Understanding Regulatory Mechanisms Underlying Stem Cells Helps to Identify Cancer Biomarkers

A dissertation submitted towards the degree Doctor of Engineering (Dr.-Ing) of the Faculty of Mathematics and Computer Science of Saarland University

by
Maryam Nazarieh
Saarbrücken, June 2018

Day of Colloquium	Jun 28, 2018
Dean of the Faculty	Prof. Dr. Sebastian Hack
Chair of the Committee	Prof. Dr. Hans-Peter Lenhof
Reporters	
First reviewer	Prof. Dr. Volkhard Helms
Second reviewer	Prof. Dr. Dr. Thomas Lengauer
Academic Assistant	Dr. Christina Backes

Acknowledgements

Firstly, I would like to thank Prof. Volkhard Helms for offering me a position at his group and for his supervision and support on the SFB 1027 project.
I am grateful to Prof. Thomas Lengauer for his helpful comments.
I am thankful to Prof. Andreas Wiese for his contribution and discussion.
I would like to thank Prof. Jan Baumbach that allowed me to spend a training phase in his group during my PhD preparatory phase and the collaborative work which I performed with his PhD student Rashid Ibragimov where I proposed a heuristic algorithm based on the characteristics of protein-protein interaction networks for solving the graph edit distance problem.
I would like to thank Graduate School of Computer Science and Center for Bioinformatics at Saarland University, especially Prof. Raimund Seidel and Dr. Michelle Carnell for giving me an opportunity to carry out my PhD studies.
Furthermore, I would like to thank to Prof. Helms for enhancing my experience by introducing master students and working as their advisor for successfully accomplishing their master projects. Moreover, I appreciate Prof. Marcel Schulz for accepting to be the reviewer of their theses.
I would like to thank to my deep interest to extend the field of study that guided me at all times.
I would like to thank all members of Prof. Helms group, especially my collaborators Dr. Mohamed Hamed, Dr. Christian Spaniol and Thorsten Will.
I owe gratitude to my family for their unconditional support throughout these years. I thank all my friends and relatives.

Abstract

Detection of biomarker genes play a crucial role in disease detection and treatment. Bioinformatics offers a variety of approaches for identification of biomarker genes which play key roles in complex diseases. These computational approaches enhance the insight derived from experiments and reduce the efforts of biologists and experimentalists. This is essentially achieved through prioritizing a set of genes with certain attributes.
In this thesis, we show that understanding the regulatory mechanisms underlying stem cells helps to identify cancer biomarkers. We got inspired by the regulatory mechanisms of the pluripotency network in mouse embryonic stem cells and formulated the problem where a set of master regulatory genes in regulatory networks is identified with two combinatorial optimization problems namely as minimum dominating set and minimum connected dominating set in weakly and strongly connected components. Then we applied the developed methods to regulatory cancer networks to identify disease-associated genes and anti-cancer drug targets in breast cancer and hepatocellular carcinoma (chapter 3). As not all the nodes in the solutions are critical, we developed a prioritization method to rank a set of candidate genes which are related to a certain disease based on systematic analysis of the genes that are differentially expressed in tumor and normal conditions (chapter 5). Moreover, we demonstrated that the topological features in regulatory networks surrounding differentially expressed genes are highly consistent in terms of using the output of several analysis tools (chapter 6). We compared two randomization strategies for TF-miRNA co-regulatory networks to infer significant network motifs underlying cellular identity. We showed that the edge-type conserving method surpasses the non-conserving method in terms of biological relevance and centrality overlap (chapter 7).
We presented several web servers and software packages that are publicly available at no cost. The Cytoscape plugin of minimum connected dominating set identifies a set of key regulatory genes in a user provided regulatory network based on a heuristic approach. The ILP formulations of minimum dominating set and minimum connected dominating set return the optimal solutions for the aforementioned problems. Our source code is publicly available (chapter 3). The web servers TFmiR and TFmiR2 construct disease-, tissue-, process-specific networks for the sets of deregulated genes and miRNAs provided by a user. They highlight topological hotspots and offer detection of three- and four-node FFL motifs as a separate web service for both organisms mouse and human (chapter 4,8).

Kurzfassung

Die Gendetektion von Biomarkern spielt eine wesentliche Rolle bei der Erkennung und Behandlung von Krankheiten. Die Bioinformatik bietet eine Vielzahl von Ansätzen zur Identifizierung von Biomarker-Genen, die bei komplizierten Erkrankungen eine Schlüsselrolle spielen. Diese computerbasierten Ansätze verbessern die Erkenntnisse aus Experimenten und reduzieren den Aufwand von Biologen und Forschern. Dies wird hauptsächlich erreicht durch die Priorisierung einer Reihe von Genen mit bestimmten Attributen. In dieser Arbeit zeigen wir, dass die Identifizierung von Krebs-Biomarkern leichter gelingt, wenn wir die den Stammzellen zugrunde liegenden regulatorischen Mechanismen verstehen. Dazu angeregt wurden wir durch die regulatorischen Mechanismen des Pluripotenz-Netzwerks in embryonalen Maus-Stammzellen. Wir formulierten und haben das Problem der Identifizierung einer Reihe von Master-Regulator-Genen in regulatorischen Netzwerken mit zwei kombinatorischen Optimierungsproblemen, nämlich als minimal dominierende Menge und als minimal zusammenhängende dominierende Menge in schwach und stark verbundenen Komponenten. Die entwickelten Methoden haben wir dann auf regulatorische Krebsnetzwerke angewandt, um krankheitsassoziierte Gene und Zielproteine für Medikamenten gegen Brustkrebs und hepatozelluläres Karzinom zu identifizieren (Kapitel 3). Im Hinblick darauf, dass nicht alle Knoten in den Lösungen wesentlich sind, haben wir basierend auf der systematischen Analyse von Genen, die unterschiedlich bei Tumor- und Normalbedingungen reagieren, eine Priorisierungsmethode entwickelt, um einen Satz von Kandidatengenen in eine Reihenfolge zu bringen, die einer bestimmten Krankheit zugeordnet sind (Kapitel 5). Darüber hinaus haben wir gezeigt, dass die topologischen Eigenschaften in regulatorischen Netzwerken, die die deregulierte Gene umgeben, sehr einheitlich in Bezug auf den Einsatz verschiedener Analysewerkzeuge sind (Kapitel 6). Wir haben zwei Randomisierungsstrategien für TF-miRNA-Co-regulatorische Netzwerke verglichen, um signifikante Netzwerkmotive herauszufinden, welche zellulärer Identität zugrunde liegen. Wir haben gezeigt, dass die Edge-Type-Erhaltungsmethode, die nicht-erhaltende Methode in Bezug auf biologische Relevanz und zentrale Überlappung übertrifft (Kapitel 7).
Wir haben mehrere Softwarepakete und Webserver vorgestellt, die allgemein und kostenlos zugänglich sind. Das Cytoscape Plugin für die Identififizierung, der minimal verbundener dominierenden Mengen identifiziert einen Satz von regulatorischen Schlüsselgenen in einem vom Benutzer bereitgestellten regulatorischen Netzwerk basierend auf einem heuristischen Ansatz. Die ILP Formulierungen, der minimal dominierenden Menge und der minimal verbundenen dominierenden Menge liefern die optimalen Lösungen für die oben vorgenannten Probleme. Unser Quellcode hierfür ist öffentlich verfügbar (Kapitel3). Die Webserver TFmiR und TFmiR2 erzeugen Krankheits-, Gewebe- und prozessspezifische Netzwerke für die von einem Benutzer bereitgestellten deregulierten Gene und miRNAs. Außerdem verwenden die Webserver topologische Merkmale, um Hotspot-Knoten

hervorzuheben und bieten die Erkennung von drei und vier Knoten FFL Motiven als separaten Web-Service für beide Organismen, Maus und Mensch (Kapitel 4,8).

Contents

List of Figures

List of Tables

Chapter 1

An Introduction to Cellular Regulatory Networks

Cancer is a disease that affects a very large number of people around the world with the potential to cause death. Being diagnosed with cancer has profound emotional effects on the patients and their families (de Leeuw et al., 2000; Cook et al., 2018). There are a variety of treatments to address this disease such as chemotherapy, radiotherapy, immunotherapy and etc, (Urruticoechea et al., 2010). Reprogramming of terminally differentiated cells to induced pluripotent stem cells (IPSCs) has recently become feasible by inducing overexpression of a few transcription factors (TFs) Oct4, Sox2, Myc and Klf4 (Takahashi and Yamanaka, 2006). This finding opened promising strategies for patient-specific regenerative medicine (Robinton and Daley, 2012). Understanding the regulatory mechanisms underlying reprogramming events helps to increase the reprogramming efficiency (Artyomov et al., 2010). To achieve this goal, we worked from the opposite direction with the aim to first understand the regulatory mechanisms underlying stem cells and the hope that this may help to identify cancer biomarkers and dominating pathways in the cancer networks. This chapter provides a basic introduction to cellular regulatory networks, specifically underlying the cellular identity in embryonic stem cells (ESCs). Then I describe the common characteristics between the behaviours of stem cells and that of cancer stem cells. Finally I provide an overview over the databases that collected regulatory interactions between TFs, miRNAs and target genes.

1.1 Regulatory networks

Gene expression is the basis of cellular identity. Basically, genes are transcribed into mRNA molecules under regulation of TFs and the resulting mRNAs are translated subsequently into the proteins. There are other regulators in the cell that affect the gene expression such as miRNAs. miRNAs are small non-coding RNAs which bind to mRNAs after transcription, leading usually to translational repression or target degradation and gene silencing (Kusenda et al., 2006; Bartel, 2009; Le et al., 2013), see Figure 1.1. Inside the cell, TFs interact with other TFs, as well as with their target genes. These interactions form a cell-specific gene regulatory network (GRN) which governs the particular cellular identity. TFs affect the transcription level of their target genes either in form of activation

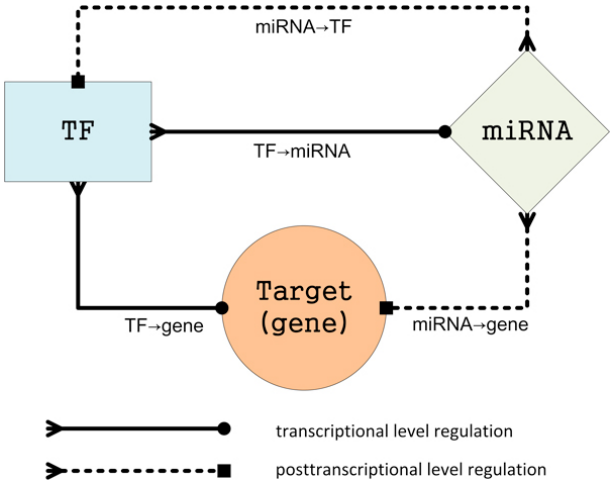

Figure 1.1: A graphical description of the regulatory effects of TFs and miRNAs on the target genes. The regulatory interactions are TF → miRNA, TF → gene, miRNA → gene and miRNA → TF. The figure is from the publication by (Xu et al., 2013).

or repression.
Modern GRNs consider the regulatory effect of miRNAs in the network. In TF-miRNA co-regulatory networks, target genes can be regulated by TFs and miRNAs. Therefore, four types of edge types corresponding to four regulatory interactions (TF → miRNA, TF → gene , miRNA → gene and miRNA → TF) are considered between TFs, miRNAs and their target genes in these networks.

1.2 Stem cells

In multicellular organisms, there are different cell types, expressing different sets of proteins with different functions. Stem cells are distinguished from other cells in the body, mainly based on the two characteristics of self-renewal and the potential for cellular differentiation. Stem cells are classified into several groups based on their different potential for cellular differentiation. Totipotent stem cells have the maximum capability of cellular differentiation. They can generate any type of cells in the body. Unipotent stem cells have the least capacity. They can generate solely cells of the same type. The reason such cells are termed stem cells is mainly based on the self-renewal ability. In between, there are pluripotent stem cells mainly in embryos that contain pluripotency networks. Pluripotency networks in mouse ESCs are maintained by a few directly interacting TFs which share many target genes (Kim et al., 2008), see Figure 1.2. Slight changes in the expression level of these TFs lead the stem cells to differentiation (Kim et al., 2008). There is hope that understanding the genetic and epigenetic states of the pluripoteny networks may help to enhance the efficiency of reprogramming, facilitating a treatment for cancer. Using this treatment strategy, terminally differentiated cells can be reprogrammed into IPSCs under the over-expression of few TFs. Creating IPSCs is essential for creating patient-specific stem cells for the purpose of regenerative medicine (Artyomov et al., 2010). Moreover, understanding the mechanisms underlying stem cells might help to target the key regulators and pathways which are responsible for tumor growth in

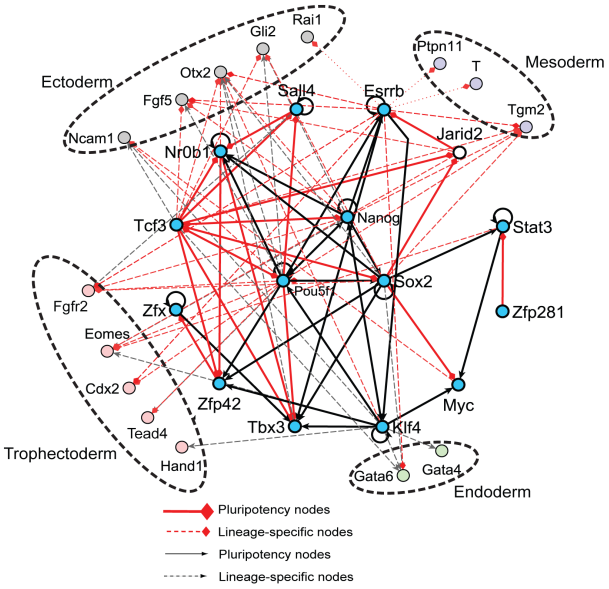

Figure 1.2: A graphical description of the pluripotency network in a mouse ESC. The nodes colored blue display the pluripotency nodes. The other colors show the lineage markers. Arcs indicate that the node at the tail regulates the expression of the head node. The figure is from the publication by (Xu et al., 2014).

cancer stem cells (Nazari et al., 2018).

1.3 Cancer

Cancer constitutes a group of diseases involving abnormal cell-growth with the potential of causing metastasis, see Figure 1.3. These diseases are generally divided into benign tumors as well as malignant tumors. Malignant tumors are usually progressive. It has been shown that a minimal set of deregulated biological processes such as deregulated cell proliferation and suppression of apoptosis are enough to render the cancer progressive (Evan and Vousden, 2001). There is a hypothesis that malignant cancers originate either from tissue-specific cells or progenitors that have abnormal indefinite divisions (Cohnheim, 1867).

Cancer stem cells are special stem cells mainly because they can generate an indefinite number of cancer cells and they can also generate different types of tumors (Reya et al., 2001). Based on cancer stem cell hypothesis that a subpopulation of tumor cells have characteristic similar to ESCs and cause the tumor growth (Reya et al., 2001; Rahman et al., 2011), researchers and therapists target the cancer stem cells than the whole tumors (Tan et al., 2006). Therefore, identification of key regulators of the cancer stem cells which are responsible for tumor growth is a promising strategy for cancer treatment (Nazari et al., 2018).

In this thesis, I worked mainly on two types of cancers, breast cancer (breast neoplasms) and liver cancer (hepatocellular carcinoma). Breast cancer is the most common invasive group of diseases in women that develops from breast tissue (McGuire et al., 2015). It occurs due to abnormal proliferation of abnormal breast cells. Damaged cells can infect

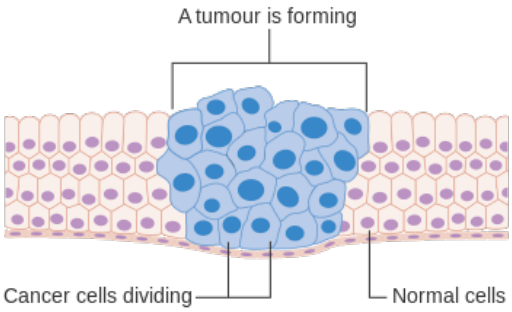

Figure 1.3: Illustration of tumor forming. The figure is from the (Cancer research UK)

other tissues by spreading through the body. Hepatocellular carcinoma is one of the most common cancers which occurs in the liver and affects more men than women (Seton-Rogers, 2014).

1.4 Regulatory Databases

To construct TF-miRNA co-regulatory disease networks, TFmiR and TFmiR2 web servers use a variety of regulatory interaction databases for the case of TF → gene, TF → miRNA, miRNA → gene, miRNA → miRNA and gene-gene interactions. The networks constructed based on these databases are filtered to disease-specific networks using disease databases.

Transcription factor-target gene interaction (TF → gene)
TRANSFAC is a manually curated database of TFs, their target genes and regulatory binding sites for a variety of eukaryotic organisms including mouse and human (Matys et al., 2003, 2006).
OregAnno is an open-source and open-access TF → gene interaction database which utilizes a literature curation system for annotation of experimentally identified regulatory binding sites from published papers for a variety of species including mouse and human (Griffith et al., 2008; Lesurf et al., 2016).
TRED is a database of TF-target gene pairs providing experimentally-validated and computationally predicted regulatory elements such as promoters and binding motifs for rat, mouse and human (Jiang et al., 2007).

Transcription factor-miRNA interaction (TF → miRNA)
TransmiR is a regulatory database of TF → miRNA interactions. It compiles manually curated experimentally validated interactions from a wide variety of publications determining their associations in tumors or other diseases (Wang et al., 2010).
ChIPBase employs the ChIP-Seq technology which combines chromatin immunoprecipitation with next-generation DNA sequencing to identify transcription factor binding site with high sensitivity in diverse tissues and cell lines for six organisms including mouse and human. Based on these predictions a large database of regulatory interactions between TFs and miRNAs were compiled (Yang et al., 2013; Zhou et al., 2017).

miRNA-target gene interaction (miRNA → gene)
TarBase (Sethupathy et al., 2006; Vergoulis et al., 2012), miRTarBase (Hsu et al., 2010; Chou et al., 2018) and miRecords (Xiao et al., 2009) are manually curated databases for miRNA-target interactions with experimentally tested resources derived from the literature for different species including mouse and human.
StarBase utilizes high-throughput CLIP-Seq and degradome sequencing methods to identify the sites of Argonaute interaction and miRNA cleavage sites to detect the miRNA target interactions for six organisms including mouse and human (Yang et al., 2011; Li et al., 2014a).

miRNA-miRNA interaction (miRNA → miRNA)
PmmR infers miRNA-miRNA interactions for human from the induced miRNA-TF regulatory network that was constructed by combining all possible regulations between miRNAs and TFs. Topological analysis of the network reveals many TF modules that are enriched in different functional categories. Many of the identified miRNAs modules are significantly associated with common diseases. A putative miRNA inter-regulatory network is derived from the induced regulatory miRNA-TF network (Sengupta and Bandyopadhyay, 2011).

gene-gene interaction
Mentha integrates manually curated experimental detection of physical protein-protein interactions data avoiding genetic and inferred interactions for many species including mouse and human (Calderone et al., 2013). In contrast, STRING is a database that integrates information extracted by text mining and prediction methods. The database contains functional links between proteins that are inferred from genomic associations in form of gene-gene interactions. In addition to direct interactions, STRING takes into account the functionally associated interactions whose genes were regulated at the same time. The third group are genes whose protein products aggregate in the cell to form protein complexes. STRING integrates the three types of interactions and assigns a confidence score to each interaction indicating the significance of the interaction predictions (von Mering et al., 2003; Szklarczyk et al., 2017).

1.5 Topological Measures in Disease-specific Networks

Identification of disease genes among lots of candidates is a very challenging task. A disease network is defined by a set of potential disease genes that are connected with each other to make a specific disease while carrying known phenotypes of the disease (Goh et al., 2007). Although the effect of hub-degree genes in disease networks is considerable, a large number of disease genes are not among the hubs (Goh et al., 2007; Liu et al., 2011; Nazarieh et al., 2016). Therefore a prioritization method which identifies the most promising disease genes with considering the disease network and giving priority to both hubs as well as non-hubs is very necessary. This idea is the base of our TopControl method which we describe later in chapter 5.

1.5.1 Disease databases: DisGeNET, HMDD

DisGeNET (Bauer-Mehren et al., 2010; Piñero et al., 2017) is a Cytoscape plugin that provides access to a gene-disease database. The database is built by integrating several public databases. It enables the user to query and analyze the gene-disease networks for human. The gene-disease network is represented by undirected bipartite graphs of two disjoint independent sets of genes and diseases, respectively. DisGeNET allows multiple edges between two set of nodes representing different source types. The networks are constructed from the bipartite graph by connecting the nodes via edges if the two genes or diseases share a disease or gene in the bipartite graph.
The human miRNA disease database (HMDD) provides manually collected miRNA-disease associations from the literature. Similar to DisGeNET, a miRNA-disease network is represented by a bipartite graph that connects two sets of nodes that represent miRNAs and diseases. The network is constructed from the bipartite graph by connecting two disease nodes with an edge if the respective two diseases share at least one common associated miRNA (Lu et al., 2008; Li et al., 2014b).

