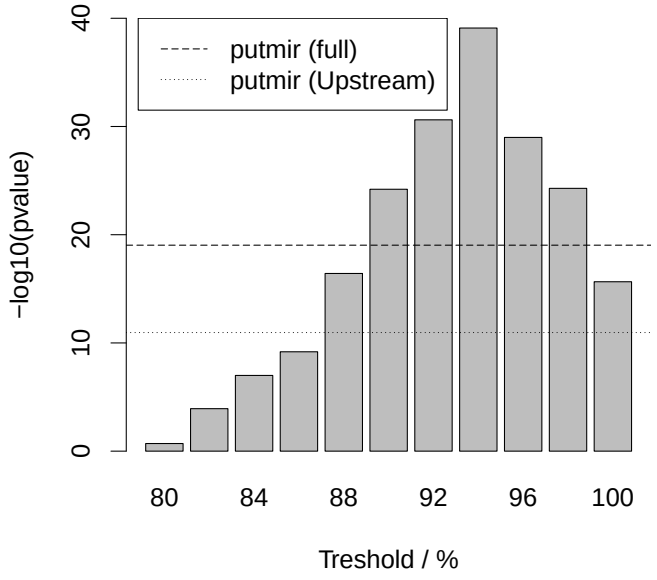


Figure 8: Comparison between the network created by using PWMs from *TRANSFAC* and *PuTmiR*. For *TRANSFAC* different thresholds are shown in order to find the optimum. The quality of the networks was calculated based on the comparison to the data from *ChIPBase*. The two dashed lines show the *p-values* for *PuTmiR*. The top one is based on the full data available in PuTmiR, the bottom one only TFBS upstream of the miRNA were used.

∩	**ModelIns**	$\overline{\textbf{ModelIns}}$
ChIPBase	944	1749
$\overline{\textbf{ChIPBase}}$	1306	4321

(a)

∩	**TRANSFAC**	$\overline{\textbf{TRANSFAC}}$
ChIPBase	538	2460
$\overline{\textbf{ChIPBase}}$	1072	10765

(b)

∩	**TRANSFAC**	$\overline{\textbf{TRANSFAC}}$
TransmiR	64	20
$\overline{\textbf{TransmiR}}$	1643	1341

(c)

∩	**PuTmiR**	$\overline{\textbf{PuTmiR}}$
ChIPBase	172	2826
$\overline{\textbf{ChIPBase}}$	1125	10712

(d)

∩	**PuTmiR(full)**	$\overline{\textbf{PuTmiR(full)}}$
ChIPBase	316	2682
$\overline{\textbf{ChIPBase}}$	2020	9817

(e)

Table 1: Contingency table for the comparison of the predicted networks to data from *TransmiR* and *ChIPBase*. In the case of data from *ModelInspector* (ModelIns) the TFs of the compared database was first converted the TF families as used by *Genomatix*. For the *TRANSFAC* PWM data the optimal threshold of 0.94 was used. *PuTmiR* contains only data of upstream TFBS. By *PuTmiR(full)* all available data was used. The number of TF and miRNAs used for the matrices: (c) miRNA:118 and TF:26, (a) miRNA:520 and TFfam:16, (b,d,e) miRNA:645 and TF:23. As for *TransmiR* and *ModelInspector* no TF was in both networks, it could no be compared.

4.4 miRNAs regulated by the TF NF-κB

Too examine the influence of miRNAs regulated by *NF-κB* on pathways, *miTALOS* was used. The miRNAs were read from the network of interest and entered in the *miTALOS* website. No filtering for tissues or certain pathways was applied. Only the default settings were used, so the organism searched was *human* and the miRNA targets were searched using *TargetScan* [15].

Using the *TRANSFAC* data with a threshold of 0.94, 25 miRNAs are predicted as being regulated from *NFKAPPAB*. The same TF returns 27 miRNAs using *PuTmiR* upstream data and 54 using *PuTmiR* full. As the numbers are more similar, only the upstream data will be used in the following.

miTALOS returns 10 pathways for *TRANSFAC* and 15 for *PuTmiR*. 7 out of 10 found for *TRANSFAC* are directly related to cancer. For *PuTmiR* it is only in 4 out of 15 results the case. This shows an enrichment of cancer related pathways being influenced by NF-κB through miRNAs, according to the *TRANSFAC* based network and *miTALOS*. 3 pathways are found for both networks used ("Melanoma", "Prostate cancer" and "Endometrial cancer", all from *KEGG*).

Also interesting is the *Notch Pathway*, as it has the lowest p-value for the *Enrichment Score* of all results for the *TRANSFAC* based network, which are not directly related to cancer. Additionally the p-value for its *Proximate Score* are lower as the ones from the cancer related pathways. From the four proteins targeted in the *Notch pathway* by the miRNAs, one is *Notch* it self and the other three directly regulate *Notch* (see Figure 9).

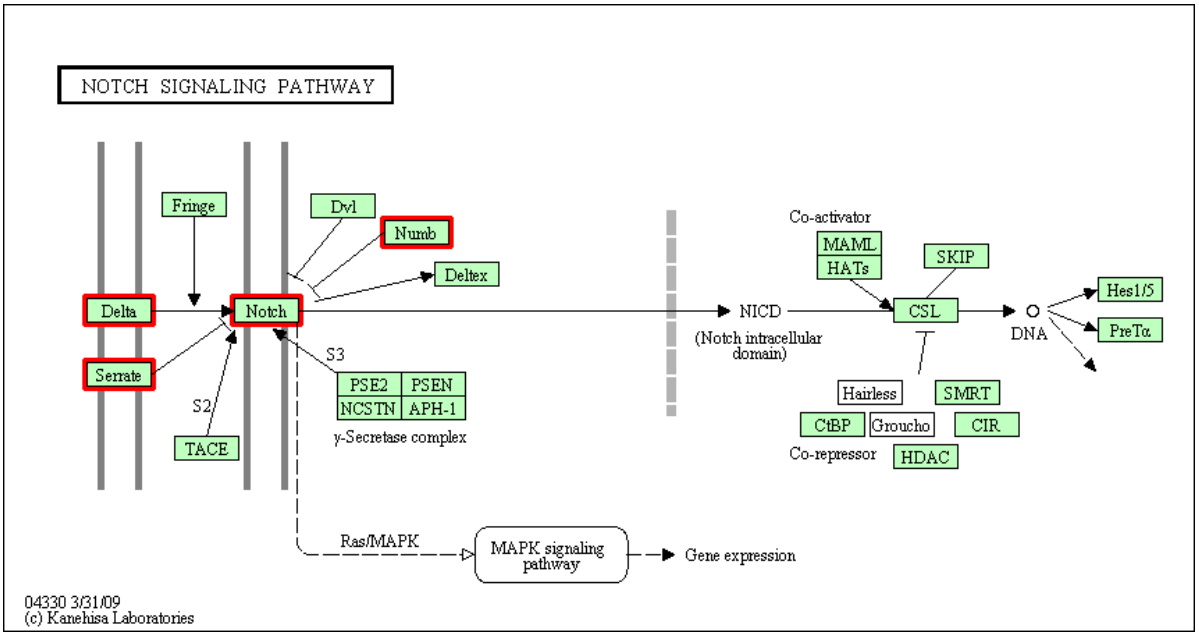


Figure 9: Example for a pathway as returned by *miTALOS*. The four proteins with a red border are suppressed from miRNAs which are regulated by NF-κB. One of the proteins is *Notch*. The other three directly interact with it. (source miTALOS [52] and KEGG [53])

5 Discussion

We created a TF-miRNA network based on the genomic sequence and the genomic positions of all known miRNAs of an organism. Additional information about TFs in this organism were needed like PWMs for their known binding sites.