1.5.2 Centrality Measures

Centrality is a measure of importance of a node in a network. There are several ways to measure centrality in regulatory networks. Degree centrality in PPI networks describes the number of interactions that a protein has. In case of GRNs which are modeled with directed graphs, degree centrality can refer to either indegree (the number of incoming edges) or outdegree (the number of outgoing edges) or both of them, see equation 1.1.

$$C_{degree}(u) = deg(u) \tag{1.1}$$

Betweenness centrality describes the number of times that a gene acts as a connector along the shortest path between any pair of nodes in the network as shown in equation 1.2. $\sigma_{st}(u)$ stands for the number of times that node u stands on the shortest path from node s to node t and σ_{st} shows the number of shortest paths between nodes s and t.

$$C_{betweenness}(u) = \sum_{s \neq u \neq t \in V} \frac{\sigma_{st}(u)}{\sigma_{st}} \tag{1.2}$$

Closeness centrality in the connected network describes the inverse of the sum of the distances denoted by $d(u, v)$ based on the shortest path between node u and all other nodes in the network, see equation 1.3.

$$C_{closeness}(u) = \frac{1}{\sum_v d(u, v)} \tag{1.3}$$

Chapter 2

Computational Methods

Although deterministic algorithms with high speed are very demanding for the complex structure of computational systems, biological systems require simple interpretable models. The corresponding algorithms need to make a trade-off between speed, robustness and accuracy (Navlakha and Bar-Joseph, 2014). The robustness of the biological algorithms is affected by the topology of biological networks. Dense topologies like cliques are preferred to sparser topologies for the networks with little noise (Milo et al., 2002; Yu et al., 2007). We discuss in chapter 3 more in detail about these issues.
This chapter begins with an explanation of optimization problems which I used for the mathematical modeling of topological features in regulatory networks. Then I describe the network motifs and the recent perspectives for considering these network modules. I explain differentially expressed (DE) genes as I took them into account as potential candidates for detecting disease-associated genes.

2.1 Combinatorial Optimization Problems

An optimization problem is a problem of minimizing (or maximizing) a function (called the objective function) given a set of constraints. In optimization problems, we search for the best solution among all the feasible solutions. An optimization problem can be written in this form:

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && g_i(x) \leq 0, i = 1, \ldots, m \\ &&& h_i(x) = 0, i = 1, \ldots, p \end{aligned} \tag{2.1}$$

where $f(x) : R^n \rightarrow R$ is the objective function to be minimized over the variable x, $g_i(x) \leq 0$ are called inequality constraints, and $h_i(x) = 0$ are called equality constraints. A maximization problem can be turned into a minimization problem by negating the objective function $-f(x)$. If the objective function is a linear function of the variables and the equality and inequality constraints are also linear, it is called linear problem. In an integer linear program (ILP), the objective function and the constraints are linear and all the variables are restricted to be integers. In a mixed integer linear program (MILP) some variables are restricted to be integer. If a problem is solvable in polynomial time in the

worst case, then it is contained in the class P. The class NP contains all problems that are solvable in polynomial time by a non-deterministic Turing machine (Cormen et al., 2003). The solution of these problem are verifiable in polynomial time. Optimization problems relate to decision problems, where a given optimization problem can be considered also as a decision problem (answer yes or no) by setting bound on the optimal value (Cormen et al., 2003). A decision problem is NP-hard if every problem in class NP can reduce to it by polynomial time reduction (Gross and Yellen, 2005). A problem is NP-complete if it is NP-hard and also contained in NP.

NP-complete Problems

Since the algorithms which solve NP problems can take a long time to execute, other types of algorithms exist which search for a close to optimal solution in shorter time.

- Approximation: The cost of the solution is within a factor of optimal.
- Fixed parameter tractability: Such an algorithm solves the problem more quickly if certain parameters are fixed than in general case.
- Heuristic: Such an algorithm returns a reasonably good result in a short time.
- Randomization: Such an algorithm reduces the running time using random numbers.

2.2 Network Motifs

Network motifs are recurring patterns of interactions between a predefined number of elements in regulatory networks which occur at higher frequencies in the real biological network than in random networks with the same size and connectivity characteristics (Alon, 2007). Regulatory motifs have been found in a variety of organisms from bacteria to human. The importance of network motifs is principally due to their biological function. Autoregulation motifs where a TF activates or represses the transcription of its own gene modulate the speed of response to biological signals by facilitating the synthesis of the amount of proteins that is required at the appropriate time. Positive autoregulation motifs speed up the process and increase cell-cell variation in protein levels whereas negative autoregulation motifs slow down the process and decrease cell-cell variation in protein levels (Alon, 2007). In contrast to autoregulation motifs, feedforward loops (FFLs) comprise of at least two regulators and one target gene which is regulated by both regulators, see Figure 2.1. Since the regulatory interactions correspond to upregulator and downregulator regulatory factors, therefore eight FFLs exist in the transcriptional regulatory networks (Alon, 2007).

In case of two target genes for the regulators, FFLs of size four are generated. Composite FFLs, in which two regulators regulate each other were found in developmental networks. Cascade motifs in which the target gene is activated by a regulator in which the regulator itself is regulated by another regulator are useful for passing and processing the information (Alon, 2007).
The computational process of network motif detection typically consists of three steps. The first step is to detect all predefined structural subgraphs in the network. To find the overrepresented subgraphs which occur in real network, more frequently than in random

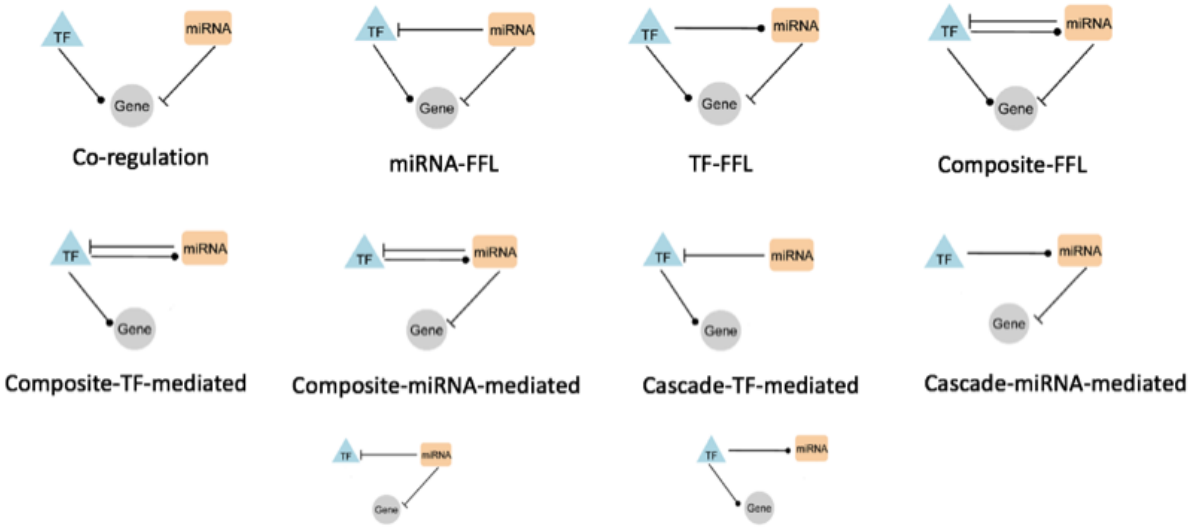

Figure 2.1: 3-node motifs in miRNA & TF synergistic regulatory networks. In FFLs the gene is regulated via two paths: (1) a direct regulation by a main regulator (TF/miRNA) and (2) an indirect regulation through an intermediate regulator (miRNA/TF) which is itself regulated by the main regulator. Composite-TF/miRNA-mediated: mutual regulation of TF and miRNA besides regulation of the target gene by only one of them. Cascade-TF/miRNA-mediated: are non-loop forms, including an indirect effect of the main regulator (TF/miRNA) on the target gene only via another type of regulator (miRNA/TF). The 3rd row shows two non-cooperative motifs where the target gene is not cooperatively regulated. The figure is from (Sadegh et al., 2017).

networks, a large number of randomized networks need to be generated. The randomization strategy affects the significance of the motifs (Sadegh et al., 2017). p-value and z-score are the statistical tests which may be used to indicate the significance of the findings.

A novel challenge in this area is to identify network motifs underlying cellular identity (Liang et al., 2015). (Megraw et al., 2013) detected sustained input switches among TFs and miRNAs that constitute composite FFLs in TF-miRNA co-regulatory networks in Arabidopsis network. When two regulators are activators, then the gene circuit ensures stable downstream gene expression and functions and as a noise repression role when both regulators are repressors (Megraw et al., 2013). To identify the network motifs that underlie cellular identity, we considered the network motifs that involve a considerable number of master regulators and key drivers detected by MCDS method (Sadegh et al., 2017).

The meaningfulness of the network motifs relies on the strategy which generates randomized networks. Reliable results come from randomized networks that do not generate a certain random network more often (Megraw et al., 2013). Variances of count distributions of subgraphs in randomized networks are an indicator for inferring uniform sampling. The best randomization strategy is the one that is capable of detecting a few significant motifs with high biological relevance. Fast network motif detection (FANMOD) method applies the edge-switching method with considering both edge-conserving and non-conserving variations maintains the in/out degree of the nodes during the randomization. Edge switching works by selecting two edges from the biological network, swap their endpoints and repeat this process for a predefined number of times. In case of the conserving approach, edge switching is performed on edges of same type of interaction. The non-conserving method has no obligation regarding the choice of edges. The number of required edge swaps is described by $Q * L$, where L is the number of edges and Q is the number of iterations. Some tools like TFmiR (Hamed et al., 2015a) suggest a default number for the number of iterations. (Liang et al., 2015) suggested maximum difference between a given network and each of the randomized networks. To avoid under-shuffling, we suggested the number of edge swaps should satisfy minimum similarity between the biological network and each of the randomized networks (Sadegh et al., 2017). The goal in both the approaches is to generate randomized networks which are fully shuffled (Liang et al., 2015; Sadegh et al., 2017). Unlike the edge-switching method which swaps the edges with potential of under-shuffling and over-shuffling, weighted and reverse swap (WaRSwap) method generates randomized networks without replacement (Megraw et al., 2013). This way of randomization creates a trade-off between uniform sampling and speed. To generate randomized networks, WaRSwap breaks the biological network into multiple layers corresponding to different edge types. For each layer, it applies the WaRSwap algorithm which preserves the target indegree distribution (Megraw et al., 2013). The algorithms consists of mainly three steps, see Figure 2.2. 1) it sorts the source nodes in each layer based on their out-degrees in descending order. For each source node S_i, it computes attraction weight for each target T_j. 2) it matches the source node S_i to target nodes T_j using edges if they are weighted proportionally to sampling weights if the degree of S_i is less than the number of unsaturated target nodes. 3) if the degree of S_i is greater than the number of unsaturated target nodes, then it matches the source node to each target that has available capacity. In case of unplaced edges, it performs swapping (Megraw et al., 2013).

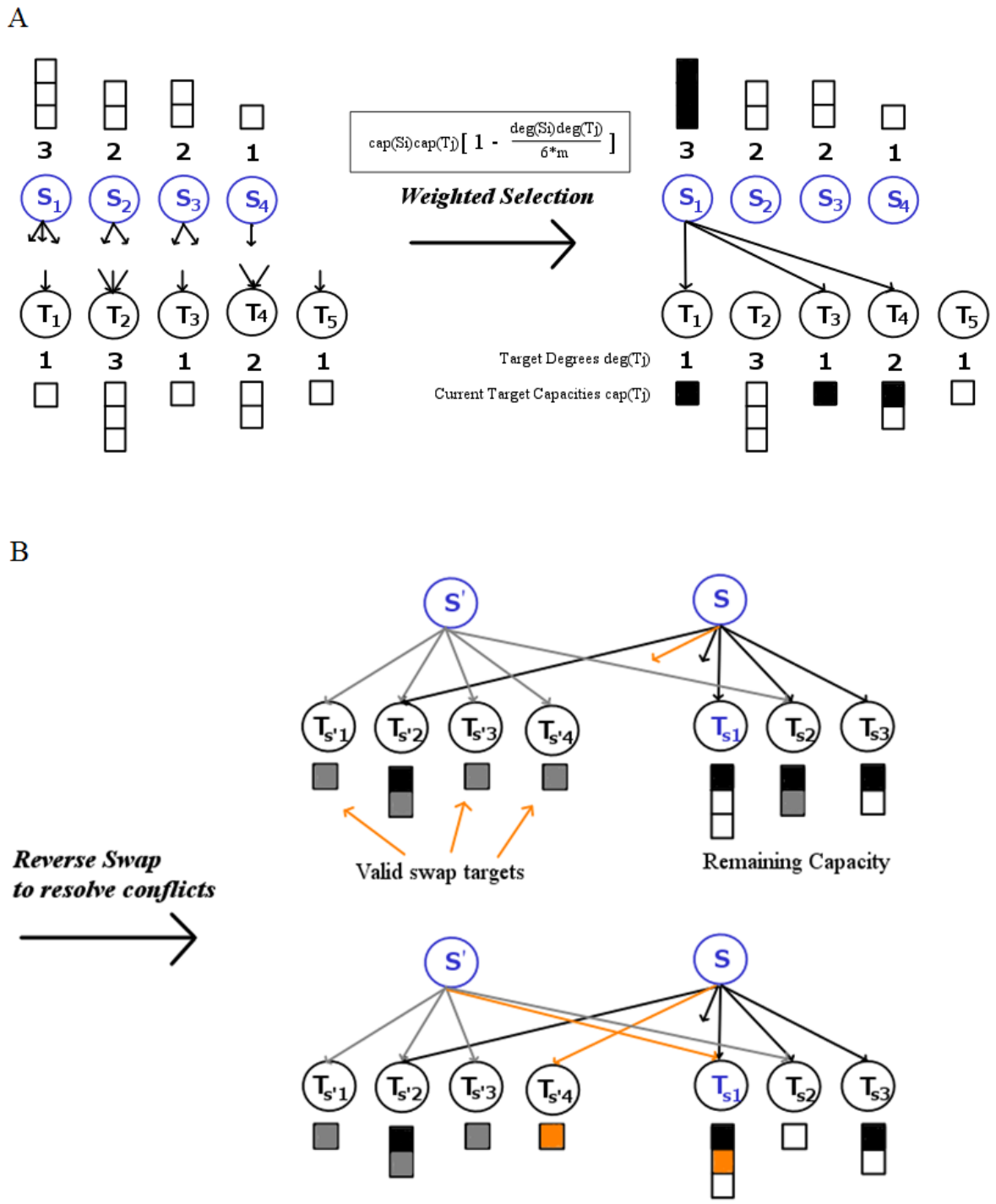

Figure 2.2: A graphical description of WaRSwap (Weighted-and-Reverse-Swap) algorithm. The algorithms comprises of two main strategies A) weighted selection B) reverse swap to resolve conflicts. The figure is from (Megraw et al., 2013).

2.3 Identification of DE Genes

There are mainly two ways to measure the expression of a gene using the RNA-Seq technology. Read counts refer to the number of reads mapping to gene segments in the DNA sequence. The data are not comparable across the samples, due to different sequencing depth, total number of reads, and sequencing biases (Conesa et al., 2016). Therefore normalization of the data would be the first step before any downstream analysis. In contrast, measures which remove the dependency of the data on the library size and gene length such as read per kilobase of exon model per million reads (RPKM) and fragments per kilobase of exon model per million mapped reads (FPKM) and transcript per million (TPM) are most commonly used for the comparison of gene expression values across different conditions (Conesa et al., 2016). In this dissertation, we focused on the read count data for the purpose of inferring DE genes. The main reason for this was that the TCGA data are available in the form of read counts. We exploited the four differential gene expression analysis methods DESeq, edgeR, voom and VST which take read counts as input and return p-values as an indicator of statistically significant DE gene. The above-mentioned methods use DESeq and TMM normalization methods.
To identify DE genes, which are transcripts with different abundances between two samples, differential gene expression analysis methods get RNA-Seq reads from two different samples and transcript sequences. To identify statistically significant DE genes, a statical test is necessary which is based on modelling of the data distribution.
Since RNA-Seq data have discrete values, it is not possible to model the data with a normal distribution. The Poisson distribution has just one parameter, μ which does not allow for dealing with overdispersed data. Thus, it has been proposed to model RNA-Seq count data with a negative binomial distribution (NB) with the two parameters mean μ and variance σ^2 as follows $K_{ij} \sim NB(\mu_{ij}, \sigma^2_{ij})$, where μ_{ij} is the mean of the gene i in sample j and variance is $\sigma^2_{ij} = \mu_{ij} + \mu_{ij}\phi_{ij}$ where ϕ_{ij} is the dispersion parameter, controlling the overdispersion level (Soneson and Delorenzi, 2013).

Poisson Distribution
Count data are modeled naturally by the Poisson distribution (Cameron and Trivedi, 2007). The Poisson distribution is described by the following formula where $E[K] = \mu$ and $V[K] = \mu$ which imposes that mean and variance are equal.

$$Pr(K = k) = \frac{e^{-\mu} * \mu^k}{k!} \tag{2.2}$$

Negative Binomial Distribution
Sampling data from an appropriate Poisson distribution is challenging due to the problem of accurately estimating μ (Cameron and Trivedi, 2007). Moreover, there is overdispersion in count data due to unobserved heterogeneity which makes the model very restrictive to model the read count (Cameron and Trivedi, 2007). An integer-valued random variable K follows a NB distribution with parameters $p \in (0, 1)$ and $r \in (0, \infty)$ if it follows equation

2.3 as mentioned in (Cameron and Trivedi, 1998; Anders and Huber, 2010).

$$Pr(K = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k \tag{2.3a}$$

$$p = \frac{\mu}{\sigma^2} \tag{2.3b}$$

$$r = \frac{\mu^2}{\sigma^2 - \mu} \tag{2.3c}$$

2.3.1 Differential Gene Expression Analysis Methods

A gene with observed difference in read counts between two conditions is termed a DE gene. To decide whether, for a given gene, an observed difference in read counts is statistically significant and did not occur due to random chance, a p-value is used. To find the set of DE genes from RNA-Seq data, mean and variance need to be estimated for distribution of expression values for each gene. But since the number of samples is often too small, it is difficult to measure these parameters reliably.

DESeq

DESeq is a DE analysis method which takes the read count data as input. The data are given as an $n * m$ table of counts k_{ij}, where $i = 1, \ldots, n$ relate to indexes of the genes, and $j = 1, \ldots, m$ relate to indexes of the samples. DESeq models count data with a negative binomial distribution (NB) $K_{ij} \sim NB(\mu_{ij}, \sigma^2_{ij})$, where K_{ij} is the number of reads in sample j that are mapped to gene i with two parameters of mean μ_{ij} and variance σ^2_{ij}. μ_{ij} is proportional to the mean of the of read counts of gene i under experimental condition $\rho(j)$ and size factor s_j as denoted by $\mu_{ij} = q_{i,\rho(j)} s_j$. The size factor s_j denotes sampling depth of sample j. The variance denoted by $\sigma^2_{ij} = \mu_{ij} + s^2_j \nu_{i,\rho(j)}$ shows a linear relationship between the two parameters of the NB model. With respect to the low number of replicates which does not lead to a reliable estimate of the variance for gene i from just the data available for this gene, DESeq assumes that the per-gene raw variance parameter $\nu_{i,\rho(j)} = \nu_\rho(q_{i,\rho(j)})$ is a smooth function of q_i and ρ (Anders and Huber, 2010). To derive the set of statistically significant DE genes from the data, DESeq estimates mean and variance for each gene separately. It applies a normalization step ahead, since the samples have been generated with different sequencing depth. DESeq calculates a size factor for m times related to m samples in the data. The size factor is estimated by taking the median of the ratios of read counts 2.4.

$$\hat{s}_j = \text{median}_{(i)} \frac{k_{ij}}{(\prod_{\nu=1}^{m} k_{i\nu})^{\frac{1}{m}}} \tag{2.4}$$

For each experimental condition ρ, there are n read counts $q_{i\rho}$. They reflect the read counts for gene i under condition ρ. That is, the mean of gene i is proportional to $q_{i\rho}$ as illustrated below in equation 2.5, where m_ρ is the number of samples of condition ρ and the sum runs over these samples.

$$\hat{q}_{i\rho} = \frac{1}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{k_{ij}}{\hat{s}_j} \tag{2.5}$$

To calculate variance ν_ρ, DESeq calculates at first sample variances on the common scale as in equation 2.6.

$$w_{i\rho} = \frac{1}{m_\rho - 1} \sum_{j:\rho(j)=\rho} (\frac{k_{ij}}{\hat{s}_j} - \hat{q}_{i\rho})^2 \tag{2.6}$$

and defines $z_{i\rho} = \frac{\hat{q}_{i\rho}}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{1}{\hat{s}_j}$. It was shown that $w_{i\rho} - z_{i\rho}$ is an unbiased estimator for the raw variance parameter $\nu_{i\rho}$ (Anders and Huber, 2010).
DESeq uses a test statistic similar to Fisher's exact test (nbinomTest) to obtain statistically significant results. The null hypothesis is $q_{iA} = q_{iB}$, where q_{iA} is the read count for the samples of condition A and q_{iB} for condition B. To this end, (Anders and Huber, 2010) define the total counts in each condition, $K_{iA} = \sum_{j:\rho(j)=A} k_{ij}$, $K_{iB} = \sum_{j:\rho(j)=B} k_{ij}$ and their overall sum $K_{iS} = K_{iA} + K_{iB}$. Then it uses any pairs (a, b), where $K_{iA} = a$ and $K_{iB} = b$ and $a + b = k_{iS}$ to calculate the p-value. The p-value of a pair of observed count sums (k_{iA}, k_{iB}) is then the sum of all probabilities less or equal to $p(k_{iA}, k_{iB})$, given that the overall sum is k_{iS} (Anders and Huber, 2010). The probability is calculated using the formula 2.7.

$$p_i = \frac{\sum_{a+b=k_{iS}, p(a,b) \leq p(k_{iA},k_{iB})} p(a, b)}{\sum_{a+b=k_{iS}} p(a, b)} \tag{2.7}$$

VST + limma

The Variance Stabilizing Transformation (VST) takes the variance-mean dependence $w(q)$ computed by DESeq and applies a transformation function 2.8 to remove the dependency. The monotonous mapping function produces data whose variance is independent from the mean (Anders and Huber, 2010).

$$\tau(K) = \int^K \frac{dq}{\sqrt{w(q)}} \tag{2.8}$$

VST uses the limma package for performing the statistical tests and inferring the set of DE genes.

edgeR

edgeR assumes a NB distribution for modelling the count data (Robinson et al., 2010). Before the analysis, TMM normalization (Robinson and Oshlack, 2010) is applied to the data. To calculate μ_{ij}, edgeR considers total number of reads in sample j and the relative abundance of gene i in the sample (Robinson et al., 2010). edgeR estimates common dispersion before gene-wise dispersion by accounting the values of all genes. Gene-wise dispersion is estimated from common dispersion ϕ as shown in equation 2.9 in (Robinson and Smyth, 2007, 2008). Gene-wise dispersion gives a unique dispersion value to each gene. edgeR estimates the gene-wise dispersion by conditional maximal likelihood, conditioning on the total count $z_i = \sum_{j=1}^{n_i} Y_{ij}$, where n_i is the total number of samples for gene i and y_{ij} denotes the observed count for gene i in sample j.

$$l_g(\phi) = \sum_{j=1}^{2} (\sum_{j=i}^{n_i} log\Gamma(y_{ij} + \phi^{-1}) + log\Gamma(n_i\phi^{-1}) - log\Gamma(z_i + n_i\phi^{-1}) - n_i log\Gamma(\phi^{-1})) \tag{2.9}$$

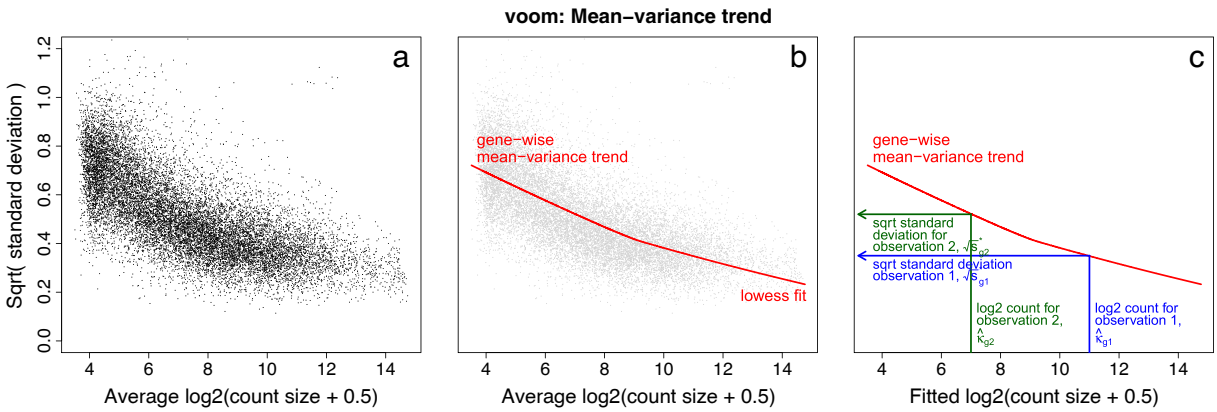

Figure 2.3: Mean-variance modelling of voom (Law et al., 2014).