After grouping these miRNAs into clusters, which are expected to underlie the same regulation, *Eponine* was used to predict the TSS. In the case of a misfit between these TSS and the annotated ones in *MiRStart*, we did not use this prediction and simply the start of the cluster was used as TSS.

Based on these TSS we searched the promoter. Finally we searched for TFBS in this promoter regions to retrieve the TF-miRNA network.

The comparison to *PuTmiR* through the use of experimentally validated data from *ChIPBase* and *TransmiR* showed the practicability of our approach. Also pathways influenced by miRNAs regulated by the TF NF-κB show to be often related to cancer. This corresponds to the fact that this TF is known to play a role in cancer cells [58].

5.1 Problems of Transcription Start Site search

The comparison between the *Eponine* prediction and the *miRStart* data shows no correlation. The question arising from this observation is, whether *Eponine* is not good in finding TSS in general, it can not predict TSS for *miRNA* or the comparison to *miRStart* is not able to evaluate the prediction.

The main reason why the evaluation may fail to compare the prediction to the annotation correctly is that more than one TSS is possible for each gene. If the pri-miRNAs have more than one TSS and a correct one is predicted but differ from the one annotated, which is also a real TSS, this would lead to a difference between the two data sets. As all predictions from the density method are compared and only the closest one is used, this cannot explain the observation.

Another explanation would be difficulties in predicting TSS for miRNAs in general. Bhattacharyya et al. showed that prediction tools developed for TSS of protein coding genes perform badly on miRNAs [59]. The training of a prediction tool especially for finding miRNAs, as described in their paper, may improve the result of TFBS prediction as it would lead to a better starting point for promoter prediction.

5.2 Quality of the Network

5.2.1 Influence of Promoter Size on TFBS Prediction

The comparison of the miRNA degrees for those with promoter prediction from *ModelInspector* to those without shows a higher degree for those without. This could be explained by the larger region used in the case no prediction was available.

This differences are lower for TFBS predictions of *ModelInspector* than for the ones created by the use of simple *PWMs*. Also for *ModelInspector* exists a overlap region, which degrees are present in many miRNAs with prediction and without prediction.

Pathway (Source)	$P(E)$	$P(P)$
Melanoma (KEGG)	0.0001	0.9531
Prostate cancer (KEGG)	0.0005	0.9961
Thyroid cancer (KEGG)	0.0018	0.6570
Bladder cancer (KEGG)	0.0023	0.7654
Non-small cell lung cancer (KEGG)	0.0073	0.8678
Colorectal cancer (KEGG)	0.0111	1.0000
Endometrial cancer (KEGG)	0.0136	0.8027
Notch signaling pathway (KEGG)	0.0160	0.0987
Adherens junction (KEGG)	0.0230	0.8120
Botulinumtoxin (NCI)	0.0313	0.2790

(a) Network constructed using Transfac PWMs

Pathway (Source)	$P(E)$	$P(P)$
Neurotrophin signaling pathway (KEGG)	0.0001	0.0716
Long-term potentiation (KEGG)	0.0001	0.4167
Prostate cancer (KEGG)	0.0001	0.4708
mTOR signaling pathway (KEGG)	0.0002	0.0143
ErbB signaling pathway (KEGG)	0.0004	0.4964
Chronic myeloid leukemia (KEGG)	0.0008	1.0000
Melanoma (KEGG)	0.0026	1.0000
Dorso-ventral axis formation (KEGG)	0.0056	1.0000
Endometrial cancer (KEGG)	0.0091	0.5260
Type II diabetes mellitus (KEGG)	0.0121	0.1139
Hdac class iii (NCI)	0.0322	0.2862
Smad2 (NCI)	0.0331	0.7667
Maturity onset diabetes of the young (KEGG)	0.0346	1.0000
Foxm1 (NCI)	0.0385	0.0801
Telomerase (NCI)	0.0396	0.6281

(b) Putmir (full)

Table 2: Result of *miTALOS* for miRNAs regulated by NF-κB. $P(E)$ and $P(P)$ are the p-values for the *Enrichment score* and the *Proximate score* as calculated by *miTALOS*.