Gene-wise dispersions is shrunk towards a common dispersion using an empirical Bayes procedure. The common dispersion estimator maximizes the common likelihood $l_C = \sum_{g=1}^{G} l_g(\phi)$ where G is the number of genes (Robinson and Smyth, 2008). Finally, differential expression is assessed for each gene using an exact test similar to Fisher's exact test, but adapted for data that have overdispersion (Robinson and Smyth, 2008; Robinson et al., 2010).

voom+limma

voom (variance modelling at the observation level) attempts to estimate the mean-variance relationship robustly and without any parameter from data at the level of individual observations. It transforms count data to log-cpm(counts per million) values for the purpose of normalization (Law et al., 2014). This transformation is performed because read counts show increasing variance with increasing count size, while log-counts typically show a decreasing mean-variance trend. To estimate the mean-variance trend at level of individual observation, it computes a residual standard deviation for each gene, See Figure 2.3 panel (a). After fitting a robust trend to the residual standard deviations, see Figure 2.3 panel (b), standard deviation for an individual observation is predicted by interpolating the standard deviation trend based on its predicted count size, see Figure 2.3 panel (c). Finally, the inverse square of the predicted standard deviation for each observation and log-cpm values are given to limma's standard differentiation pipeline as input to obtain the set of statistically significant DE genes (Law et al., 2014).

2.4 Applied Programming Languages

I used a variety of programming languages to develop softwares, process and analyze the data and develop web servers including R, Java, Perl, PHP, JavaScript and SageMath.

- SageMath: it is a free open-source software system under the terms of GNU general public license which was developed initially by William Stein, at the University of Washington (The Sage Developers, 2015). The software uses many open-source packages such as NumPy, SciPy. It is accessible either through Python language

or directly through interfaces of wrappers. I implemented the ILP formulations of minimum dominating set and minimum connected dominating set in SageMath.

- Java: it is a general-purpose class-based and object oriented language. The platform independent language lets the Java programs run on any combination of software and hardware systems because of the java virtual machine. I implemented the heuristic approximate approaches of minimum dominating set and minimum connected dominating set in Java.

- R: it is a free programming language and software environment for statistical computing and graphics which is widely used for analysis of large data sets. I implemented the backend programs of TFmiR2 web server in R. Moreover, I used the language for TCGA data analysis.

- Perl: it is a scripting language facilitating easy manipulation of text files and defining and searching string patterns. I used these facilities to download the YPA data from the web and to process the files.

- PHP: it is a server-side scripting language mainly designed for web development and usually embedded into HTML. I implemented the front-end of the TFmiR2 web server (server-side) in PHP.

- JavaScript: it is a client-side scripting language developed to run in web browsers which is embedded usually with HTML and CSS. I implemented the front-end of TFmiR2 web server (client-side) in JavaScript.

Chapter 3

Identification of Key Regulatory Genes in Gene Regulatory Networks

This chapter is based on our paper entitled "Identification of Key Regulatory Genes in Gene Regulatory Networks " by Maryam Nazarieh, Andreas Wiese, Thorsten Will, Mohamed Hamed and Volkhard Helms published in the journal of BMC Systems Biology (Nazarieh et al., 2016).
Maryam Nazarieh mapped the problem to minimum dominating set and minimum connected dominating set, designed and implemented the algorithms, performed data analysis and wrote the manuscript. Dr. Andreas Wiese extended the established the ILP formulations and edited the manuscript. Thorsten Will developed the Cytoscape plugin for the heuristic approach of MCDS and improved the implementation of the algorithm. Mohamed Hamed assisted with the functional enrichment analysis. Prof. Volkhard Helms proposed the biological motivation of the paper and helped with designing the study, data analysis and the manuscript.
The initial results of the mathematical modelling was presented at the German Stem Cell Network conference in Nov.2014 in Heidelberg http://www.gscn.org/Conferences/2014/Program.aspx.
Identifying the gene regulatory networks governing the workings and identity of cells is one of the main challenges in understanding processes such as cellular differentiation, reprogramming or cancerogenesis. One particular challenge is to identify the main drivers and master regulatory genes that control such cell fate transitions. In this work, we reformulate this problem in terms of the optimization problems of computing a Minimum Dominating Set (MDS) and a Minimum Connected Dominating Set (MCDS) for directed graphs.
Both MDS and MCDS are applied to the well-studied gene regulatory networks of the model organisms *E. coli* and *S. cerevisiae* and to a pluripotency network for mouse ESCs. The results show that a MCDS can capture most of the known key player genes identified so far in the model organisms. Moreover, this method suggests an additional small set of TFs as novel key players for governing the cell-specific gene regulatory network which can also be investigated with regard to diseases. To this aim, we investigated the ability of MCDS to define key drivers in breast cancer. The method identified many known drug targets as members of the MDS and MCDS.

The Java implementation of the heuristic algorithm explained in this chapter is available as a Cytoscape plugin at http://apps.cytoscape.org/apps/mcds. The SageMath programs for solving integer linear programming formulations used in the paper are available at
https://github.com/maryamNazarieh/KeyRegulatoryGenes and as supplementary material.

3.1 Background

Although all the cells in multicellular organisms basically share the same DNA sequence with the same set of genes, in each cell type only a particular set of genes is actively expressed which then defines its specific morphology and function. Thus, different types of cells are controlled by different sets of active genes and by the interactions between them (Chen and Rajewsky, 2007; Bossi and Lehner, 2009; Neph et al., 2012; Will and Helms, 2014). Inside each cell, a set of target genes and regulatory genes, namely the TFs, interact with each other and form a gene regulatory network (GRN). Typically, GRNs topologically comprise a highly connected component and a few nodes with low connectivity (Barabasi and Oltvai, 2004). ESCs, for example, can be distinguished from other cells mainly based on their pluripotency network. This network in ESCs is spanned up by few connected TFs which share many target genes (Kim et al., 2008). A slight change in the expression levels of such a tightly interwoven network of TFs leads to ESC differentiation (Kim et al., 2008).
Of particular interest are the groups of key driver genes and master regulatory genes in condition-specific and unspecific gene regulatory networks. Key driver genes are basically those genes that control the state of the network (Liu et al., 2011; Zhang et al., 2015b; Hamed et al., 2015b). The term master regulatory gene was introduced by Susumu Ohno over 30 years ago. According to his definition, a master regulator is a gene which stands at the top of a regulatory hierarchy and is not regulated by any other gene (Ohno, 1979). Later on, this term was redefined to involve a set of genes which either directly govern the particular cellular identity or are at the inception of developmental lineages and regulate the expression of a cascade of genes to form specific lineages (Ohno, 1979).
To address the problem of computational identification of key and master regulatory genes, we have used the notion of network controllability (Liu et al., 2011) in terms of dominating set, modeled and solved two optimization problems named Minimum Dominating Set (MDS) and Minimum Connected Dominating Set (MCDS) on the GRNs. We compared these sets against well-known centrality measures such as degree, betweenness and closeness centrality as described in (Freeman, 1978). These attribute the importance of genes to their centrality in the networks. However, it is unclear whether high centrality genes exercise a full control over the underlying network.
A recent study derived a minimum input theorem based on structural control theory which can be applied to directed graphs to fully control the network (Liu et al., 2011). For this purpose, the authors introduced a deep relation between structural controllability and maximum matching. The idea is to control the whole network by covering all the regulatory interactions with a minimum number of genes. Their results show that a few nodes are sufficient to control dense and homogeneous networks, but this number increases dramatically when the nodes in the network are sparsely connected.
An MDS is a related concept in which the goal is to control the network by covering all

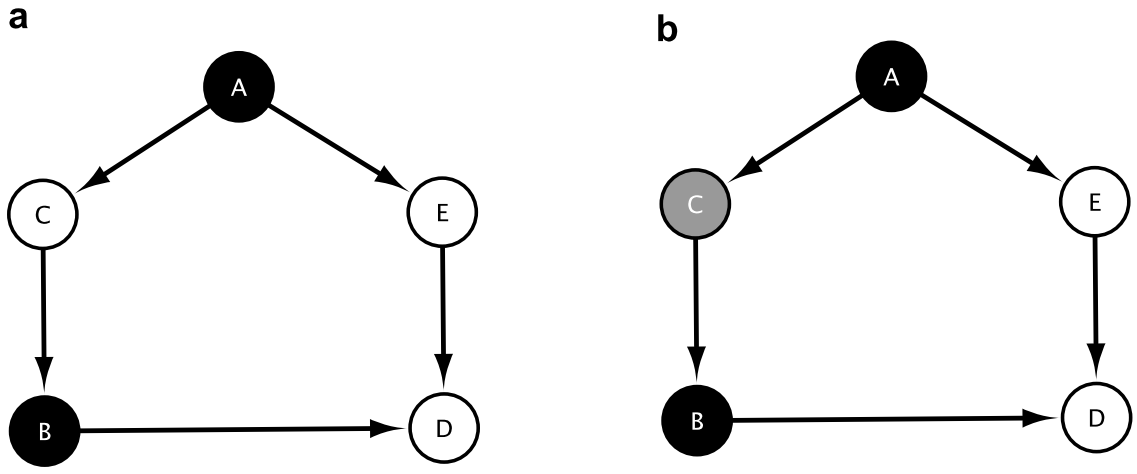

Figure 3.1: A graphical representation that illustrates the MDS and MCDS solutions of an example network. The network can be controlled by MDS and MCDS nodes. In the case of a GRN, directed arcs symbolize that a TF regulates a target gene. In panel (a), the MDS nodes {A,B} are the dominators of the network. Together, they regulate all other nodes of the network (C, E, D). Panel (b) visualizes the respective set of MCDS nodes (black and gray). Here, node C is added in order to preserve the connection between the two dominators A and B to form an MCDS.

expressed genes with a minimum number of TFs. Since each node that does not belong to the MDS is adjacent to at least one node in the MDS, full control over the network is provided by the MDS solution. Our group has previously applied the concept of MDS to the area of complex diseases. The results showed that this method can capture several important disease and drug target genes (Hamed et al., 2015b,a). The MDS method can be applied to any connected or disconnected regulatory network to identify key dominator nodes. In this work, we use MDS in directed graphs to identify key driver genes. Besides the MDS concept, we suggest to also consider the task of identifying a set of master regulatory genes in terms of another optimization problem, namely that of constructing an MCDS. We suggest to apply MCDS mainly to networks that are related to cell fate transitions such as the pluripotency network of an ESC. This idea is motivated by the observation that the pluripotency network in mouse ESCs is maintained by a few connected TFs which share many target genes (Kim et al., 2008). The concepts of MDS and MCDS are visualized for a small toy network in Figure 3.1.

The concept of MCDS has already been applied to protein-protein interaction networks (which are represented by undirected graphs). There, the proteins which compose a MCDS solution contributed significantly to related biological processes (Milenković et al., 2011). In this work, we show how the MCDS concept can be applied to GRNs (represented by directed graphs) to detect the TFs and target genes which determine a specific cellular identity. We start with the model organisms *E. coli* and *S. cerevisiae* because their GRNs have been extensively characterized in experimental studies. Then, we present applications to a mouse pluripotency network and to a breast cancer regulatory network.

3.2 Methods

3.2.1 Minimum Dominating Set

A dominating set (DS) in an undirected graph $G = (V, E)$ is a subset of nodes $D \subseteq V$ with the property that for each node $v \in V$ we have that $v \in D$ or that there is a node $u \in D$ and an edge $\{u, v\} \in E$. We call a set $D \subseteq V$ a minimum dominating set (MDS) if it is a dominating set and it has minimum cardinality among all dominating sets for G. Computing a MDS is known to be an NP-complete problem (Garey and Johnson, 1979). In biological networks, the set of dominators can provide full control over the whole network. Since each node that does not belong to the MDS is at least adjacent to one node in the MDS, full control over the network can be obtained by the MDS solution. To address GRNs which are represented by directed graphs, we define an MDS for a directed graph $G = (V, E)$ to be a set $D \subseteq V$ of minimum cardinality such that for each node $v \in V$ we have that $v \in D$ or that there is a node $u \in D$ and an arc $(u, v) \in E$. The integer linear programming (ILP) formulation of MDS for directed graphs is given below. Here, for each node $v \in V$ we denote by $\delta^-(v)$ the set of incoming nodes of v, i.e., the set of nodes u such that $(u, v) \in E$.

$$\begin{aligned} \text{minimize} \quad & \sum_{v \in V} x_v \\ \text{subject to} \quad & x_u + \sum_{v \in \delta^-(u)} x_v \geq 1 \quad \forall u \in V \\ & x_v \in \{0, 1\} \qquad\quad \forall v \in V \end{aligned} \tag{3.1}$$

Here, variables x_u and x_v are binary variables associated to the nodes u and v in the graph. Using this formulation, we select a node v as a dominator if its binary variable x_v has value 1 in the computed solution and otherwise we do not select it. Since our objective function is to minimize $\sum_{v \in V} x_v$ this yields a minimum dominating set. For all networks considered here, MDS solutions were constructed in less than 1 minute of running time.

3.2.2 Minimum Connected Dominating Set

A minimum connected dominating set (MCDS) for a directed graph $G = (V, E)$ is a set of nodes $D \subseteq V$ of minimum cardinality that is a dominating set and that additionally has the property that the graph $G[D]$ induced by D is weakly connected, i.e., such that in the underlying undirected graph between any two nodes $v, v' \in D$ there is a path using only vertices in D. Computing an optimal MCDS in undirected graphs is known to be NP-hard (Garey and Johnson, 1979). Since GRNs are represented by directed graphs, we are interested in MCDSs for directed graphs.

Optimal Solution via ILP

To this end, we modified the existing integer linear programming (ILP) formulation of MCDS in undirected graphs (Simonetti et al., 2011) to determine a MCDS for directed graphs. This work introduced a set of valid inequalities for the undirected graphs which can be modified to be used for the directed MCDS on the directed graphs (Simonetti et al., 2011).
As before, the set V is the set of vertices and E is the set of edges in the input graph.

For a set $S \subseteq V$, the set $E(S)$ stands for all the edges connecting two vertices u, v with $u, v \in S$. The binary valued y_v variables indicate whether node v is selected to belong to the minimum connected dominating set. The binary variables x_e for the edges then yield a tree that contains all selected vertices and no vertex that was not selected. Thus, the selected vertices form a connected component. The first constraint guarantees that the number of edges is one unit less than the number of nodes. This is necessary for them to form a (spanning) tree but is not sufficient. The second constraint guarantees that the selected edges imply a tree. The third constraint guarantees that the set of selected nodes in the solution forms a dominating set of the graph. For dense undirected graphs, this formulation provides a quick solution, but in the case of sparse graphs, finding the optimal solution may take considerable running time (Simonetti et al., 2011).

$$\begin{aligned}
\text{minimize} \quad & \sum_{v \in V} y_v \\
\text{subject to} \quad & \sum_{e \in E} x_e = \sum_{i \in V} y_i - 1 \\
& \sum_{e \in E(S)} x_e \leq \sum_{i \in S \setminus \{j\}} y_i \quad \forall S \subset V, \forall j \in S \\
& y_u + \sum_{v \in \delta^-(u)} y_v \geq 1 \qquad \forall u \in V \\
& y_v \in \{0, 1\} \qquad \forall v \in V \\
& x_e \in \{0, 1\} \qquad \forall e \in E
\end{aligned} \tag{3.2}$$

The above IP formulation contains an exponential number of constraints since it has one constraint for each subset $S \subseteq V$. Therefore, already for relatively small instances it is impractical to generate all its inequalities. Instead, we used the following approach: we generate the first constraint and all constraints of the third type (i.e., $\sum_{e \in E} x_e = \sum_{i \in V} y_i - 1$ and $y_u + \sum_{v \in \delta^-(u)} y_v \geq 1$ for each $u \in V$). Then we compute the optimal IP solution subject to these constraints. Then we check whether the found solution satisfies all constraints of the above IP (even those that we did not add to our formulation). This is the case if and only if the computed set of vertices yields a connected dominating set. If this is the case then we found the optimal solution and we stop. Otherwise, we add (violated) constraints of the second type (i.e., $\sum_{e \in E(S)} x_e \leq \sum_{i \in S \setminus \{j\}} y_i$ for some subset V and some node j) to our formulation and compute the optimal IP solution to this stronger formulation and repeat. If the computed set of vertices has more than one connected component then we add such a constraint for each connected component S and for each vertex $j \in S$. In order to improve the running time of our procedure, we added some valid inequalities to our initial formulation. These inequalities discard all the solutions that select an edge $e = \{u, v\}$ (i.e., $x_e = 1$) such that not both of its incident vertices were selected (i.e., not both $y_u = 1$ and $y_v = 1$). Formally, for each edge $e = \{u, v\}$ we added the inequalities

$$\begin{aligned}
x_e &\leq y_u \\
x_e &\leq y_v
\end{aligned} \tag{3.3}$$

Despite adding these valid inequalities, some problem instances were not solved in appropriate time. To overcome this problem, we also considered a heuristic approach. It

is known that an approximate MCDS in undirected graphs can be found by heuristic approaches in polynomial time (Wightman et al., 2011; Rai et al., 2009). For graphs with low number of nodes and high node degree, the optimal ILP solution can be found at comparable running times as such heuristic solutions (Wightman et al., 2011). However, the heuristic solution outperforms the ILP for graphs with high node density and low node degree in terms of running time (Wightman et al., 2011). In this work, all computations were conducted on a single threaded Intel XEON CPU at 2.2 Ghz. We determine the ILP solution using the glpk solver version 4.35 (Makhorin, 2008). In cases where the network is very sparse we used the heuristic algorithm (see next section).

Heuristic Solution

In this study, we computed the heuristic solution for all networks except for the modules of a breast cancer network. There, the optimal MCDS solution could be obtained within a few minutes to several hours of compute time. We adapted the heuristic algorithm presented in (Rai et al., 2009) that was inspired by one of the two general approximation approaches mentioned in (Guha and Khuller, 1998) to find solutions for MCDS. We modified the algorithm to determine a MCDS for directed graphs rather than an undirected graph. The algorithm has three main phases as described in the following. Initially, all nodes are white. In the first phase, a white node with the highest outdegree is selected as a dominator and colored black. In cases where multiple nodes have the same outdegree, we select the node with the highest indegree. This selection guarantees higher connectivity compared to nodes with smaller indegree. Its (directed) child neighbors are colored gray to indicate that they are already dominated. This step is repeated until all nodes are either black or gray. From these, we check if the (black) set of dominators forms a connected dominating set. If yes, we move to the third phase, otherwise we move to the second phase. In the second phase, a node with maximum number of arcs to black nodes, that we term a connector, is colored dark gray. This dark gray node is then added to the connected dominating set if it belongs to a path between two connected components that are not connected so far. This step is repeated until all black and dark gray nodes form a connected component in the underlying undirected graph. In the third phase, the size of the connected dominating set is reduced as much as possible by repeatedly removing a node with smallest outdegree while making sure that the dominating set remains connected and the graph remains covered by the connected dominating set. In cases where multiple nodes have the same outdegree, we again select the node with highest indegree. One can also interpret the algorithm biologically in the context of GRNs. We start by selecting a TF with the most target genes as a dominator. This process is repeated until all the genes are either selected as dominators or as target genes. If the dominating set is not connected, the next step is to connect the dominators by adding a few number of connector genes. This step is motivated by the modularity of cellular networks (Singh et al., 2008). We will investigate below whether defining a connected set of dominator nodes is beneficial for the biological interpretability of the control hierarchy. As connectors, we consider TFs as well as target genes. The last step is to reduce the size of the connected dominating set. Then, the connected dominating set comprises of dominators and connectors, whereby all dominators are TFs and the connectors comprise of TFs and/or target genes. Note that the set of MCDS identified as dominators or connectors provides potential candidates for key drivers and master regulatory genes.