This is not the case for the *TRANSFAC* based data, for which miRNA with and without prediction can be separated clearly.

An explanation for this effect could be a higher specificity of *ModelInspector*, as it would lead to a lower number of falsely detected TFBS. A perfect method, only finding true TFBS, would only be influenced slightly by the length of the searched region, as most of the TFBS are located in the promoter and nearly no TFBS would by found around it. On the other hand, a complete random method would find more TFs binding to a long sequence than to a short one, as it is possible to find a binding at more positions. These extreme examples suggest that a more specific method should only predict slightly more TFBS for a larger region around the promoter than for the promoter alone.

As the idea behind the use of TFBS models, which are used by *ModelInspector*, is the increase of specificity, this could explain the observed effect. Therefore the use of *ModelInspector* seems to be more reasonable if the number of false connections should be low. The drawback is that no TF, but only TF families are found and that TFBS being not part of a known model will not be found.

5.2.2 More hits than PuTmiR

Comparing the numbers of connections verified and unverified by *ChIPBase* between the network constructed here and *PuTmiR* revealed that *PuTmiR* has less verified connections. Even if up- and downstream data are used, the number for the ones predicted by the *TRANSFAC* based one is higher. But the use of both data sets also increase the number of unverified connections.

As the region examined by *PuTmiR* is much larger than the one used here, 20 kb compared to maximal 1.5 kb, the result is unexpected. Searching for the binding site of a certain TF in a larger region should increase the number of result, but less are observed.

This can be explained by the fact that *PuTmiR* only uses conserved TFBS being present in all three examined species. As not all miRNA present in human exist also in mouse and rat, not all TFBS can be found by this method. Regulation not conserved but present in human are not recognised, too.

The higher number of unverified interactions on the other hand can be explained by the large region that were examined. They are large enough that TFBS of other genes could be part of them. Of course this binding sites do not belong to the miRNAs and therefor are false positives.

5.3 NFKAPPAB and Notch

The TF NF-κB is known to have a negative influence on the effectiveness of cancer therapies, like ionizing radiation or the use of daunorubicin [58]. As our network shows a connection between this TF and several cancer pathways, miRNAs seems to play a role in the connection between NF-κB and cancer. This is also supported by the fact, that miRNA deregulation is found in human cancer cells [60].

Searching *PubMed* for papers about the relationship of TFs and miRNAs in cancer ("miran[Title] OR microrna[Title]) AND cancer[Title] AND "transcription factor"[Title]") resulted in only three results [61, 62, 63], where only one is about the TF-miRNA interaction in cancer [61]. Therefore better TF-miRNA networks may be able improve our understanding of this relation in cancer cells.

Also the *Notch Pathway*, mentioned above for its high *Proximate Score*, is interesting in the aspect of cancer. As it is known to influence the survival rate by cancer [64], *Notch* is a promising target for cancer treatment [65]. So, also this pathway is related to cancer, what increases the number of cancer related pathways influence by NF-κB through miRNAs according to the combination of our network and *miTALOS*. The other targeted gene products of this pathway could also give more insight into the connection of the influence of this TF and *Notch* on cancer.

This shows that a TF-miRNA network could reveal interesting aspects of the known correlation between TFs and human diseases, like cancer. A connection of such a network to regulatory pathways, using a tool like *miTALOS*, could give insight into unknown aspects of cellular deregulation leading to diseases.

5.4 Outlook

5.4.1 Tissue Specificity

Our current pipeline does not consider any tissue our state specific data. As *TRANSFAC* [33] contains tissue- and state-specific matrices, this information could be used to generate networks for different cell types or conditions.

With the help of this PWMs, our pipeline could be used to predict these networks. For researchers, which are interested in a specific cell type or different cells states, these networks may reveal additional insight into gene regulation.

5.4.2 Uses in other Projects

Our network could be used to improve methods based on gene regulatory networks, like the *Multilevel ONtology Analysis* (MONA) [66]. It could add an additional layer of regulation which might help to improve the inference of biological processes among a given set of genes.

If this application of our network would improve the results of such a method, it would prove the success of our work.

Acknowledgement

I thank Philipp Bruns for developing the density based method to analyse *Eponine* data we used here. I also thank Kinga Balazs, Hagen Fritsch, Gudrun Idrissou and My Nguyenly who started this project. Of course I also thank Steffen Sass and Nikola S. Mueller for the help during the project and Fabian J. Theis for offering me the opportunity to write this thesis at his lab.

References

[1] R. C. Lee, R. L. Feinbaum, V. Ambros, *et al.*, "The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14," *Cell*, vol. 75, no. 5, pp. 843–854, 1993.

[2] L. He and G. J. Hannon, "Micrornas: small rnas with a big role in gene regulation," *Nature Reviews Genetics*, vol. 5, no. 7, pp. 522–531, 2004.

[3] J. McEntyre and D. Lipman, "Pubmed: bridging the information gap," *Canadian Medical Association Journal*, vol. 164, no. 9, pp. 1317–1319, 2001.

[4] D. S. Latchman, "Transcription factors: an overview," *The international journal of biochemistry & cell biology*, vol. 29, no. 12, pp. 1305–1312, 1997.

[5] H. Yu, K. Tu, Y.-J. Wang, J.-Z. Mao, L. Xie, Y.-Y. Li, and Y.-X. Li, "Combinatorial network of transcriptional regulation and microrna regulation in human cancer," *BMC Systems Biology*, vol. 6, no. 1, p. 61, 2012.

[6] J. Wang, M. Lu, C. Qiu, and Q. Cui, "Transmir: a transcription factor–microrna regulation database," *Nucleic acids research*, vol. 38, no. suppl 1, pp. D119–D122, 2010.

[7] P. J. Farnham, "Insights from genomic profiling of transcription factors," *Nature Reviews Genetics*, vol. 10, no. 9, pp. 605–616, 2009.

[8] S. Bandyopadhyay and M. Bhattacharyya, "Putmir: a database for extracting neighboring transcription factors of human microrna," *BMC bioinformatics*, vol. 11, no. 1, p. 190, 2010.

[9] Y. Lee, M. Kim, J. Han, K.-H. Yeom, S. Lee, S. H. Baek, and V. N. Kim, "Microrna genes are transcribed by rna polymerase ii," *The EMBO journal*, vol. 23, no. 20, pp. 4051–4060, 2004.

[10] T.-C. Chang, E. A. Wentzel, O. A. Kent, K. Ramachandran, M. Mullendore, K. H. Lee, G. Feldmann, M. Yamakuchi, M. Ferlito, C. J. Lowenstein, *et al.*, "Transactivation of mir-34a by p53 broadly influences gene expression and promotes apoptosis," *Molecular cell*, vol. 26, no. 5, p. 745, 2007.

[11] Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim, *et al.*, "The nuclear rnase iii drosha initiates microrna processing," *nature*, vol. 425, no. 6956, pp. 415–419, 2003.

[12] G. Hutvagner, J. McLachlan, A. E. Pasquinelli, É. Bálint, T. Tuschl, and P. D. Zamore, "A cellular function for the rna-interference enzyme dicer in the maturation of the let-7 small temporal rna," *Science Signalling*, vol. 293, no. 5531, p. 834, 2001.

[13] P. Jin, D. C. Zarnescu, S. Ceman, M. Nakamoto, J. Mowrey, T. A. Jongens, D. L. Nelson, K. Moses, and S. T. Warren, "Biochemical and genetic interaction between the fragile x mental retardation protein and the microrna pathway," *Nature neuroscience*, vol. 7, no. 2, pp. 113–117, 2004.