For the networks considered here, the running time for the heuristic MCDS solution was less than 1 minute.

3.2.3 Components

Unlike MDS, the task of computing an MCDS only makes sense for input graphs that are connected since otherwise there can be no solution. Therefore, if we are given a disconnected undirected graph, we compute MCDSs for connected components of the graph. For directed graphs, we distinguish between strongly connected components and (weakly) connected components.

Strongly Connected Component

A component is called a strongly connected component (SCC) in a directed graph if each of its nodes is reachable via directed edges from every other node in the component. In a SCC, there is a path between each pair of nodes in the component. Here, we implemented Tarjan's algorithm to find SCCs as described in (Tarjan, 1972).

Largest Connected Component

A component is a (weakly) connected component if in the underlying undirected graph, there exists a path between any pair of nodes of this component. The connected component of highest cardinality is termed the largest connected component (LCC). The connected components were found by breadth first search (BFS) as described in (Hopcroft and Tarjan, 1973). Note that each strongly connected component is also a (weakly) connected component but the converse is not necessarily true. Since a MCDS does not exist in graphs that are not connected, we consider the LCC and the largest strongly connected component (LSCC) in such cases, see Figure 3.2. We compared the results of MCDS when the network has only one connected component to those obtained with a directed version of MDS in terms of the size of the result set and enrichment analysis.

Criteria to Select the Component

MDS is always applied to the whole network. If the input network is not connected, we select either LCC or the LSCC as the input for MCDS. If the cardinality of the network is equal to the LCC of the network, we select the whole network. Otherwise, we consider the component density of LCC (the number of other components were few with very small size for the considered GRNs in this study) and LSCC. For a directed graph $G = (V, E)$, the component density is defined as $\frac{|E|}{|V|(|V|-1)}$, where E denotes the set of edges and V denotes the set of nodes in the component. The component density is equal to the ratio of existing edges (interactions) $|E|$ in the component to the total number of possible edges (interactions). According to the definitions in (Sant, 2004), in a dense graph the number of edges is close to the maximal number of edges which is in contrast to a sparse graph. In this study, an MCDS is then derived for the component (LCC or LSCC) with highest density, as we were interested to find the minimum number of genes. High density components are more promising in this regard, because they need a smaller number of connectors to connect the dominators.

Enrichment Analysis

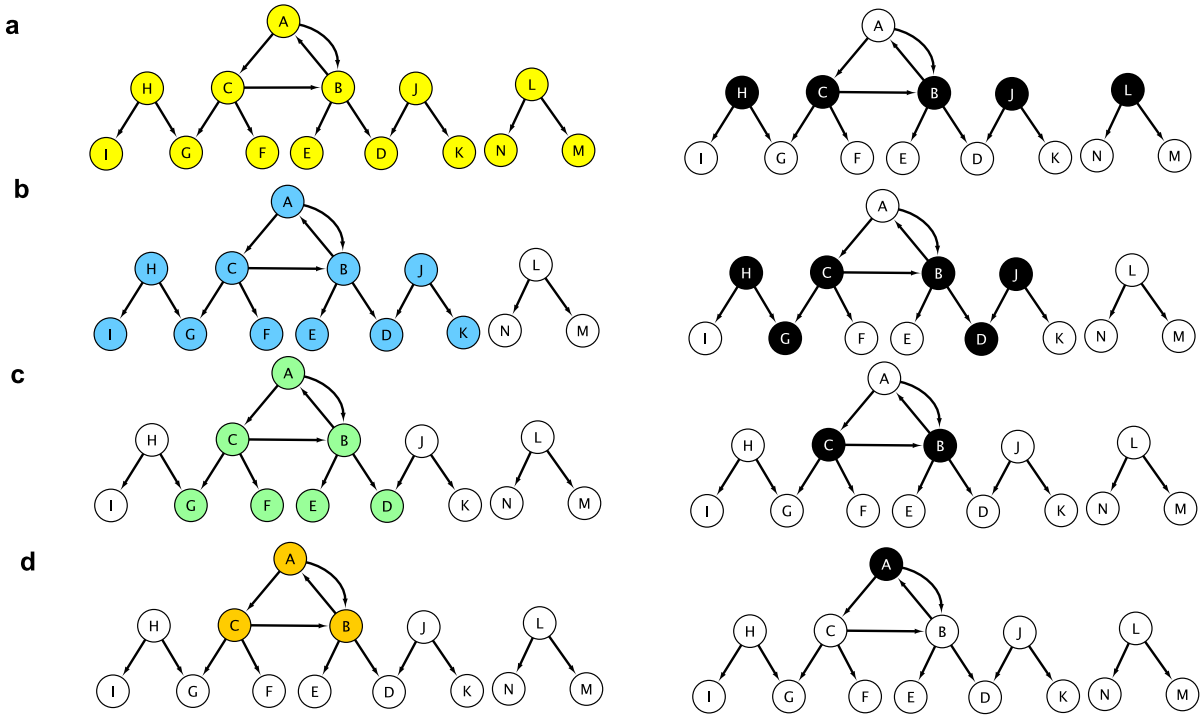

Figure 3.2: A graphical representation that illustrates the concept of MDS on a toy network. In addition, the MCDS nodes are colored black on three types of components (LSCC, LCC of the underlying directed graph and LCC of the underlying undirected graph) in the toy network. The above toy network includes 14 nodes and 14 edges as shown in yellow in panel (a). The nodes {J, B, C, H, L} are the dominators of the network obtained by computing a MDS (right panel). The nodes colored blue in panel (b), make up the LCC of the underlying undirected graph. MCDS nodes for this component are {J, D, B, C, G, H}. Green colored nodes in panel (c) are elements of the LCC underlying the directed graph. The two nodes {B, C} form the MCDS for this component. The nodes colored orange in panel (d) show the LSCC in the network. Here, the node A is the only element of the MCDS.

The biological relevance of the results obtained by the directed forms of MDS and MCDS was evaluated using the enrichment analysis tool provided at the DAVID portal of NIH (Huang et al., 2009). p-values below the threshold 0.05 obtained by the hypergeometric test were adjusted for multiple testing using the Benjamini & Hochberg (BH) procedure (Benjamini and Hochberg, 1995).

Functional Similarity

Functional similarity was examined based on Gene Ontology (GO) Biological Process (BP) terms among the pairs of MCDS nodes. This was then compared to the functional similarity of gene pairs from the entire network as described in (Hamed et al., 2012). The permutation test was repeated 100 times and Kolmogorov-Smirnov test was applied to get the p-value.

Hypergeometric Test

The statistical significance of the results was assessed using the hypergeometric test which is based on sampling without replacement. The p-value for the test is calculated from the following formula:

$$p\text{-value} = 1 - \sum_{i=0}^{x} \frac{\binom{k}{i}\binom{M-k}{N-i}}{\binom{M}{N}}$$

where M is the total number of genes in the network, N is the sample size which is equal e.g. to the size of the MCDS, k is the number of genes in M with a specific property and x is the number of genes in the MCDS having that property. The cutoff value was set to $p = 0.05$ to report a set obtained by MCDS as a significant result. To apply the test, we used the online tool (GeneProf) which is described in (Halbritter et al., 2011).

Data and software

We tested the presented approaches to identify key player and master regulatory genes in several GRNs for *E. coli*, *S. cerevisiae*, a human breast cancer network and for the pluripotency of mouse ESC. We will present the obtained results one by one in the next section.

The dataset of *E. coli* is a GRN of the *Escherichia coli* strain K-12 that was downloaded on 22-July-2014 from RegulonDB (Salgado et al., 2013). It contains curated data for 1807 genes, including 202 TFs.

The dataset of *S. cerevisiae* was taken from the Yeast Promoter Atlas (YPA) downloaded on 26-March-2014 (Chang et al., 2011). It contains 5026 genes including 122 TFs. In this database, the target genes for each TF is a set of genes whose promoter regions contain the associated TF binding site for the TF binding motif.

The dataset for mouse is a manually curated GRN of mouse (*Mus musculus*) ESCs. It consists of 274 mouse genes/proteins and 574 molecular interactions, stimulations and inhibitions (Som et al., 2010). The network consists of genes that are involved in either induction, maintenance or loss of the pluripotency state and is thus termed pluripotency network throughout the text.

The breast cancer network used here was generated in (Hamed et al., 2015b) using a Bayesian learning approach that was coupled to an integrative network-based approach based on whole-genome gene expression profiling, DNA methylome, and genomic mutations of breast cancer samples from TCGA. The GRN networks were constructed via three steps: first the co-expression network was generated based on the topological overlap matrix as a distance measure. Then, we connected the co-expression interactions to regulatory information retrieved from publicly available regulatory databases accompanied

with motif search for all known binding motifs of the TFs represented in the co-expression network against the promoter regions of all genes in the network. Finally, a causal probabilistic Bayesian network was inferred from the co-expression modules utilizing the directed edges obtained from the previous step as a start search point to infer directionality between nodes. Clustering yielded ten network modules of dysregulated genes (Hamed et al., 2015b). Each module turned out to have distinct functional categories, cellular pathways, as well as oncogene and tumor suppressor specificity. We also extracted breast cancer specific subnetworks from the human genome regulatory interactome induced by the dysregulated mRNAs.
We implemented the ILP formulas for the directed forms of MDS and MCDS in the SageMath software system (The Sage Developers, 2015) version 6.8 using the glpk solver (Makhorin, 2008). We implemented the heuristic algorithm in Java and made it available as a plugin for the popular biological network analysis platform Cytoscape (Shannon et al., 2003). It is available at `http://apps.cytoscape.org/apps/mcds`.

3.3 Results

3.3.1 Global E. coli GRN

The GRN for *E. coli* studied here contains 1807 genes, including 202 TFs and 4061 regulatory interactions. This set of regulatory interactions in *E. coli* forms a general network which controls all sorts of responses which are needed in different conditions. With network density 0.001, the network can be considered as sparse. Due to this sparsity, MDS deems 199 TFs to be necessary to control the network. The network does not have any SCC with size larger than 5 nodes. For computing an MCDS, we therefore used the LCC underlying directed graph that contains 1198 genes. Based on the directed form of the LCC, target genes are placed at the bottom level and a set of TFs comprises the MCDS. In the LCC, the algorithm identified an heuristic MCDS containing 34 genes (11 dominators and 23 connectors) that cover the entire component, see Table S1.
Figure 3.3 illustrates the hierarchical structure between the 34 TFs contained in the MCDS. The hierarchical structure was drawn based on generalized hierarchies using breadth-first search as described in (Yu and Gerstein, 2006). A previous study that was based on an earlier version of RegulonDB identified 10 global regulators that regulate operons in at least three modules (Ma et al., 2004). Two of them, H-NS and CspA, do not belong to the LCC considered here. Two other global TFs identified previously (RpoS and RpoN) are no longer contained in the list of regulators in the version of RegulonDB used here. Out of the six remaining genes, the five genes IHF, CRP, FNR, ArcA and NarL are among the nine top genes in Table S1 and the sixth gene OmpR is found a bit further below in the list. Table S2 lists enriched KEGG and GO terms for the 34 genes in the MCDS of the *E. coli* gene regulatory network. As expected, the strongest enrichment is found for processes related to transcriptional regulation. The second most enriched term is related to two-component systems which enables *E. coli* to respond to changes arising from different environmental conditions (Stock et al., 2000).

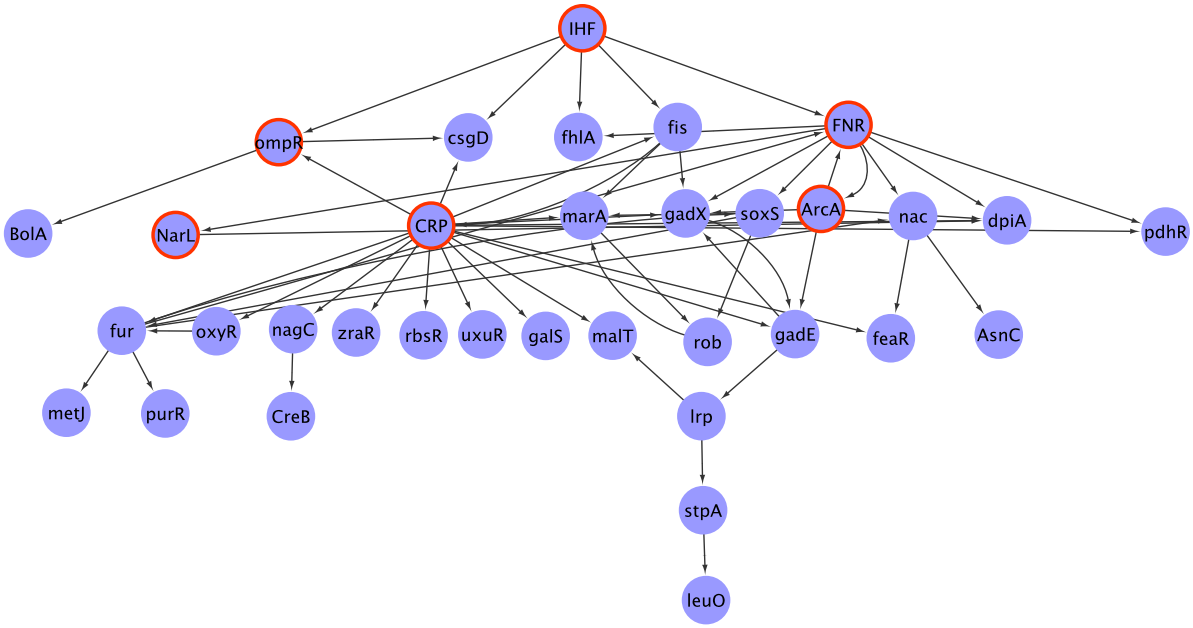

Figure 3.3: Connectivity among the genes in the connected dominating set of the LCC of the *E.coli* GRN. In this component, TFs construct the set of dominators and connectors. The red nodes are master regulatory genes identified as global regulators in [32].

3.3.2 Cell-cycle specific S. cerevisiae GRN

Next, we retrieved regulatory interactions in *S. cerevisiae* involving 122 TFs from the YPA (Chang et al., 2011). From this set of regulatory interactions, we extracted a cell-cycle specific subnetwork of 302 genes that are DE along the cell cycle of yeast as described in (Spellman et al., 1998). The 302 genes already form the LCC of this subnetwork. This set of genes is controlled by a MDS including 12 TFs and a heuristic MCDS including 14 TFs and 3 target genes. The MDS and MCDS elements are listed in Tables S3 and S4, respectively. Most of the TFs identified to belong to the MDS and MCDS have been identified before by experimental methods to be associated with the cell cycle (Lee et al., 2002). Figure 3.4 shows the GRN of the cell cycle activity of *S. cerevisiae* controlled by these 14 TFs. Table S5 lists enriched KEGG and GO terms for the 17 genes in this MCDS. As expected and similar to what we found for the *E. coli* network, the strongest enrichment was found for processes related to transcriptional regulation. 9 of the 17 genes (PMA2, YOX1, ACE2, SWI5, SWI4, ORC1, STB1, FKH1, TID3) are annotated to cell-cycle related GO terms, namely GO:0051329 $\sim$ interphase of mitotic cell cycle and GO:0000278 $\sim$ mitotic cell cycle and to the KEGG pathway sce04111:Cell cycle.

3.3.3 Pluripotency Network in Mouse ESCs

Next, we applied the MDS and MCDS methods to a manually curated GRN of mouse ESCs that consists of 274 mouse genes/proteins and 574 molecular interactions, stimulations and inhibitions (Som et al., 2010). We found that the heuristic MCDS of the LSCC (80 genes) of this network contains 29 TFs. The connectivity among these 29 TFs is displayed in Figure 3.5. The MCDS elements are listed in Table S6, respectively. Among

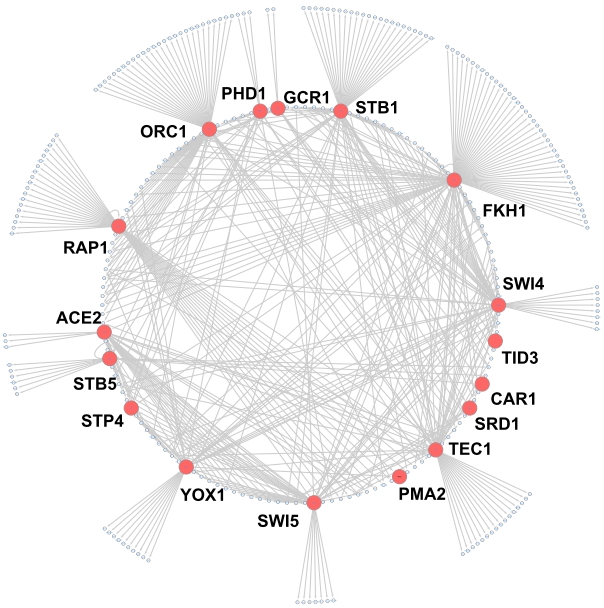

Figure 3.4: Tightly interwoven network of 17 TFs and target genes that organize the cell cycle of *S. cerevisiae*. Shown on the circumference of the outer circle are 164 target genes that are DE during the cell cycle. The inner circle consists of the 14 TFs from the heuristic MCDS and of 123 other target genes that are regulated by at least two of these TFs.

the set of regulators, 7 TFs including Pou5f1, Nanog, Sox2, Stat3, Esrrb, Tcf3, Sall4 are in common with an experimentally validated regulatory network controlling pluripotency that consists of 15 experimentally validated TFs (Xu et al., 2014). Such a result is unlikely to be obtained by chance (hypergeometric test p-value = 0.004) in a network with 176 TFs.

Next, we evaluated the ability of the MCDS method to detect a cooperative biologically functional backbone within the entire network. For this, we examined the functional similarity according to the Wang measure in the GoSemSim R package (Yu et al., 2010) (explained in chapter 4), among the pairs of MCDS nodes and compared this to the functional similarity of gene pairs from the mouse network, see Figure S1. This figure shows the cumulative distribution of the functional similarity scores between pairs of MCDS nodes of the mouse pluripotency network (in red) compared to the similarity scores of all possible pairs between genes of this network (in black). The Kolmogorov-Smirnov test revealed that the MCDS genes were functionally significantly more homogeneous than the randomly selected gene pairs of the whole network with p-value of 6.41e-05. This hints at the ability of the MCDS method to extract a functionally homogeneous network backbone that is expected to have an important role in maintaining the pluripotency state in early developmental stages. Table S7 lists enriched KEGG and GO terms for the 29 genes in the MCDS of the mouse ESC pluripotency network. In this case, GO terms related to developmental processes are stronger enriched than GO terms related to transcriptional regulation. The set of genes (Nanog, Cdx2, Esrrb, Pou5f1, Sox2 and TCl1) annotated with GO:0019827 are responsible for stem cell maintenance. The genes annotated with

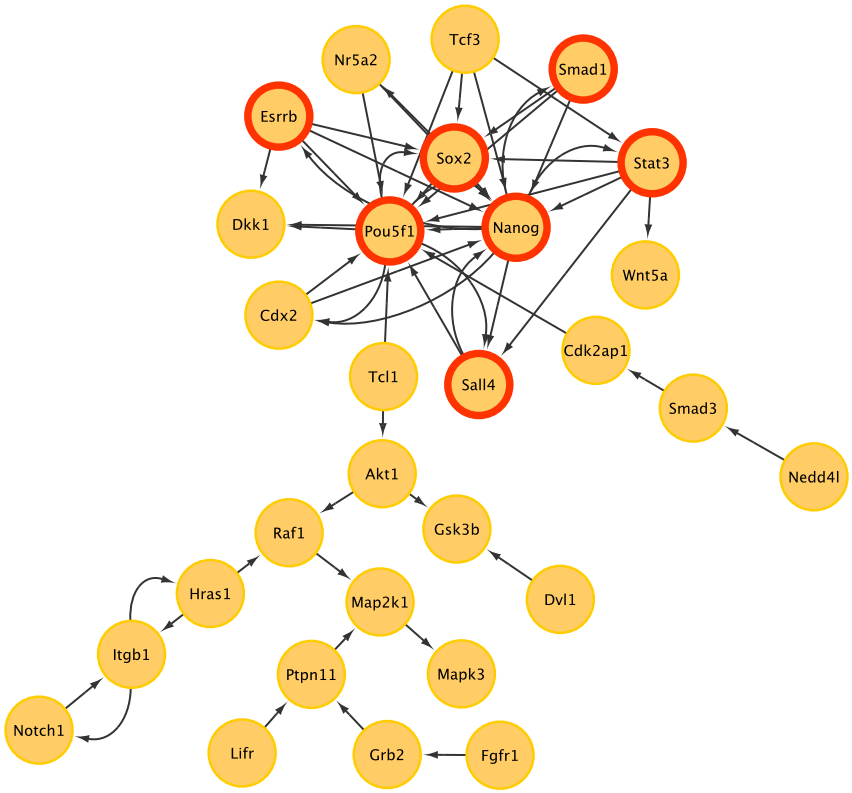

Figure 3.5: Connectivity among TFs in the heuristic MCDS in the LSCC of a GRN for mouse ESCs. The red circle borders mark the 7 TFs belonging to the set of master regulatory genes identified experimentally in (Xu et al., 2014).