[14] H. Guo, N. T. Ingolia, J. S. Weissman, and D. P. Bartel, "Mammalian micrornas predominantly act to decrease target mrna levels," *Nature*, vol. 466, no. 7308, pp. 835–840, 2010.

[15] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets," *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[16] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements," *Nature Reviews Genetics*, vol. 5, no. 4, pp. 276–287, 2004.

[17] D. Nikolov and S. Burley, "Rna polymerase ii transcription initiation: a structural view," *Proceedings of the National Academy of Sciences*, vol. 94, no. 1, pp. 15–22, 1997.

[18] S. M. Paranjape, R. T. Kamakaka, and J. T. Kadonaga, "Role of chromatin structure in the regulation of transcription by rna polymerase ii," *Annual review of biochemistry*, vol. 63, no. 1, pp. 265–297, 1994.

[19] A. G. Pedersen, P. Baldi, Y. Chauvin, S. Brunak, *et al.*, "The biology of eukaryotic promoter prediction-a review," *Computers & Chemistry*, vol. 23, no. 3-4, pp. 191–207, 1999.

[20] S. Griffiths-Jones, H. K. Saini, S. Van Dongen, and A. J. Enright, "mirbase: tools for microrna genomics," *Nucleic acids research*, vol. 36, no. suppl 1, pp. D154–D158, 2008.

[21] A. Kozomara and S. Griffiths-Jones, "mirbase: integrating microrna annotation and deep-sequencing data," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D152–D157, 2011.

[22] A. Rodriguez, S. Griffiths-Jones, J. L. Ashurst, and A. Bradley, "Identification of mammalian microrna host genes and transcription units," *Genome research*, vol. 14, no. 10a, pp. 1902–1910, 2004.

[23] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, *et al.*, "Ensembl 2012," *Nucleic acids research*, vol. 40, no. D1, pp. D84–D90, 2012.

[24] J. Ashurst, C.-K. Chen, J. Gilbert, K. Jekosch, S. Keenan, P. Meidl, S. Searle, J. Stalker, R. Storey, S. Trevanion, *et al.*, "The vertebrate genome annotation (vega) database," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D459–D465, 2005.

[25] V. Curwen, E. Eyras, T. D. Andrews, L. Clarke, E. Mongin, S. M. Searle, and M. Clamp, "The ensembl automatic gene annotation system," *Genome research*, vol. 14, no. 5, pp. 942–950, 2004.

[26] R. J. Kinsella, A. Kähäri, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, *et al.*, "Ensembl biomarts: a hub for data retrieval across taxonomic space," *Database: the journal of biological databases and curation*, vol. 2011, 2011.

[27] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[28] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber, "Biomart and bioconductor: a powerful link between biological databases and microarray data analysis," *Bioinformatics*, vol. 21, no. 16, pp. 3439–3440, 2005.

[29] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, "Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart," *Nature protocols*, vol. 4, no. 8, pp. 1184–1191, 2009.

[30] S. Pelengaris, M. Khan, and G. Evan, "c-myc: more than just a matter of life and death," *Nature Reviews Cancer*, vol. 2, no. 10, pp. 764–776, 2002.

[31] L. Zhang and Y. Zhao, "The regulation of foxp3 expression in regulatory cd4+ cd25+ t cells: multiple pathways on the road," *Journal of cellular physiology*, vol. 211, no. 3, pp. 590–597, 2007.

[32] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent, "The ucsc table browser data retrieval tool," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D493–D496, 2004.

[33] V. Matys, E. Fricke, R. Geffers, E. Goessling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, *et al.*, "Transfac®: transcriptional regulation, from patterns to profiles," *Nucleic acids research*, vol. 31, no. 1, pp. 374–378, 2003.

[34] P. J. Park, "Chip–seq: advantages and challenges of a maturing technology," *Nature Reviews Genetics*, vol. 10, no. 10, pp. 669–680, 2009.