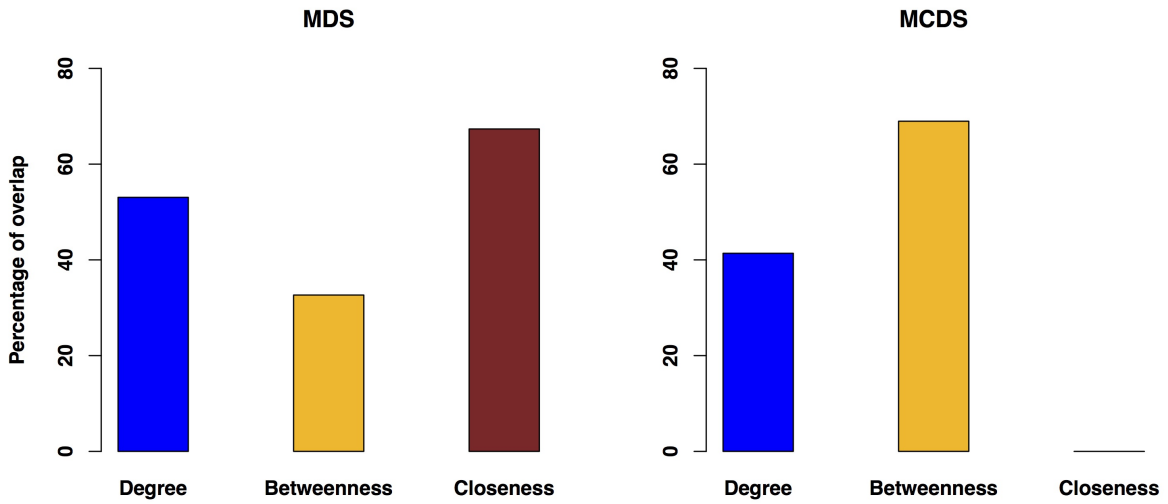

Figure 3.6: Percentage overlap of the genes of the MDS and MCDS with the list of top genes (same size as MCDS) according to 3 centrality measures. Shown is the percentage of genes in the MDS or MCDS that also belong to the list of top genes with respect to degree, betweenness and closeness centrality.

other GO terms are mainly related to embryonic development and other tissue-specific development.
To check the centrality significance of the MCDS genes in the LSCC, we selected the same number of genes as the size of MCDS with respect to degree, betweenness and closeness centrality. The centralities were measured using the igraph package (Csardi and Nepusz, 2006). We considered only outdegree nodes in the directed network. The results show that most of the genes contained in the heuristic MCDS are among the top nodes according to at least one centrality (degree, betweenness, closeness), see Figure 3.6. Among them, the top nodes of the MCDS have the highest overlap with the top nodes of the degree centrality and the betweenness centrality. Six out of 10 connector nodes in MCDS belong to the top 29 nodes with highest betweenness centrality according to Jaccard's index.

3.3.4 Human Disease Network

Finally, we applied the MCDS method to the LCC of ten breast cancer network modules where each LCC covers the whole module (Hamed et al., 2015b), see methods section. Table 3.1 lists the identified MDS and MCDS sets for the nine out of ten modules. One module (grey) could not be solved in appropriate time using ILP. In total, the MDS and MCDS sets of the nine modules contain 68 and 70 genes, respectively. Then, we looked up the known anti-cancer drugs that target any of the 70 proteins coded for by these genes based on experimentally validated drug-target databases as described in (Hamed et al., 2015b). In the network with 1169 genes including 228 drug target genes, we found that 20 of the 70 drug target genes belong to the genes identified using the MCDS. This is statistically significant with a p-value of 0.03 obtained from the hypergeometric test. Sixteen out of the 68 proteins belonging to the MDS genes are binding targets of at least

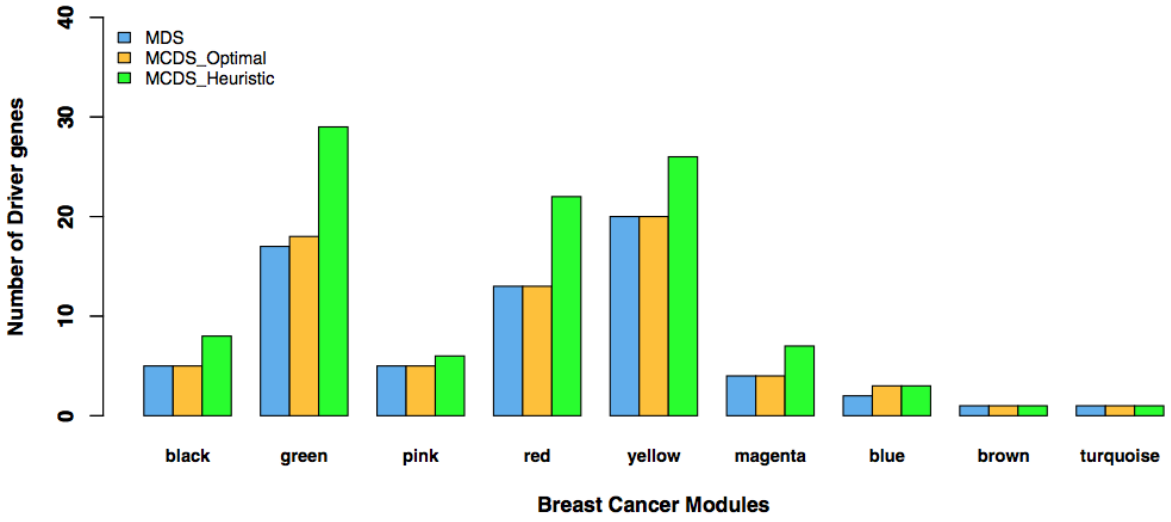

Figure 3.7: Number of MCDS genes determined by the heuristic approach or by the ILP formulation and in the MDS. Shown are the results for 9 modules of the breast cancer network.

one anti-breast cancer drug, see Table 3.1.

Next, we compared the set size of the optimal and heuristic solutions of MCDS and MDS for 9 out of the 10 modules. One module (grey) could not be solved in appropriate time using ILP. Table S8 displays the density and running time for the ILP solutions for the mentioned modules implemented in Sage. The running time was not correlated with the size or density of the networks. Figure 3.7 shows that the optimal solutions of MCDS and MDS contain almost the same number of genes for all modules. In comparison, the heuristic MCDS solutions (see, Table S9) contain 10-50 % more genes than the solutions of the other two approaches. We also compared the heuristic approach with the optimal solution in terms of overlapped identified genes. Table 3.2 indicates that according to Jaccard's index the solutions overlap approximately by about 60% in a range from 40% to 75%. Table S9 shows the results obtained by the heuristic approach of MCDS. Table S10 lists enriched GO BP terms and KEGG pathways in the MCDS genes obtained by heuristic approach. 12 genes (AKT1, RASSF5, WNT5B, ETS1, PDGFA, TP53, SPI1, NFKB1, TCEB1, MYC, TGFB1, DAPK1) belong to a known cancer pathway (p-value = 0.004). We hypothesize that the products of the some of the remaining identified MCDS protein coding genes may open up new avenues for novel therapeutic drugs.

3.3.5 Directed Random Networks

To characterize the size of problems which can be solved using the MCDS ILP formulation, multiple Erdos-Renyi random digraphs were generated using the Java code DigraphGenerator available in (Sedgewick and Wayne, 2015) with different sizes and densities. We discarded the networks whose running times exceeded 2 days. Table 3.3 shows that the size of MCDS reduces when the network density increases. A low density for networks of size more than 110 nodes leads to a dramatic increase in the computation time.

Table 3.1: Identified genes in the MDS and MCDS (ILP) for 10 modules of the breast cancer network.

Method	Module	Network Size	Result size	Key driver genes
MCDS	black	41	5	ZNF254, KIAA1632, ZNF681, SEC24B, ZNF615
MDS			5	ZNF254, KIAA1632, ZNF681, SEC24B, ZNF615
MCDS	blue	247	3	FAM54A,**ACAN**, GLDC
MDS			2	**ACAN**, FAM54A
MCDS	brown	195	1	**AATK**
MDS			1	**AATK**
MCDS	green	110	18	**ADPRHL2**, **AKT1**, LTBR, MAN2C1, SH3GLB2, UTP14A, WDR55, MADD, **B4GALT7**, **OS9**, MYO1C, **CDC34**, **CDC37**, RBM19, **MARS**, **CCDC22**, MAP2K2, DAP
MDS			17	**ADPRHL2**, LTBR, HMG20B, **HK1**, SH3GLB2, UTP14A, ELK1, MED6, **B4GALT7**, **OS9**, MYO1C, **CDC34**, CLN3, **INPPL1**, DAP, PLXNB1, TIMM44
MCDS	magenta	26	4	ILF2, **BGLAP**, POGK, **ATF6**
MDS			4	ILF2, **BGLAP**, **ATF6**, VPS72
MCDS	pink	30	5	TCEB1, RAB2A, ZNF706, TMEM70, **ATP6V1C1**
MDS			5	TCEB1, RAB2A, TMEM70, TCEA1, **ATP6V1C1**
MCDS	red	93	13	SIX4, **SP1**, **ATP1B1**, PCGF1, SUMF2, EPN3, GTF3A, RAP1B, FHL3, RPS3A, **ABCB8**, GFAP, **ANXA5**
MDS			13	LSM11, SIX4, PCGF1, SUMF2, EPN3, ZNRF2, GTF3A, RAP1B, FHL3, RPS3A, **ABCB8**, GFAP, NAGA
MCDS	turquoise	295	1	**ABHD10**
MDS			1	**ABHD10**
MCDS	yellow	132	20	**CASP10**, TSPAN2, **ACSL6**, HDAC11, SLC7A7, TRAF3IP3, GZMK, PAG1, LAP3 , HTRA4, **CD79B**, SPI1, GCET2, WAS, DFNA5, LRRC33, FCRL2, LCP2, TCTEX1D1, FUT4
MDS			20	**CASP10**, TSPAN2, **ACSL6**, HDAC11, TLR9, SLC7A7, FAM129C, TRAF3IP3, HTRA4,SPI1, CPXM2, GCET2, **FASN**, SLFN11, DFNA5, ETS1, PLS3, LCP2, TCTEX1D1, FUT4

The genes, whose protein products are known to be targeted by drugs, are marked in bold.

Table 3.2: Overlapping genes between the heuristic and optimal solutions of MCDS for modules of the breast cancer network. The names of the modules were introduced in the original ref. (Hamed et al., 2015b).

Module	shared genes	count
black	SEC24B, ZNF254, ZNF681	3
green	UTP14A, LTBR, SH3GLB2, OS9, CDC34, CDC37, AKT1	7
magenta	BGLAP, ATF6, ILF2	3
pink	ZNF706, TCEB1, TMEM70	3
red	FHL3, SUMF2, RPS3A, PCGF1, EPN3, GTF3A, ATP1B1	7
yellow	FUT4, SPI1, DFNA5, CASP10, PAG1, HDAC11, LCP2, TRAF3IP3, HTRA4, TSPAN2, GZMK	9
blue	ACAN, FAM54A	2

The modules brown and turquoise have only 1 mcds gene and give 100% overlap.

3.4 Summary and Discussion

Experimental identification of a set of key regulatory genes among large sets of genes is very time-consuming and costly. Therefore, computational methods such as the ones presented here are helpful to condense and shape a list of candidate genes to more promising candidates before planning and starting expensive experimental work. Such follow-up works could e.g. validate the regulatory roles of these genes by siRNA knockdown experiments, by over-expressing genes e.g. under the control of the highly inducible GAL1 promoter in yeast, or by CRISPR-type genome editing of promoter sequences containing TF binding sites. We presented three novel approaches (ILP formulation for the directed form of MDS, ILP formulation for the directed form of MCDS and heuristic algorithm for the directed form of MCDS) to identify driver genes and master regulatory genes responsible for a particular cellular identity. In the notion of network controllability, MDSs and MCDSs of biological networks are likely enriched in key regulatory genes. The results of these optimization problems can thus aids in pruning the network to the potentially more important nodes. We applied our method to the established GRNs of *E. coli* and *S. cerevisiae* and also to a pluripotency network of mouse ESC. The characteristics of these methods appear to be well suited, on the one hand, to the topology of approximately scale-free biological networks that contain a small number of high degree hub nodes and, on the other hand, to the observed tendency of these hubs to interact with each other. We showed that the networks can be controlled by a fairly small set of dominating TFs. A notable number of known master regulatory genes are detected in the connected dominating set of the components.
The number of driver genes obtained by the directed form of MDS and MCDS depends on the connectivity of the network. Networks with low connectivity yield a higher number of driver genes compared to networks with higher connectivity. The application of the MCDS method to modules of a regulatory network for a breast cancer network identified 70 key driver genes that could possibly drive the tumorigenesis process. Twenty of them are already known targets of available cancer drugs. The remaining dominating genes may be suitable candidates as news drug targets that may warrant further experimental validation.

Table 3.3: Runtime to determine an optimal solution for generated directed random networks with differing number of nodes and edges. Listed is also the resulting component density. All computations were conducted on a single threaded Intel XEON machine running at 2.2 Ghz. The networks whose running times exceeded 2 days were discarded.

nodes	edges	density	mcds_size	mcds_time (s)
10	9	0.1	6	1.94
10	27	0.3	4	2.00
10	45	0.5	2	1.81
10	63	0.7	2	1.80
10	81	0.9	1	1.85
30	87	0.1	9	2.53
30	261	0.3	4	2.21
30	435	0.5	3	2.03
30	609	0.7	2	2.07
30	783	0.9	1	2.17
50	245	0.1	11	4.43
50	735	0.3	5	3.83
50	1225	0.5	3	8.77
50	1715	0.7	2	4.47
50	2205	0.9	1	3.03
70	483	0.1	11	5.69
70	1449	0.3	5	25.56
70	2415	0.5	3	19.89
70	3381	0.7	3	61.69
70	4347	0.9	2	43.32
90	801	0.1	12	35.16
90	2403	0.3	6	1467.69
90	4005	0.5	4	1022.77
90	5607	0.7	3	137.33
90	7209	0.9	2	42.01
110	1199	0.1	13	497.21
110	3597	0.3	5	1761.15
110	5995	0.5	4	3132.90
110	8393	0.7	3	455.06
110	10791	0.9	2	27.90
130	1677	0.1	13	4706.06
130	5031	0.3	6	8625.99
130	8385	0.5	4	9903.08
130	11739	0.7	3	959.93
130	15093	0.9	2	279.81
150	2235	0.1	13	5902.89
150	6705	0.3	6	21610.52
150	11175	0.5	4	24067.34
150	15645	0.7	3	1994.68
150	20115	0.9	2	810.58
170	2873	0.1	-	-
170	8619	0.3	-	-
170	14365	0.5	4	44398.62
170	20111	0.7	3	2867.04
170	25857	0.9	2	675.49
190	3591	0.1	-	-
190	10773	0.3	-	-
190	17955	0.5	4	85180.81
190	25137	0.7	3	4738.96
190	32319	0.9	2	854.05

3.5 Remarks

(Golipour et al., 2012) showed that distinct sets of TFs are required for the transition of somatic cells to induced pluripotent stem cells (IPSCs) versus the maintenance. At this point, I proposed the hypothesis that the underlying networks define the role of dominators, whether they control the network in order to maintain the cellular identity or they are responsible for cellular transition. To check the hypothesis, I suggested Thorsten Will to apply the set cover approach on the hematopoiesis network to identify the set of transcriptomes that are responsible for cellular differentiation (Will and Helms, 2017). This suggestion was based on the one-to-one correspondence of MDS for directed graphs with the set cover problem by considering only the outgoing edges (Chlebík and Chlebíková, 2008) with respect to the constraints of protein complexes.

3.6 Availability of Data and Software

- **MDS:** This file includes the implementation of ILP formulation for MDS problem using glpk solver.
  available at https://github.com/maryamNazarieh/KeyRegulatoryGenes.
- **MCDS:** This file includes the implementation of ILP formulation for MCDS problem using glpk solver.
  Available at https://github.com/maryamNazarieh/KeyRegulatoryGenes.
- **User guide:** This guide (in the supplementary) contains instructions for users to use the MDS and MCDS programs to find the optimal solution in a directed network. It also includes two GRNs for modules of the breast cancer networ which can be used as input networks for ILP programs.
  Available at https://github.com/maryamNazarieh/KeyRegulatoryGenes.
- **Cytoscape plugin:** The Java implementation of the heuristic algorithm explained in this paper is available as a Cytoscape plugin at http://apps.cytoscape.org/apps/mcds, see Figure S2 with three examples, see Figures S3, S4, S5. Based on the request by Cytoscape users, see Figures S6, S7, we were informed by Cytoscape organizers that there is a need for an updated version of MCDS Cytoscape. This update will be available soon, which enables users to call Cytoscape from Python and R workflows.
- **Data:** This folder contains GRN files for *E. coli*, *S. cerevisiae*, pluripotency network of mouse ESC and breast cancer network modules.
  Available at https://github.com/maryamNazarieh/KeyRegulatoryGenes.

Chapter 4

TFmiR: A Web server for Constructing and Analyzing Disease-specific Transcription factor and miRNA co-regulatory Networks

This chapter is based on our paper entitled "TFmiR: A web server for constructing and analyzing disease-specific transcription factor and miRNA co-regulatory networks" by Mohamed Hamed, Christian Spaniol, Maryam Nazarieh and Volkhard Helms (Hamed et al., 2015a) published in the journal of Nucleic Acids Research.
Mohamed Hamed compiled the regulatory databases for human, developed the backend of the web server in R language and wrote the manuscript. Christian Spaniol developed the frontend of the web server in PHP and Java-script. Maryam Nazarieh modelled the problem of identifying hotspot nodes with a MDS problem and introduced the effect of centrality measures in regulatory networks as an alternative approach to the controllability approach via a minimum dominating set and implemented the heuristic approach of dominating set in Java for TF-miRNA co-regulatory networks and wrote the text of the paper in Latex. Prof. Volkhard Helms proposed the biological motivation of the paper, helped in designing the study and edited the manuscript.
TFmiR is a freely available web server for deep and integrative analysis of combinatorial regulatory interactions between TFs, microRNAs and target genes that are involved in disease pathogenesis. Since the inner workings of cells rely on the correct functioning of an enormously complex system of activating and repressing interactions that can be perturbed in many ways, TFmiR helps to better elucidate cellular mechanisms at the molecular level from a network perspective. The provided topological and functional analyses promote TFmiR as a reliable systems biology tool for researchers across the life science communities. TFmiR web server is accessible through the following URL: `http://service.bioinformatik.uni-saarland.de/tfmir`

4.1 Introduction

Among many genetic regulators, TFs and microRNAs (miRNAs) are the essential key players for regulating gene expression (Hobert, 2008). Together they play important roles in regulating virtually all cellular processes such as differentiation, proliferation, survival, and apoptosis (Esquela-Kerscher and Slack, 2006). Also genetic disorders and complex

diseases including cancer are mostly associated with perturbations of the interwoven regulatory circuit between TFs and miRNAs (Hanahan and Weinberg, 2011; Poos et al., 2013). TFs and miRNAs frequently form Feed Forward Loops (FFLs) and other network motifs to regulate cellular transcription in a connective manner (Poos et al., 2013; Yan et al., 2012). Therefore, utilizing the combined regulatory information on TFs and miRNAs as well as their target genes could shed light on key driver genes and miRNAs in human diseases and, in turn, suggests novel therapeutic strategies in disease treatment (Yan et al., 2012; Hamed et al., 2015b).
Several databases have been developed in order to facilitate research on transcriptional and post-transcriptional interaction types between TFs, miRNAs and target genes. For instance, TransFac (Matys et al., 2003), OregAnno (Griffith et al., 2008), and MsigDB (Liberzon et al., 2011) provide compilations of TFs regulating genes (TF $\rightarrow$ gene). TransmiR (Wang et al., 2010) provides information on which TFs regulate miRNAs (TF $\rightarrow$ miRNA). mirTarBase(Hsu et al., 2010), TarBase (Sethupathy et al., 2006) and miRecords (Xiao et al., 2009) collect target genes of miRNAs (miRNA $\rightarrow$ gene) in different organisms. Although still little is known about miRNA-mediated miRNA regulations, recent studies reported plausible evidences that miRNAs may regulate the expression of other miRNAs as well as their target genes (Yan et al., 2012; Matkovich et al., 2013). Thus, miRNA $\rightarrow$ miRNA interactions were computationally predicted and made available in the PmmR database (Sengupta and Bandyopadhyay, 2011).
Despite the general availability of such databases, generalized repositories integrating different kinds of molecular interactions and enabling to analyze their contributions to diseases are still missing. To this end, we present TFmiR, a web server that allows for integrative and comprehensive analysis of interactions between a set of deregulated TFs/-genes and a set of deregulated miRNAs within the relevant pathways of a certain disease. The tool unravels the disease-specific co-regulatory network between TFs and miRNAs and performs over representation analysis (ORA) for the involved TFs/genes and miRNAs. Our web server also detects feed forward loops (FFLs) consisting of miRNAs, TFs, and co-targeted genes (TF-miRNA co-regulatory motifs) and statistically assesses the functional homogeneity between the co-regulated targets. Furthermore, TFmiR utilizes seven different methods for identifying key network players that could possibly drive oncogenic processes of diseases and thus could act as potential drug targets. Especially when combined with experimental validation, these putative key players as well as the novel TF-miRNA co-regulatory motifs could promote novel insights to develop new therapeutic approaches for human diseases.