[35] J.-H. Yang, J.-H. Li, S. Jiang, H. Zhou, and L.-H. Qu, "Chipbase: a database for decoding the transcriptional regulation of long non-coding rna and microrna genes from chip-seq data," *Nucleic Acids Research*, vol. 41, no. D1, pp. D177–D187, 2013.

[36] E. P. Consortium *et al.*, "A user's guide to the encyclopedia of dna elements (encode)," *PLoS Biol*, vol. 9, no. 4, p. e1001046, 2011.

[37] N. Nègre, C. D. Brown, L. Ma, C. A. Bristow, S. W. Miller, U. Wagner, P. Kheradpour, M. L. Eaton, P. Loriaux, R. Sealfon, *et al.*, "A cis-regulatory map of the drosophila genome," *Nature*, vol. 471, no. 7339, pp. 527–531, 2011.

[38] M. B. Gerstein, Z. J. Lu, E. L. Van Nostrand, C. Cheng, B. I. Arshinoff, T. Liu, K. Y. Yip, R. Robilotto, A. Rechtsteiner, K. Ikegami, *et al.*, "Integrative analysis of the caenorhabditis elegans genome by the modencode project," *Science*, vol. 330, no. 6012, pp. 1775–1787, 2010.

[39] C.-H. Chien, Y.-M. Sun, W.-C. Chang, P.-Y. Chiang-Hsieh, T.-Y. Lee, W.-C. Tsai, J.-T. Horng, A.-P. Tsou, and H.-D. Huang, "Identifying transcriptional start sites of human microrna based on high-throughput sequencing data," *Nucleic Acids Research*, vol. 39, no. 21, pp. 9345–9356, 2011.

[40] T. Shiraki, S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, *et al.*, "Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage," *Proceedings of the National Academy of Sciences*, vol. 100, no. 26, pp. 15776–15781, 2003.

[41] M. Harbers and P. Carninci, "Tag-based approaches for transcriptome research and genome annotation," *Nature methods*, vol. 2, no. 7, pp. 495–502, 2005.

[42] H. Kawaji, J. Severin, M. Lizio, A. Waterhouse, S. Katayama, K. M. Irvine, D. A. Hume, A. Forrest, H. Suzuki, P. Carninci, *et al.*, "The fantom web resource: from mammalian transcriptional landscape to its dynamic regulation," *Genome Biol*, vol. 10, no. 4, p. R40, 2009.

[43] R. Yamashita, H. Wakaguri, S. Sugano, Y. Suzuki, and K. Nakai, "Dbtss provides a tissue specific dynamic view of transcription start sites," *Nucleic acids research*, vol. 38, no. suppl 1, pp. D98–D104, 2010.

[44] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, K. Zhao, *et al.*, "High-resolution profiling of histone methylations in the human genome," *Cell*, vol. 129, no. 4, pp. 823–837, 2007.

[45] T. A. Down and T. J. Hubbard, "Computational detection and location of transcription start sites in mammalian genomic dna," *Genome research*, vol. 12, no. 3, pp. 458–461, 2002.

[46] S. Knudsen, "Promoter2. 0: for the recognition of polii promoter sequences.," *Bioinformatics*, vol. 15, no. 5, pp. 356–361, 1999.

[47] T. Abeel, Y. Saeys, E. Bonnet, P. Rouzé, and Y. Van de Peer, "Generic eukaryotic core promoter prediction using structural features of dna," *Genome research*, vol. 18, no. 2, pp. 310–323, 2008.

[48] M. Scherf, A. Klingenhoff, T. Werner, *et al.*, "Highly specific localization of promoter regions in large genomic sequences by promoterinspector: a novel context analysis approach," *Journal of molecular biology*, vol. 297, no. 3, pp. 599–606, 2000.

[49] G. D. Stormo, "Dna binding sites: representation and discovery," *Bioinformatics*, vol. 16, no. 1, pp. 16–23, 2000.

[50] H. Pages, P. Aboyoun, R. Gentleman, and S. DebRoy, *Biostrings: String objects representing biological sequences, and matching algorithms*. R package version 2.26.2.