4.2 Materials and Methods

4.2.1 Description

TFmiR integrates genome-wide transcriptional and post-transcriptional regulatory interactions to elucidate human diseases. For a specified disease and based on user-supplied

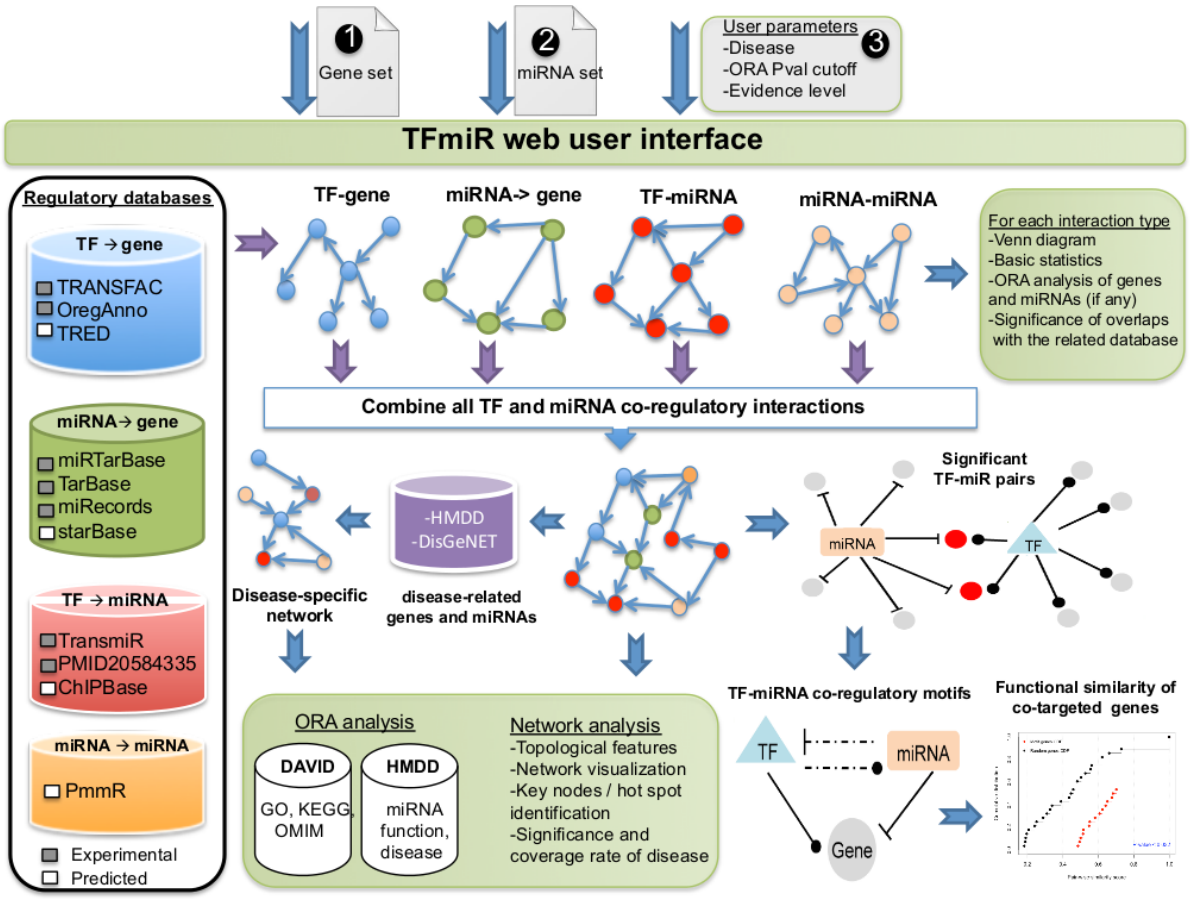

Figure 4.1: A system level overview of the TFmiR architecture describing the incorporated databases, data flows and output downstream analysis.

lists of deregulated genes/TFs and miRNAs, TFmiR investigates four different types of interactions, TF → gene, TF → miRNA, miRNA → miRNA, and miRNA → gene. It also unravels the circuitry between miRNAs, TFs and target genes with respect to specified diseases. For each interaction type, TFmiR utilizes information provided by established and curated regulatory databases of both predicted and experimentally validated interactions (see Figure 4.1) whereby all duplicate interactions were removed. For TF → miRNA interactions, we also integrated manually curated regulatory relationships compiled by the authors from (∼ 5000) published papers (Qiu et al., 2010). From the predicted miRNA → miRNA interactions in the PmmR database (Sengupta and Bandyopadhyay, 2011), we considered only the best hits having score < 0.2, which was computed as the normalized path length between the two involved miRNAs. The incorporated predicted miRNA → gene interactions were retrieved from starBase (Yang et al., 2011) by selecting only those predictions confirmed by three out of five prediction algorithms (targetScan (Bartel, 2009), picTar (Krek et al., 2005), RNA22 (Miranda et al., 2006), PITA (Kertesz et al., 2007), and miRanda (John et al., 2004)). Table S11 lists the included databases and the number of regulations available for each interaction type. In total, TFmiR currently integrates information on almost 10.000 genes, 1856 miRNAs, ∼ 3000 diseases including subtypes, and more than 111.000 interactions.

4.2.2 TFmiR user input Scenarios

TFmiR can be called in two ways. If the user submits two RNA sets (a set of deregulated mRNAs/genes and a set of deregulated miRNAs), the tool will return regulatory interactions based on the provided deregulated genes and deregulated miRNAs. In the second scenario, a user submits only a set of deregulated genes. In this case, TFmiR identifies the set of miRNAs whose target genes as well as regulator TFs are significantly enriched within the input deregulated genes using the hypergeometric distribution function followed by the BH adjustment with a cutoff value of 0.001. Sample input files of the deregulated genes and miRNAs are provided in the supplementary Figures S8 and S9. The user can optionally set the p-value cutoff (default is 0.05) for ORA on the resulting network nodes (genes/miRNAs). Finally, the user can control the evidence level (experimentally validated, predicted, or both) for the constructed regulatory interactions that will be used in the subsequent network analysis.

4.2.3 Functionality of TFmiR

TFmiR pools all four interactions types (TF $\rightarrow$ miRNA, miRNA $\rightarrow$ TF, miRNA $\rightarrow$ gene, miRNA $\rightarrow$ miRNA) based on the significant TF(gene)-miRNA pairs from the input deregulated genes and miRNAs and accordingly generates the entire combinatorial regulatory network. If a disease was selected, TFmiR uses data retrieved from the human miRNA disease database (HMDD) (Lu et al., 2008) as well as DisGeNET (a database for gene-disease association) (Bauer-Mehren et al., 2010) as sources for disease-associated miRNAs and genes, respectively. Interactions whose target nodes or regulator nodes are known to be associated with the disease compose the putative disease-specific network. TFmiR then offers three levels of downstream analysis: (1) the regulatory subnetwork of the four interaction types, (2) the combined network of all interaction types, and (3) the disease-specific network (if disease was selected).
For each interaction type subnetwork of *regulator* $\rightarrow$ *target* links, we display the total number of targets and regulators in the corresponding interaction databases, a Venn diagram depicting the overlap between the input deregulated targets (miRNAs/genes), and the targets of the input deregulated regulators (genes/miRNAs) available from the database. The significance of overlap is computed using the hypergeometric distribution test. To avoid the effect of false-positives in the *regulator* $\rightarrow$ *target* databases and to account for a different number of targets for the input deregulated regulators, a randomization test is conducted (n=1000). Furthermore, TFmiR carries out ORA for both gene analyses and miRNA sets comprising the interaction subnetwork.
For gene set analysis, TFmiR employs DAVID (Huang et al., 2009) to check for enrichment of GO terms (BP subcategory), KEGG pathways, and OMIM diseases as well as for clustering the genes based on their functional similarities. For miRNA set analysis, we used the miRNA-functional association data and miRNA-disease association data from HMDD to statistically relate the functional and disease terms to the miRNA set.
For the combined and disease-specific networks, TFmiR calculates for each network basic topological features, relevance to the disease-associated genes/miRNAs by testing the overlap significance with the network nodes, degree distribution plot, ORA analyses for both gene and miRNA nodes, network key nodes, and detects 3-node motifs. To measure the strength of correlation between the potential disease-specific network, the input disease, and the input deregulated genes and miRNAs, we compute a coverage ratio C_R

between the nodes of the disease-specific network and the nodes of the entire combined network.

$$C_R = \frac{N_d}{N_t}$$

Here, N_d represents the number of disease-specific network nodes, and N_t represents the total number of nodes in the entire network. We also calculate the C_R ratio between the edges of the two networks. All resulting networks are visualized using the interactive Cytoscape-web viewer (Lopes et al., 2010).

4.2.4 Identification of Network Key Nodes

We defined the key nodes as the top 10% highest centrality nodes of the TFs, miRNAs, and genes in the disease-specific and whole network. TFmir uses degree centrality, closeness centrality, betweenness centrality and eigenvector centrality as well as the common and union sets of the key nodes identified by these four measures. We also determine the minimal set of dominating nodes that regulate the entire network as explained in chapter 3. To solve such an optimization problem, we apply the directed version of the algorithm presented by (Rai et al., 2009) to search for the dominating set on the directed graphs.

4.2.5 Identification of TF-miRNA co-regulatory Motifs

Feed Forward Loops (FFLs) are interconnection patterns that recur in many different parts of a network and form key functional modules (Yan et al., 2012; Shen-Orr et al., 2002). They have been demonstrated as one of the most important motif patterns in transcriptional regulation networks (Shen-Orr et al., 2002) that govern many aspects of normal cell functions and diseases (He et al., 2007; Li et al., 2009). Here, TFmiR identifies four types of 3-nodes motifs (3 FFLs and 1 co-regulation motif) consisting of a TF, a miRNA, and their co-targeted gene and defines them as TF-miRNA co-regulatory motifs, see Figure 4.2. (1) The Coregulation-FFL includes only TF regulation of a target gene as well as miRNA repression of that target gene. (2) The TF-FFL includes TF regulation of the expression of both a miRNA and a target gene and it also includes miRNA repression of that target gene. (3) The miRNA-FFL includes miRNA repression of both a TF and a target gene, as well as TF regulation of this target gene. (4) The Composite-FFL describes TF regulation of both a miRNA and a target gene as well as miRNA suppression of that TF and that target gene.

1-Identifying significant TF-miRNA co-occuring pairs

We identified statistically significant TF and miRNA pairs that cooperatively regulate the same target gene using the hypergeometric distribution and evaluated p-values:

$$p\text{-value} = 1 - \sum_{i=0}^{x} \frac{\binom{k}{i}\binom{M-k}{N-i}}{\binom{M}{N}}$$

where k is the number of target genes of a certain miRNA, N is the number of genes regulated by a certain TF, x is the number of common target genes between these TF and miRNA, and M is the number of genes in the union of all human genes targeted by human miRNAs and all human genes regulated by all human TFs in our databases. Then,

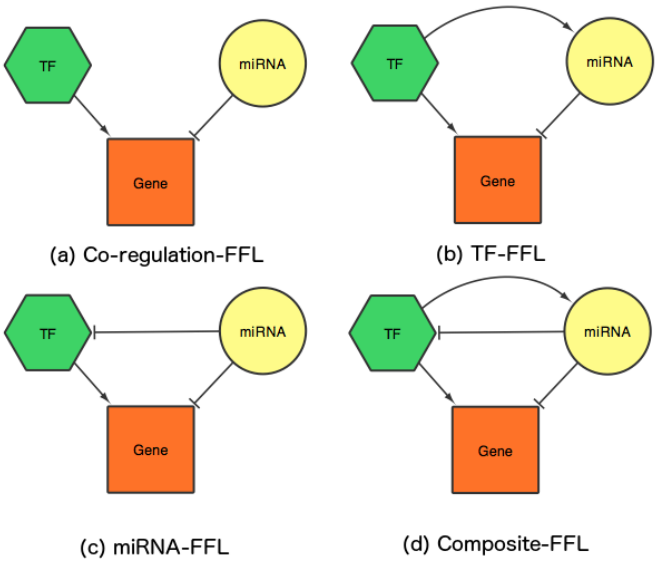

Figure 4.2: Schematic illustration of the four motif types detected in TFmiR. All motifs contain a TF, a miRNA, and a common target gene.

multiple test correction was performed by determining FDR according to BH (Benjamini and Hochberg, 1995) method and only those pairs with a adjusted p-value less than 0.05 were selected as significant TF-miRNA pairs.

2-Construction of candidate TF-miRNA-gene FFLs

All interactions associated with the significant TF-miRNA pairs are represented as connectivity matrix, M, such that $M_{ij} = 1$ if regulator i regulates target j where $i \in$ (TF, miRNA) , and $j \in$ (TF, miRNA, gene). Then, we scan all the $3 * 3$ submatrices of M that represent each type of the four considered FFL topologies, see Figure 4.2.

3-Significance of the FFL motifs

To evaluate the significance of each FFL motif type, we compare how often they appear in the real network to the number of times they appear in randomized ensembles preserving the same node degrees. The random networks were constructed 100 times and compared to the real network. The p-value is calculated as

$$p\text{-value} = \frac{N_h}{N_r}$$

where N_h is the number of random times that a certain motif type is acquired more than or equal to its number in the real network, and N_r is 100. We also calculate the z-score for each motif type to examine by how many standard deviations the observed real motif was above or below the mean of the random ones.

$$z\text{-score} = \frac{N_o - N_m}{\sigma}$$

Here N_o is the number of motifs observed in the real network, whereas N_m, and σ are the mean and standard deviation of the motif occurrence in 100 random networks, respectively.

4.2.6 Functional Homogeneity

In order to evaluate the biological evidence of the identified TF-miRNA co-regulatory motifs and better understand their functional roles, TFmiR allows the user to analyze the GO semantic similarity for all pairs of genes targeted by the same TF and miRNA pair or for all pairs of genes regulated by the TF or the miRNAs of that TF-miRNA pair, see Figure S10. The GoSemSim R package (Yu et al., 2010) is used to compute the semantic similarity scores according to the GO annotations. Statistical significance is determined by randomly selecting the same number of genes (co-targeted genes or co-regulated genes) from all Entrez genes with GO annotations, and computing their similarity scores. The permutation procedure is repeated 1000 times. Then, we carry out a Kolmogorov-Smirnov test to check whether the functional similarity scores of all gene pairs from the FFL motif are significantly higher than that of randomly selected pairs.

4.3 Results

4.3.1 Case Study

TFmiR was applied to several data sets related to complex diseases such as cancer, Alzheimer and diabetes. In a recent study on breast cancer (Hamed et al., 2015b), the authors identified 1262 deregulated genes and 121 deregulated miRNAs using gene and miRNA expression data from the TCGA portal (https://tcga-data.nci.nih.gov/tcga/). These two sets of deregulated genes and miRNAs are the default sample input files provided by the TFmiR web server. Next, TFmiR was used to reveal the co-regulation network between the deregulated genes/TFs and deregulated miRNAs and to better understand the pathogenic mechanisms associated with breast tumorigenesis. As user input parameters we set the p-value cut off to 0.05, disease was set to breast neoplasms, and the evidence level was set to both experimentally validated and predicted interactions.
For this data set TFmiR constructed a total of 427 regulatory interactions comprising 263 nodes of deregulated miRNAs and deregulated TFs/genes. The breast cancer-specific network involved 345 interactions and 212 nodes of deregulated miRNAs and genes with node and edge coverage ratios C_R of 80.6% and 80.8%, respectively. The provided ORA analysis of the disease network nodes revealed their implications in many cancer types as well as cancer-related KEGG pathways. Moreover, ORA analysis of the network miRNAs showed their involvement in cancerogenesis of multiple organs such as lung neoplasms, ovarian cancer, and adenocarcinoma, see Table S12. Additionally, TFmiR identified 22 key network players (10 genes and 12 miRNAs) based on the union set of four centrality measures described above, see Table S13. Interestingly, some of the identified key genes such as BRCA2, ESR1, AKT1, and TP53 were previously implicated and significantly mutated in breast cancer samples (Koboldt et al., 2012). More importantly, the protein products of the genes ESR1, TP53, TGFB1, AKT1, and BRCA2 are binding targets for anti-breast cancer drugs (Hamed et al., 2015b), see Table S14.
The effect of MDS on the breast cancer co-regulatory network has been explored in

(Hamed et al., 2015b). Next, we examined the TF-miRNA co-regulatory motifs that were significantly enriched in the entire interaction network. We identified 53 FFL motifs (3 composite-FFLs, 2 TF-FFLs, 6 miRNA-FFLs, and 42 coreg-FFLs). An interesting motif involving the TF SPI1, the miRNA hsa-mir-155, and the target gene FLI1 reveals how FFL motifs may help to better understanding the pathogenicity of breast cancer, see Figure S11 from the tool. Recent studies reported that the oncogene SPI1 is involved in tumor progression and metastasis (Guo et al., 2005; Rimmelé et al., 2010). However, the co-regulation of the oncogene FLI1 (Sakurai et al., 1784) by both SPI1 and the oncomiR hsa-mir-155 was not reported before. As the co-regulated genes of SPI1 and hsa-mir-155 have significantly more similar cellular functions than randomly selected genes (see Figure S12), this FFL motif provides novel insights on SPI1-miRNA network's alteration in breast cancer and suggests a cooperative functional role between SPI1 and potential miRNA partners.

4.4 Summary and Discussion

We developed TFmiR as a comprehensive web server for integrative analysis of the molecular interactions between TFs/genes and miRNAs and their interwoven critical roles in the pathology of human diseases. TFmiR provides an extended downstream analysis, a variety of user parameters, use case scenarios, and incorporates information from various well-established regulatory databases. TFmiR is based on user-provided sets of deregulated genes and/or miRNAs regardless of the data producing technologies of either microarray experiments, NGS, or PCR. We showed that unlike the traditional separate analysis of gene expression profiles (Bertucci et al., 2004; Chang et al., 2003) or the aberration of miRNA expression in cancer tissues (Yang et al., 2009; Xi et al., 2006), this integrated molecular analysis of deregulated miRNAs and genes using TFmiR was able to uncover literature confirmed core regulators as well as important new aspects of the TF/gene-miRNA interactomes, their co-regulation mechanisms, and the underlying pathogenesis of human breast cancer. The novel hub nodes of TFs/miRNAs could be further experimentally investigated as new potential drug targets. TFmiR was also able to characterize important TF miRNA co-regulatory motifs whose co-regulated genes form cooperative functional modules in breast oncogenesis processes.
Compared to the web services of related databases and tools such as Transmir (Wang et al., 2010), ChIPBase (Yang et al., 2013), CircuitsDB (Friard et al., 2010), starBase (Yang et al., 2011), miR2Disease (Jiang et al., 2009), and cGRNB (Xu et al., 2013), our TFmiR web server has several distinctive features: (1) TFmiR performs integrative analysis of molecular interactions between a set of deregulated genes and a set of deregulated miRNAs within or without the pathogenic pathways of a certain disease. In contrast, the above mentioned web tools only search the regulatory interactions of a single gene or a single miRNA. (2) TFmiR performs a rich network analysis, TF-miRNA co-regulatory motif detection, network visualization, statistical significance of the extracted interactions, and ORA analysis for each interaction type, the combined interaction network, and the disease network. Such an integrated analysis is not provided by other web tools. (3) TFmiR allows the user to retrieve either experimentally validated or predicted interactions or both. Such an option is not available using the other tools. In a somehow similar fashion, DisTMGneT (Sengupta and Bandyopadhyay, 2013) was developed for obtaining cancer-specific network based on user-selected sets of deregulated genes and miRNAs. However,

it lacks the downstream analysis, the varieties of user input parameters, and it is limited to a predefined set of miRNAs and genes as well as cancer disease. Also miRTrail (Laczny et al., 2012) performs ORA and Gene Set Enrichment (GSEA) analyses of interactions of genes and miRNAs based on expression profiles. However, it explores only miRNA → gene interactions.

4.5 Outlook and Perspective

TFmiR is planned to be integrated with other useful ORA tools such as KeyPathwayMiner (Alcaraz et al., 2011), GiGA (Thomas, 2010), HotNet (Vandin et al., 2011) and jActiveModules (Ideker et al., 2002) to allow the user to benefit their advances within TFmiR. We also intend to allow for submitting multi case expression data and times series data as well as the currently supported case/control data. Finally, expanding the TFmiR to elucidate the regulatory mechanisms of cellular processes (ex. stem cell differentiation) in addition to diseases would sort TFmiR of great interest for wide range of researchers and most of life science community.

Chapter 5

TopControl: Candidate Disease Gene Prioritization

This chapter is based on the manuscript entitled "Candidate Disease Gene Prioritization based on Topological Features" written by Maryam Nazarieh. Prof. Volkhard Helms edited the manuscript.
Potentially disease-associated genes are typically identified among those genes that are DE between disease and normal conditions. This strategy typically yields thousands of DE genes. Gene prioritizing schemes boost the power for identifying the most promising disease-causing genes among a set of candidates. We introduce a novel system for prioritizing genes among those which are significantly DE between tumor and normal samples. To achieve this goal, a TF-miRNA co-regulatory network is constructed for the set of candidates, where the ranks of the candidates are determined by topological and biological factors. We tested our prioritization system on breast invasive carcinoma and liver hepatocellular carcinoma datasets to reveal the power of the system to detect sets of disease-associated genes. Our experiments show that this novel prioritization technique identifies a significant set of known disease-associated genes, while suggesting new candidates which can be investigated later as potential disease-associated genes.

5.1 Introduction

RNA sequencing (RNA-Seq) generates an abundant number of sequencing reads, which leaves a large number of experiments for identifying disease-associated genes. Candidate gene prioritization helps experimentalists to focus their follow-up experiments on the most promising candidates based on the relationship between known disease genes and candidates.
Prioritizing tools typically produce their outputs either by filtering the candidates into smaller subsets or by ranking the candidates from the most promising to least promising ones. Some ranking techniques select the relevant candidates based on their similarities with user-defined disease-associated genes (Moreau and Tranchevent, 2012).
Currently, ranking methods based on network analysis indicate most of what we already know about the disease. They combine interaction networks with functional annota-

tions to select disease related candidates. The tool ToppNet (Chen et al., 2009) takes a different approach and ranks the candidate genes based on topological features in the protein-protein interaction network. The tool utilizes three algorithms mainly developed for social networks to prioritize candidate genes (Chen et al., 2009).
Other ranking approaches like text mining techniques select the candidates related to a disease through retrieving related documents from literature focusing on certain keywords (Moreau and Tranchevent, 2012). Since most prioritization methods require access to multiple databases, they are available mostly in the form of web services, e.g Endeavor (Tranchevent et al., 2016). Unlike Endeavor, NetworkPrioritizer (Kacprowski et al., 2013) utilizes the central nodes of a network mainly based on betweenness and closeness as seed nodes. It aggregates multiple node rankings derived according to the distance to central nodes to prioritize the candidates. However, it is unclear whether high centrality genes exercise a full control over the underlying network. FocusHeuristics aims to combine the static knowledge on the PPI network with the dynamics of gene expression for the gene ranking (Ernst et al., 2017). The software aggregates three features that are derived from gene expression data and the biological network, namely fold change, the differential link score and the interaction link score. Then the tool prunes the network to those nodes that exceed at least one of the predefined thresholds of the mentioned features. Therefore, the results vary based on different thresholds.
In this work, we present a prioritization system termed TopControl which finds a significant set of disease genes among the candidates by considering the set of genes that control the disease network. TopControl does not rely on any prior knowledge for prioritization of candidate disease genes. No seed genes need to be provided by a user. TopControl combines topological features and a biological factor to give an aggregated ranking to the candidates. MDS, MCDS and hub degree nodes are the topological features and $\log_2$(fold change) of expression is a biological factor that we considered in this work, see Figure 5.1.
In previous works, we demonstrated the power of MDS and MCDS in capturing the significant set of drug target genes in the breast cancer network. Moreover, we showed that a MCDS had high overlap with betweenness central nodes in the network, whereas the closeness centrality was more dominant in the MDS in the mouse ESC network (Nazarieh et al., 2016), see chapter 3. As explained earlier, MDS and MCDS genes in the regulatory networks control the network through their interactions, whereas hub-degree nodes are the top 10% high degree nodes in the network.
Here, we propose a novel candidate gene prioritization based on systematic analysis of DE genes between tumor and normal conditions. Initially, DE genes are filtered out to yield a set of candidates which form a network. The network candidates are prioritized based on network topological features. The priority of the candidates by which we mean relevance to a disease in disease networks increases when they are either in solutions derived from computing MDS, MCDS or they are in the hub set. Since significant DE genes with high fold change are the top candidates to experimentalists (Yang et al., 2016), the set of candidates with equal topological priorities are sorted based on the $\log_2$(fold change) of expression.
As explained earlier, MDS and MCDS are optimization problems. Solving the optimization problems by exact algorithms returns optimal solutions, but these optimal solutions are not unique. Different optimal solutions can be generated with different optimization algorithms. To address this problem, (Zhang et al., 2015b) modified the ILP formulation of MDS to return optimal solutions which are biased to hubs in the protein-protein

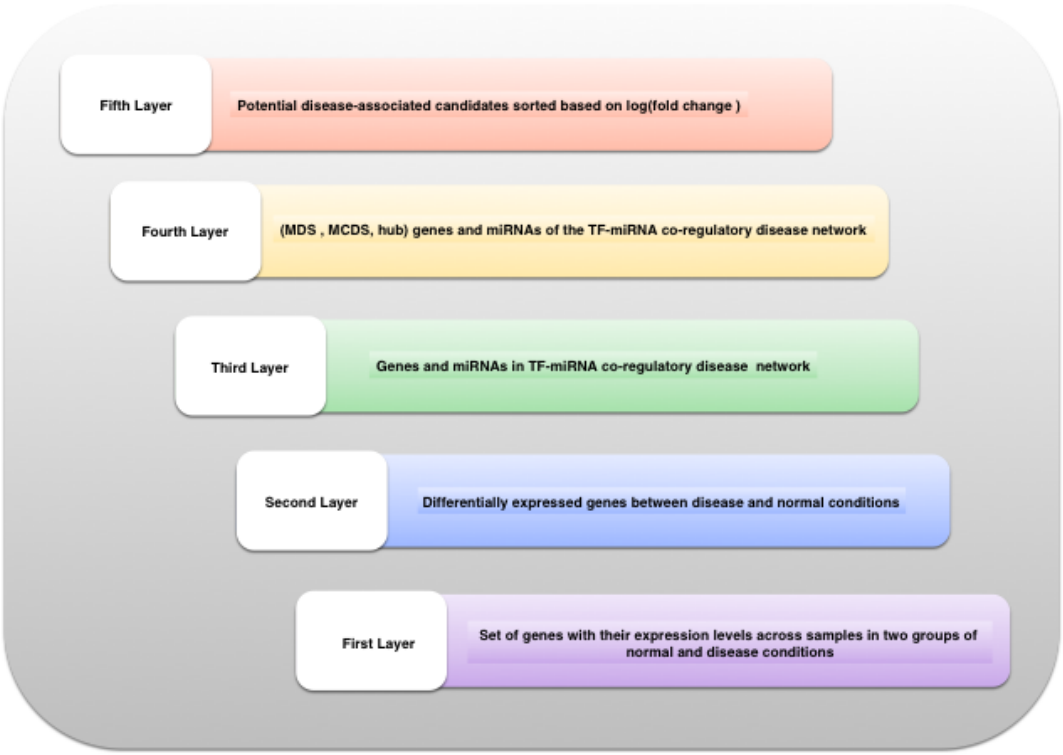

Figure 5.1: Schematic illustration of TopControl method with describing its hierarchical layers.

interaction networks. This modification cannot be applied to ILP formulations which address GRNs which are represented by directed graphs. (Liu et al., 2011) and (Nazarieh et al., 2016) realized that key drivers tend not to be among hubs in GRNs. Moreover, the heuristic approaches which are used when exact algorithms cannot find the solutions in appropriate time (Rai et al., 2009; Wightman et al., 2011), affect the optimal solution. The other problem is that not all nodes in MDS or MCDS are critical. We termed nodes that are part of both MDS and MCDS as critical nodes. TopControl addresses these issues by selecting the genes and miRNAs which are identified based on network controllability (the term was explained in chapter 3) by two different approaches. Although both MDS and MCDS consider a set of nodes which can control the state of GRNs through direct regulation of their target genes, only MCDS ensures that network controllability is achieved by a dominating pathway in the LCC of the network. To select the most promising candidates, TopControl gives priority to the candidates that are identified by both MDS and MCDS and have high degree of interactions. Therefore, the most promising candidates according to TopControl are the set of genes and miRNAs which are selected by all the three methods MDS, MCDS and hub set. They are sorted by $\log_2$(fold change) in descending order.

5.2 Materials and Methods

In this work, we use a hierarchical model of five layers to prioritize a set of candidates related to a certain disease. Genes at the first layer have the lowest priority and highest priority when they reach to the fifth layer. Basically, the genes in each layer is a subset of the genes in the lower layer. A set of miRNAs is introduced in the third layer if they interact with the selected genes from the second layer.
TopControl considers the whole set of genes in the first layer whose expression levels between two conditions have been provided. Then it selects the set of genes that are DE between disease and normal conditions based on the high potential that this set of genes carry to be associated with a related disease. With respect to the fact that disease genes interact with each other and related miRNAs, a TF-miRNA co-regulatory network is constructed for the set of DE genes. Therefore, the set of genes from the second layer that interact with each other and a selected set of miRNAs by TFmiR (the process of selection was explained in chapter 4) construct the disease-network in the third layer. Priority in the fourth layer is given to the set of genes and miRNAs which have either high interactions (hub-degree) or can control the network through their interactions. For this purpose, we consider two methods MDS and MCDS as explained in chapter 3. Top most candidates selected by TopControl in the fifth layer are the genes and miRNAs which have three roles in the network. They are in the set of hubs as well as MDS and MCDS. These genes and miRNAs are ranked up based on the absolute value of their fold changes in descending order.
We tested our method on the processed RNA-Seq data taken from the cancer genome atlas (TCGA) for matched tumor and normal samples of liver hepatocellular carcinoma (LIHC) and breast invasive carcinoma (BRCA) datasets downloaded on 15-Jun-2015. We exploited the DESeq method to identify the set of DE genes. TFmiR constructs a TF-miRNA co-regulatory network for the set of DE genes. Then the genes and miRNAs in the network were prioritized based on topological and biological factors.

5.2.1 Candidates in the First Layer

The first layer contains the whole set of genes with the corresponding expression levels across samples of tumor and normal conditions.

5.2.2 Candidates in the Second Layer

Potential disease-associated genes are typically identified among those genes that are DE between disease and normal conditions. This strategy usually yields thousands of DE genes, where the results are affected by various methods and sample size (Soneson and Delorenzi, 2013; Ching et al., 2014). Therefore, we considered DE genes as the candidates in the second layer. To identify the DE genes, we used the DESeq method (Anders and Huber, 2010).

5.2.3 Candidates in the Third Layer

DE genes and miRNAs which interact with each other to construct a network form the third layer of candidates. To construct the network, the TFmiR web server is used, see chapter 4. The set of miRNAs is selected such that target genes and regulator TFs of miRNAs are significantly enriched within the input deregulated genes using the hypergeometric distribution function followed by the BH adjustment (Benjamini and Hochberg, 1995) with a cut off value of 0.001. A complete network includes all the experimentally validated interactions between user-defined DE genes and retrieved miRNAs which are extracted from a variety of regulatory databases (Hamed et al., 2015a). The interaction types are TF $\rightarrow$ gene, TF $\rightarrow$ miRNA and miRNA $\rightarrow$ gene. In this work, we just used the complete network without specifying any disease.

5.2.4 Candidates in the Fourth Layer

Genes and miRNAs from the third layer are prioritized if they take part in one of the three sets, MDS, MCDS or hub set. Therefore, hubs, dominators and the nodes on the dominating pathway of the TF-miRNA co-regulatory network form the fourth layer. As a basis for this, we used regulatory networks involving TFs, microRNAs, and target genes that we predicted with our TFmiR web server from a set of DE genes. As hub-degree genes the web server outputs the top 10% highest degree nodes. A MDS was calculated based on the ILP formulation described in (Nazarieh et al., 2016), where MDS in a regulatory network is the minimum number of regulatory genes and miRNAs that control the whole network. A MCDS was computed based on the heuristic approach mentioned in (Nazarieh et al., 2016), where MCDS in a co-regulatory network is a connected set of genes and miRNAs that control the LCC of the network. A score is assigned to the nodes in this layer based on the number of roles they play in the network. The maximum score which is equal to three is given to a gene or miRNAs which is a hub, as well as a dominator and is on the dominating pathway of the network. Genes and miRNAs with the same score are sorted based on $\log_2$(fold changes) in descending order.

5.2.5 Candidates in the Fifth Layer

The set of candidates in the fourth layer which are selected by all three methods such as MDS, MCDS, and hubs are selected in the fifth layer and sorted in descending order

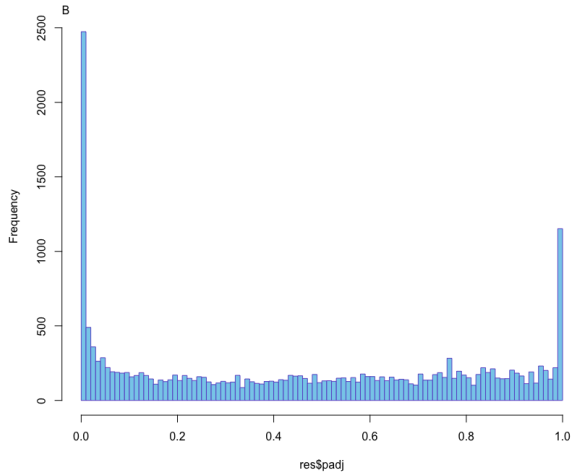

Figure 5.2: Histogram of p-values from the call to nbinomTest adjusted by BH derived by DESeq method.

based on the $\log_2$(fold changes) of expression between two conditions.

5.2.6 Biological Relevance

The biological relevance of the results obtained by the hub set, MDS, and MCDS criteria was evaluated using the enrichment analysis tool provided at the DAVID portal of NIH (version 6.8) based on the functional categories in GO Direct (Huang et al., 2009). p-values below the threshold 0.05 obtained by the hypergeometric test were adjusted for multiple testing using the BH procedure (Benjamini and Hochberg, 1995).

5.3 Results

5.3.1 LIHC dataset

The LIHC data set consisting of 20501 genes for 100 matching tumor and normal samples were given as input to the DESeq method. We selected the set of DE genes whose adjusted p-values were below 0.05. DESeq identified 3872 significant DE genes. Figure 5.2 shows the histogram of adjusted p-values visualized using the DESeq package (Anders and Huber, 2016). This set of DE genes was given to the TFmiR web server, by setting the p-value threshold to 0.05 and selecting experimentally validated resources. TFmiR constructed a complete TF-miRNA co-regulatory network with 275 genes and miRNAs and 383 regulatory interactions. 28 hub-degree nodes were obtained from the hotspot section of the web server. 61 and 68 genes and miRNAs were the results of MDS and MCDS, respectively. The union of hubs, MDS and MCDS were 82 distinct genes and miRNAs, see Table S15.
Table 5.1 shows the set of genes and miRNAs associated with hepatocellular carcinoma in the TFmiR complete network with the corresponding TopControl-assigned scores. 17 out of 33 genes and miRNAs associated with hepatocellular carcinoma (reported by DisGeNET (Bauer-Mehren et al., 2010) and HMDD database (Lu et al., 2008)) in the network were among the set selected by TopControl. This led to *sensitivity* = 52%, *specificity*

= 73% and *accuracy* = 71%. The significance of the union of genes and miRNAs selected by respective methods was assessed with the hypergeometric test, by returning a p-value of 0.004 with (0.002, 0.035, 0.013) per each set of hub, MDS, MCDS.
Enrichment analysis for these sets (hub, MDS, MCDS) demonstrates the related enriched GO terms and KEGG pathways, see Tables S16, S17 and S18. These three sets shared several GO terms like GO:0051726, GO:0008285 and GO:0042493 for regulation of cell cycle, negative regulation of cell proliferation and response to drug, respectively. Moreover, MDS and MCDS shared the GO:0010941 correspond to regulation of cell death.

Table 5.2 shows the top-most candidates in the fifth layer. 6 out of 18 proposed genes and miRNAs were reported in DisGeNET and HMDD. To biologically assess the potential of new candidates, we used LiverWiki (Chen et al., 2017), which is a comprehensive and up-to-date database for liver data. All the 12 new candidates were significantly expressed in hepatocellular carcinoma.

5.3.2 BRCA dataset

The breast invasive carcinoma (BRCA) data set with 20501 genes and 226 matching tumor and normal samples were given as input to the DESeq method. 5231 significant DE genes with adjusted p-values below 0.05 were selected by DESeq method. Figure 5.3

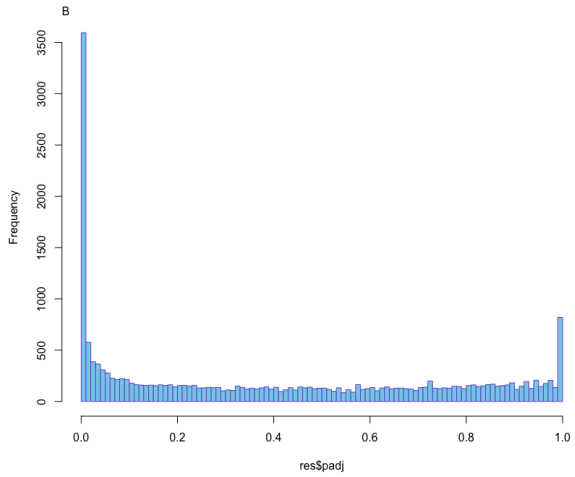

Figure 5.3: Histogram of p-values from the call to nbinomTest adjusted by BH derived by DESeq method.

shows the histogram of adjusted p-values visualized using the DESeq package (Anders and Huber, 2016).
The set of DE genes was given to the TFmiR web server, by setting the p-value threshold to 0.05 and selecting experimentally validated resources. TFmiR constructed a co-regulatory network with 463 nodes and 696 regulatory interactions. 47 hub-degree nodes were obtained as the top 10% high-degree nodes in the network. 97 and 113 genes and miRNAs were the results of MDS and MCDS, respectively. This led to a total 140 distinct genes and miRNAs. Table 5.3 shows the set of genes and miRNAs associated with breast neoplasms with the corresponding TopControl-assigned scores. 33 out of 50 genes

Table 5.1: TopControl-assigned scores of 33 disease-associated genes and miRNAs reported in DisGeNET and HMDD databases and were in the constructed TFmiR complete network for the LIHC dataset. The set of genes and miRNAs in the table were ranked up first by their TopControl-assigned scores and then sorted by LFC in descending order. D stands for the degree of the node and LFC for $\log_2$(fold change) of expression. The HCC flag denotes whether a gene or miRNA is associated with this disease (1) or not (0) according to DisGeNET and HMDD.

gene	D	hub	mds	mcds	score	LFC	HCC
E2F1	25	1	1	1	3	3.76	1
ESR1	9	1	1	1	3	2.19	1
JUN	33	1	1	1	3	1.39	1
MYC	18	1	1	1	3	1.07	1
hsa-let-7b	47	1	1	1	3	-	1
hsa-mir-29a	21	1	1	1	3	-	1
FOXM1	3	0	1	1	2	3.69	1
FOS	23	1	0	1	2	2.93	1
CEBPD	1	0	1	1	2	1.3	1
PDGFB	1	0	1	1	2	1.11	1
SREBF2	3	0	1	1	2	0.68	1
NFE2L2	4	0	1	1	2	0.66	1
TERT	6	1	0	0	1	9.17	1
RRM2	2	0	0	1	1	2.98	1
AR	3	0	1	0	1	1	1
HTATIP2	2	0	0	1	1	0.68	1
CCND1	12	1	0	0	1	0.67	1
CYP17A1	1	0	0	0	0	5.22	1
BIRC5	1	0	0	0	0	4.2	1
CCNE1	2	0	0	0	0	3.69	1
PTGS2	1	0	0	0	0	2.73	1
MT2A	2	0	0	0	0	2.71	1
HGF	1	0	0	0	0	2.37	1
GLUL	2	0	0	0	0	2.26	1
ACSL4	1	0	0	0	0	2.26	1
TFPI2	1	0	0	0	0	1.79	1
HSPB1	1	0	0	0	0	1.75	1
TYMS	1	0	0	0	0	1.73	1
MMP9	1	0	0	0	0	1.57	1
APOA1	1	0	0	0	0	1.38	1
ACE	2	0	0	0	0	1.09	1
F2	1	0	0	0	0	0.88	1
PTK2	1	0	0	0	0	0.71	1

Table 5.2: Top most candidates (genes and miRNAs) selected by TopControl in the hepatocellular carcinoma network. D stands for the degree of the node and LFC for $\log_2$(fold change) of expression. The HCC flag denotes whether a gene or miRNA is associated with this disease (1) or not (0) according to DisGeNET and HMDD.

gene	D	hub	mds	mcds	score	LFC	HCC
E2F1	25	1	1	1	3	3.76	1
EGR1	18	1	1	1	3	2.33	0
ESR1	9	1	1	1	3	2.19	1
JUN	33	1	1	1	3	1.39	1
NR1I2	9	1	1	1	3	1.37	0
MYC	18	1	1	1	3	1.07	1
JUND	6	1	1	1	3	0.99	0
STAT3	10	1	1	1	3	0.81	0
USF1	15	1	1	1	3	0.73	0
NR1H4	7	1	1	1	3	0.63	0
ETS1	9	1	1	1	3	0.61	0
IRF1	5	1	1	1	3	0.61	0
NR1I3	8	1	1	1	3	0.6	0
hsa-let-7b	47	1	1	1	3	NA	1
hsa-mir-29a	21	1	1	1	3	NA	1
hsa-mir-26a-5p	16	1	1	1	3	NA	0
hsa-mir-29a-3p	18	1	1	1	3	NA	0
hsa-mir-34a-5p	28	1	1	1	3	NA	0

and miRNAs associated with breast neoplasms (reported by DisGeNET (Bauer-Mehren et al., 2010) and HMDD database (Lu et al., 2008)) in the network were among the set selected by TopControl. This led to *sensitivity* = 66%, *specificity* = 74% and *accuracy* = 73%. The significance of the overlap was measured using the hypergeometric test with p-value of $3.15 * 10^{-8}$ with (6.68e-7, 2e-50, 6e-6) per each set (hub, MDS, MCDS) individually.

Enrichment analysis for these sets (hub, MDS, MCDS) describes the related biological process GO terms and KEGG pathways, see Tables S20, S21 and S22. These three sets shared some GO terms such as GO:0008285 and GO:0042493 for negative regulation of cell proliferation and response to drug, respectively. All of these sets shared several KEGG pathways related to different cancers. Moreover, MDS and MCDS shared several terms related to cell cycle and cell differentiation. Table S19 shows the candidates in the fourth layer proposed by TopControl. Table 5.4 shows the top most candidates with the highest scores given by TopControl, where some of them such as ESR2, FOS, E2F1, ESR1, JUN, STAT5A, ETS2, TFAP2A, hsa-mir-146a and hsa-mir-21 are disease-associated genes based on DisGeNET and HMDD databases. We found experimental validations in the literature mainly as drug target, metastasis promoter and tumor growth enhancer for other candidates such as EGR1, RUNX2, STAT1, TRAP2, IRF1 and USF1 in (Weiwei et al., 2013; Li et al., 2016; Hix et al., 2013; Reithmeier et al., 2017; Schwartz-Roberts et al., 2015; Bouafia et al., 2014).