[51] K. Frech, J. Danescu-Mayer, and T. Werner, "A novel method to develop highly specific models for regulatory units detects a new ltr in genbank which contains a functional promoter," *Journal of molecular biology*, vol. 270, no. 5, pp. 674–687, 1997.

[52] A. Kowarsch, M. Preusse, C. Marr, and F. J. Theis, "mitalos: analyzing the tissue-specific regulation of signaling pathways by human and mouse microrna," *RNA*, vol. 17, no. 5, pp. 809–819, 2011.

[53] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, *et al.*, "Kegg for linking genomes to life and the environment," *Nucleic acids research*, vol. 36, no. suppl 1, pp. D480–D484, 2008.

[54] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow, "Pid: the pathway interaction database," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D674–D679, 2009.

[55] J. F. Reid, *mirbase.db: miRBase: the microRNA database*. R package version 1.1.0.

[56] R. Shalgi, D. Lieber, M. Oren, and Y. Pilpel, "Global and local architecture of the mammalian microrna–transcription factor regulatory network," *PLoS computational biology*, vol. 3, no. 7, p. e131, 2007.

[57] C. Cheng, K.-K. Yan, W. Hwang, J. Qian, N. Bhardwaj, J. Rozowsky, Z. J. Lu, W. Niu, P. Alves, M. Kato, *et al.*, "Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data," *PLoS Computational Biology*, vol. 7, no. 11, p. e1002190, 2011.

[58] C.-Y. Wang, M. W. Mayo, A. S. Baldwin Jr, *et al.*, "Tnf-and cancer therapy-induced apoptosis: potentiation by inhibition of nf-kappab.," *Science (New York, NY)*, vol. 274, no. 5288, p. 784, 1996.

[59] M. Bhattacharyya, L. Feuerbach, T. Bhadra, T. Lengauer, and S. Bandyopadhyay, "Microrna transcription start site prediction with multi-objective feature selection," *Statistical Applications in Genetics and Molecular Biology*, vol. 11, no. 1, pp. 1–25, 2012.

[60] M. V. Iorio, M. Ferracin, C.-G. Liu, A. Veronese, R. Spizzo, S. Sabbioni, E. Magri, M. Pedriali, M. Fabbri, M. Campiglio, *et al.*, "Microrna gene expression deregulation in human breast cancer," *Cancer research*, vol. 65, no. 16, pp. 7065–7070, 2005.

[61] B. D. Aguda, "Modeling microrna-transcription factor networks in cancer," in *MicroRNA Cancer Regulation*, pp. 149–167, Springer, 2013.

[62] J. Zhang, N. Luo, Y. Luo, Z. Peng, T. Zhang, and S. Li, "microrna-150 inhibits human cd133-positive liver cancer stem cells through negative regulation of the transcription factor c-myb," *International journal of oncology*, vol. 40, no. 3, pp. 747–756, 2012.

[63] G. Wang, Y. Wang, W. Feng, X. Wang, J. Y. Yang, Y. Zhao, Y. Wang, and Y. Liu, "Transcription factor and microrna regulation in androgen-dependent and-independent prostate cancer cells," *Bmc Genomics*, vol. 9, no. Suppl 2, p. S22, 2008.

[64] K. A. Hassan, L. Wang, H. Korkaya, G. Chen, I. Maillard, D. G. Beer, G. P. Kalemkerian, and M. S. Wicha, "Notch pathway activity identifies cells with cancer stem cell-like properties and correlates with worse survival in lung adenocarcinoma," *Clinical Cancer Research*, 2013.

[65] I. Espinoza and L. Miele, "Notch inhibitors for cancer treatment," *Pharmacology & Therapeutics*, 2013.

[66] S. Sass, F. Buettner, N. Mueller, and F. J. Theis, "A modular framework for gene ontology analysis integrating multilevel omics data." currently developed at the CMB.