Table 5.3: TopControl-assigned scores of 50 disease-associated genes and miRNAs reported in DisGeNET and HMDD databases and were in the constructed TFmiR complete network for the BRCA dataset. The set of genes and miRNAs in the table were ranked up first by their TopControl-assigned scores and then sorted by LFC in descending order. D stands for the degree of the node and LFC for $\log_2$(fold change) of expression. The BC flag denotes whether a gene or miRNA is associated with this disease (1) or not (0) according to DisGeNET and HMDD.

gene	D	hub	mds	mcds	score	LFC	BC
ESR2	7	1	1	1	3	2.58	1
FOS	20	1	1	1	3	2.47	1
E2F1	25	1	1	1	3	2.34	1
ESR1	19	1	1	1	3	1.79	1
JUN	45	1	1	1	3	1.6	1
STAT5A	7	1	1	1	3	1.6	1
ETS2	6	1	1	1	3	1.15	1
TFAP2A	24	1	1	1	3	0.9	1
hsa-mir-21	44	1	1	1	3	-	1
hsa-mir-146a	31	1	1	1	3	-	1
WT1	2	0	1	1	2	5.39	1
IFNB1	9	1	0	1	2	4.13	1
FOXM1	1	0	1	1	2	3.54	1
KIT	6	1	0	1	2	2.79	1
IL6	7	1	0	1	2	2.76	1
FOXA1	3	0	1	1	2	1.87	1
RARB	6	1	1	0	2	1.1	1
TRERF1	1	0	1	1	2	1.08	1
HEY2	1	0	1	1	2	1.06	1
MEIS1	1	0	1	1	2	1.01	1
NR2F6	2	0	1	1	2	1	1
KRAS	1	0	1	1	2	0.83	1
AR	5	0	1	1	2	0.63	1
ERBB2	6	1	0	0	1	1.89	1
BRCA2	4	0	0	1	1	1.81	1
AFP	1	0	1	0	1	1.43	1
EGFR	8	1	0	0	1	1.4	1
PDGFA	2	0	0	1	1	1.33	1
BRCA1	5	0	0	1	1	1.24	1
PARP1	1	0	1	0	1	1.2	1
CCND1	15	1	0	0	1	1.01	1
PGR	5	0	0	1	1	0.87	1
ZEB1	1	0	1	0	1	0.72	1
CAV1	3	0	0	0	0	3.19	1
RAD54L	1	0	0	0	0	3.06	1
CCL20	1	0	0	0	0	2.93	1
CCNE1	1	0	0	0	0	2.5	1
MMP3	2	0	0	0	0	2.39	1
F3	2	0	0	0	0	1.77	1
STMN1	2	0	0	0	0	1.68	1
TFPI2	1	0	0	0	0	1.58	1
SYNE1	1	0	0	0	0	1.44	1
PTGS2	1	0	0	0	0	1.26	1
DLL1	1	0	0	0	0	1.09	1
ALDOA	1	0	0	0	0	0.92	1
SERPINB5	3	0	0	0	0	0.75	1
SNAI2	2	0	0	0	0	0.74	1
SOD2	1	0	0	0	0	0.74	1
CSF1	2	0	0	0	0	0.73	1
PIM1	4	0	0	0	0	0.61	1

Table 5.4: Top most candidates (genes and miRNAs) selected by TopControl in the breast neoplasms network. D stands for the degree of a node and LFC for $\log_2$(fold change). The BC flag denotes whether a gene or miRNA is associated with this disease (1) or not (0) according to DisGeNET and HMDD.

gene	D	hub	mds	mcds	score	LFC	BC
EGR1	19	1	1	1	3	2.59	0
ESR2	7	1	1	1	3	2.58	1
FOS	20	1	1	1	3	2.47	1
E2F1	25	1	1	1	3	2.34	1
CEBPA	17	1	1	1	3	2.1	0
ESR1	19	1	1	1	3	1.79	1
JUN	45	1	1	1	3	1.6	1
STAT5A	7	1	1	1	3	1.6	1
RUNX2	5	1	1	1	3	1.4	0
STAT1	34	1	1	1	3	1.28	0
ETS2	6	1	1	1	3	1.15	1
MITF	8	1	1	1	3	0.94	0
TFAP2A	24	1	1	1	3	0.9	1
NR1H3	8	1	1	1	3	0.9	0
IRF1	16	1	1	1	3	0.65	0
ARHGEF7	11	1	1	1	3	0.62	0
USF1	16	1	1	1	3	0.61	0
SRF	5	1	1	1	3	0.58	0
TFDP1	10	1	1	1	3	0.58	0
hsa-mir-1	86	1	1	1	3	-	0
hsa-mir-21	44	1	1	1	3	-	1
hsa-mir-145-5p	39	1	1	1	3	-	0
hsa-mir-21-5p	32	1	1	1	3	-	0
hsa-mir-146a	31	1	1	1	3	-	1
hsa-mir-34a-5p	27	1	1	1	3	-	0

5.4 Comparison of TopControl with Endeavor

We compared the results of TopControl with Endeavor (Tranchevent et al., 2016). The set of DE genes for the LIHC and BRCA datasets derived using DESeq method were given as input to Endeavor. We selected the same number of genes as the size of TopControl from top selected genes by Endeavor based on p-value that stands for the significance of a combination of rankings. In this comparison, we ignored the set of miRNAs from the lists of top candidates by TopControl as Endeavor does not consider miRNAs. During the training of Endeavor, we provided the minimum required number of genes for training the model, as TopControl does not rely on any prior knowledge. We selected all data sources to build models and prioritize the candidates from Endeavor. DisGeNET was used for the evaluation of the results. For the case of hepatocellular carcinoma, both methods performed equally well with detecting 15 related disease-associated genes among of 77 top candidates, see Figure 5.4 panel (A). The overlap among the identified disease-associated genes comprises seven genes (E2F1, MYC, JUN, CCND1, ESR1, TERT, SREBF2). In the case of breast neoplasms, TopControl outperformed Endeavor with detecting 31 compared to 26 related disease-associated genes out of 134. This led to an overlap of 20 genes including (ESR2, PGR, AR, JUN, CCND1, RARB, NR2F6, EGFR, FOXA1, FOS, ERBB2, E2F1, WT1, BRCA1, KRAS, TFAP2A, ZEB1, STAT5A, TRERF1, PARP1), see Figure 5.4 panel (B).

Figure 5.4: A) Comparison of identified disease-associated genes between TopControl and Endeavor for the LIHC dataset. B) Comparison of identified disease-associated genes between TopControl and Endeavor for the BRCA dataset.

5.5 Summary and Discussion

This chapter introduced the tool TopControl as a new prioritizing method based on topological and biological factors to propose a new set of disease-associated candidate genes. This led to the detection of a significant set (based on hypergeometric test) of disease-associated genes and miRNAs, while introducing a new set of candidates. As a basis for this, we used regulatory networks involving TFs, microRNAs, and target genes that we predicted with our TFmiR web server from a set of DE genes and identified the hubs in a similar way to this tool. Then we applied the ILP formulation of MDS and the heuristic approach of MCDS to find MDS and MCDS in the underlying networks. Here, we processed RNA-Seq data taken from TCGA for matched tumor and normal samples of LIHC and BRCA. DE genes were identified by the DESeq tool. The BRCA dataset has more than twice as many samples size as LIHC. This affected the sensitivity without changing the specificity.
TopControl differs from other gene prioritization tools mainly in that it does not rely on any prior knowledge to train and build models for prediction of disease-associated genes. It suggests topological features as potential candidates and prioritizes them to target critical ones.

Chapter 6

Topology Consistency of Disease Networks

This chapter is based on the manuscript entitled "Topology Consistency of Disease-specific Networks".
Maryam Nazarieh discovered the topological consistency among different bioinformatics tools, performed the experiments and wrote the manuscript. Hema Sekhar Reddy Rajula preprocessed the LIHC dataset and applied the analysis tools for differential expression under the guidance of Maryam Nazarieh (I was advisor of Hema, only on this work under supervision of Prof. Helms). Prof. Volkhard Helms helped in designing the study and edited the manuscript. The motivation of this investigation was that various methods for identifying DE genes yield quite different results. Thus, we investigated whether this affects the identification of key regulatory players in regulatory networks derived by downstream analysis from lists of DE genes that may be responsible for disease processes.
While the overlap between the sets of significant DE genes was only 26% in liver hepatocellular carcinoma and 28% in breast invasive carcinoma, we found that the topology of the regulatory networks constructed using TFmiR for the different sets of DE genes was highly consistent with respect to hub-degree nodes, MDS and MCDS. This suggests that key genes identified in regulatory networks derived from systematic analysis of DE genes may be a more robust basis for understanding diseases processes than simply inspecting the lists of DE genes.

6.1 Introduction

RNA-Seq or whole transcriptome shotgun sequencing uses next-generation sequencing technology to quantify the abundance of RNA in a biological sample at a given moment in time. Despite a high correlation between gene expression profiles using the same set of samples, RNA-Seq is capable of detecting low abundance transcripts and allowed for the detection of more differentially expressed (DE) genes with higher fold-changes than microarray data (Zhao et al., 2014).
A typical differential expression analysis consists of normalizing raw counts, dispersion estimation and performing a statistical test. Statistical tests are performed to see if the observed differences in read counts data between two groups are statistically significant. The results returned by them typically in terms of p-values reject or accept a certain null hypothesis which signifies that the mean values of the two groups are equal or that

the read counts follow the same distribution. To obtain accurate significant results, an assumption about the distribution of the underlying data is required (see section 2.3), e.g, a t-test which is widely to process microarray data assumes that the data has a normal distribution. This assumption does not work for RNA-Seq data with discrete values. Several data distributions have been suggested to model the RNA-Seq values. Among them, Poisson distribution and Negative Bionomial (NB) distribution are used most often. The Poisson distribution does not account for over-dispersion in the data and presumes that mean and variance are equal which leads to lots of false discovery rates. Therefore, the NB distribution with considering both mean and dispersion parameters is typically preferred to model RNA-Seq data. Although, several methods such as DESeq (Anders and Huber, 2010) and edgeR (Robinson et al., 2010) assume that RNA-Seq data can be modelled by the NB distribution, each of them has a different way to estimate the model parameters, mean and dispersion. This lead to different results with different sizes. The problem gets worse when the methods have different assumptions about the data distribution. (Soneson and Delorenzi, 2013) conducted a comprehensive comparison between the results of eleven differential expression analysis methods which take RNA-Seq read counts as input on both simulated and real data. They demonstrated characteristics and benefits of utilizing each method. There appears to be no general consensus among the DE genes found by the different methods (Soneson and Delorenzi, 2013). Here, we selected four methods such as DESeq, edgeR, voom and VST from the above-mentioned methods which take read counts as input and return p-values. The methods have been explained in details in section 2.3.
In this work, we applied these methods to the LIHC and BRCA datasets to get the significant DE genes. Initially, we showed the overlap among their results. Then, we showed that key regulatory players are highly consistent among different methods despite generally low similarity among their sets of DE genes.

6.2 Materials and Methods

We tested our method on the processed RNA-Seq data obtained from TCGA for matched tumor and normal samples of LIHC and the patients (see chapter 5). We exploited the R packages of DESeq, edgeR, VST and voom methods (Anders et al., 2013; Law et al., 2014) to identify the respective sets of DE genes. With TFmiR a TF-miRNA co-regulatory network was constructed for each set of DE genes. Then we used MDS and MCDS tools to get the topological features in the regulatory networks.

6.2.1 Network Construction with TFmiR

The networks were constructed using TFmiR web server, see chapter 4. In this work, we focused on disease-specific network. In this study, we selected hepatocellular carcinoma and breast neoplasms from the list of diseases for construction of the disease-specific co-regulatory networks.

6.2.2 Topology Inference

We inferred network topologies of disease-specific networks involving TFs, microRNAs, and target genes that we predicted with our TFmiR web server from a set of DE genes.

We took the same strategy as TFmiR and selected the top 10% highest centrality nodes as hub-degree nodes. A MDS was calculated based on the ILP formulation described in (Nazarieh et al., 2016), where MDS in a regulatory network is the minimum number of regulatory genes and miRNAs that control the whole network. A MCDS was computed based on the heuristic approach mentioned in (Nazarieh et al., 2016), where MCDS in a co-regulatory network is a set of genes and miRNAs that are connected and control the LCC of the network.

6.3 Results

6.3.1 Inference of DE Genes

The processed matching tumor-normal samples of LIHC and BRCA consist of 100 and 226 samples with 20501 genes, respectively. The data were given as input to the R packages DESeq, edgeR, voom and VST. We obtained adjusted p-values as output from all four packages. Based on the adjusted p-value threshold 0.05, we determined if genes are DE. The number of significant DE genes for the LIHC dataset with DESeq, edgeR, voom and VST were 3872, 11399, 10610 and 10238, respectively and for the BRCA dataset 5231, 14722, 15559 and 13918, respectively. Venn diagrams in Figure 6.1 show the number of genes which are common between these methods. The Venn diagrams were visualized using the R package VennDiagram (Chen and Boutros, 2011).

6.3.2 Reconstructed Networks

In the case of LIHC dataset, analyzed by the DESeq method, 163 nodes and 199 edges make up the hepatocellular carcinoma disease-specific network. The hubs, MDS and MCDS of the network were visualized in Figure 6.2. In the case of breast neoplasms with BRCA dataset and the same method, 227 nodes and 302 edges were retrieved from TFmiR databases considering experimental resources. The TFmiR web server also was used to construct disease-specific networks for the set of DE genes derived from edgeR, voom and VST. Tables S23 and S27 show the number of nodes, edges, hubs, MDS and MCDS for the LIHC and BRCA data sets for all the above-mentioned methods, respectively.

6.3.3 Topology Consistency

We performed pairwise comparisons between the topological features of these networks, see Tables 6.1 and 6.2. The results demonstrate the percentage overlap of hubs, MDS and MCDS between the aforementioned analysis methods. As shown in the tables, DESeq has the highest overlap with edgeR than voom and VST in both the studies, whereas the topological features of edgeR are highly overlapped with voom than VST.
Tables S24, S25, S26 show the list of consistent genes and miRNAs that are common among all the methods for hepatocellular carcinoma and in Tables 6.3, 6.4, 6.5 for breast neoplasms. The tables show a high number of consistent genes and miRNAs among the topological features of the methods. 13 out of 17 hubs selected by DESeq were identified by other methods in the case of LIHC dataset and 20 out of 23 for the case of BRCA dataset. The common MDS and MCDS make up almost 70% to 75% of the selected MDS and MCDS by the DESeq method. The number of consistent topological features among edgeR, voom and VST could increase when we disregard DESeq method, as it has the

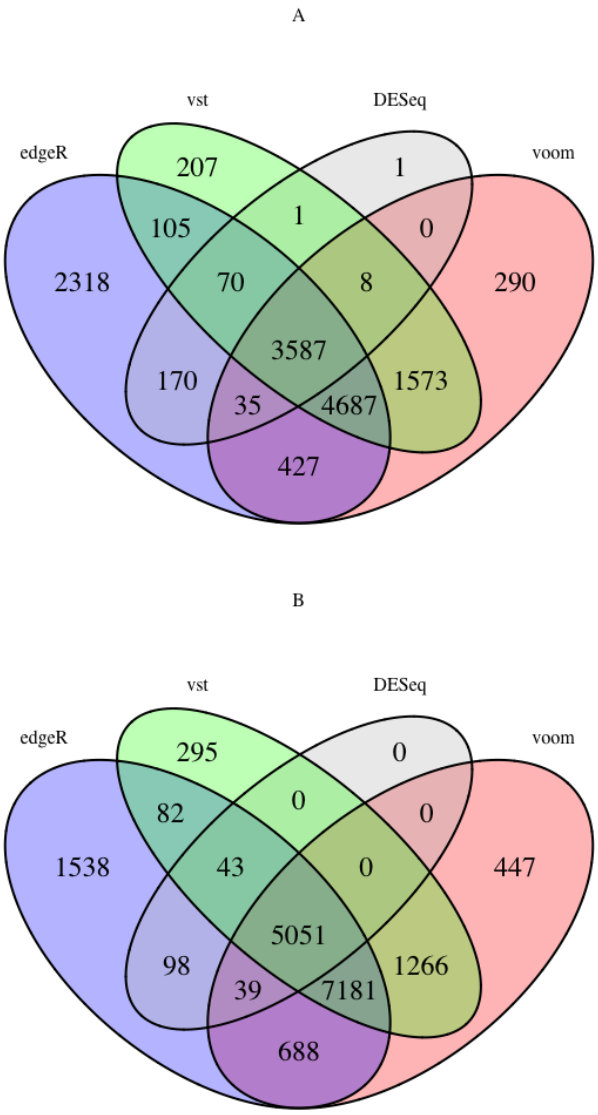

Figure 6.1: A) Venn diagram describing the number of overlapped DE genes between the results of DESeq with edgeR, voom and VST for the LIHC dataset. B) Venn diagram describing the number of overlapped DE genes between the results of DESeq with edgeR, voom and VST for the BRCA dataset.

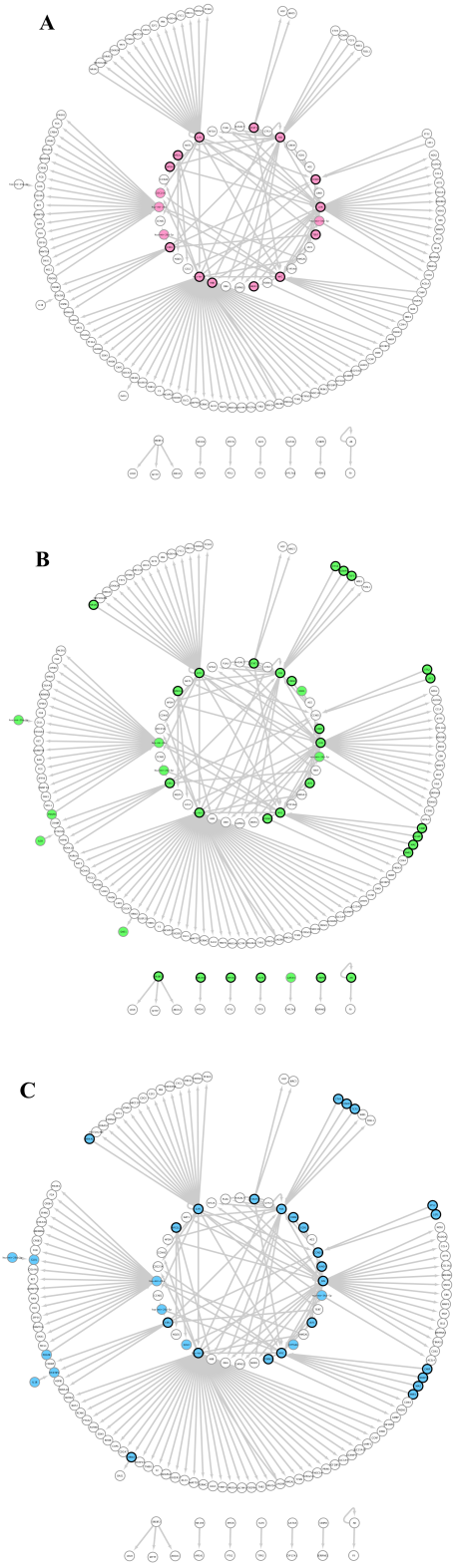

Figure 6.2: Topology consistency in the disease-specific networks for the LIHC dataset. A) hub-degree genes and miRNAs colored pink. B) MDS nodes colored green. C) MCDS nodes colored blue. The black circle borders mark the consistent genes and miRNAs between DESeq, edgeR, voom and VST.

Table 6.1: Pairwise comparison of hubs(left), MDS(middle) and MCDS(right numbers) for the networks constructed from the set of DE genes analyzed by DESeq, edgeR, voom and VST methods for the LIHC dataset.

Methods	edgeR	voom	VST
DESeq	82,84,77	88,81,74	82,81,71
edgeR	-	80,82,78	70,75,68
voom	-	-	87,92,95

Table 6.2: Pairwise comparison of hubs(left), MDS(middle) and MCDS(right numbers) for the networks constructed from the set of DE genes analyzed by DESeq, edgeR, voom and VST methods for the BRCA dataset.

Methods	edgeR	voom	VST
DESeq	96,83,81	91,80,79	96,83,80
edgeR	-	86,83,83	70,72,75
voom	-	-	83,85,88

lowest number of DE genes, the smallest network size and subsequently smallest set of hubs, MDS and MCDS among all the methods.

6.3.4 Robustness of the Results

To check the robustness and significance of the results, 100 random networks were constructed with 11000 and 14000 randomly selected genes as pseudo set of DE genes . Related networks were constructed with TFmiR. Detection of hubs, MDS and MCDS were performed as explained before. The results of DESeq were compared with the other tools, edgeR, voom and VST. We used the widely used tool, DESeq (Anders and Huber, 2010) as the base line of comparison because it appears to be a very conservative method to detect the set of DE genes (Soneson and Delorenzi, 2013; Anders et al., 2013). Moreover, we realized from the previous experiments that DESeq contains the highest number of consistent topological features among all the methods. Figure 6.3, panels (A) and (C) visualize the overlap percentage between DESeq and other methods, and panels (B) and (D) show the percentage overlap of hubs, MDS and MCDS of DESeq with random networks for hepatocellular carcinoma and breast neoplasms, respectively. If one provides more than half of all human genes as input and generates a regulatory disease-specific network, one can expect that a considerable fraction of the real key genes are recovered by chance. In the two studied cases, between 20 and almost 60% overlap with the DESeq key genes. However, the results indicate that random selection of nodes does not reach to

Table 6.3: Consistent hub genes and miRNAs for the BRCA dataset.

JUN, hsa-mir-21, E2F1, TFAP2A, FOS, ESR1, CCND1, IFNB1, EGFR, STAT5A, IL6, ESR2, KIT, ERBB2, RARB, MYC, ETS2, hsa-mir-21-5p, STAT1, BRCA1

Table 6.4: Consistent MDS genes and miRNAs for the BRCA dataset.

EGR1, JUN, RARA, RARB, BMP6, hsa-mir-21-5p, ESR2, TCF7L2, TNFSF12, FOXA1, MEIS1, TCF3, PARP1, ETV5, ESR1, TFDP1, NR2F6, TRERF1, FOXM1, THRA, ZEB1, USF1, SRF, EFNA2, GBX2, LEF1, HEY2, E2F1, LMO2, hsa-mir-34a-5p, STAT5B, SREBF1, hsa-mir-21, WT1, TFF3, IRF7, TAL1, TEAD4, CEBPD, TFAP2A, ETS2, KLF6, hsa-mir-145-5p, NR3C1, JUND, NR4A1, STAT5A, RPA3

Table 6.5: Consistent MCDS genes and miRNAs for the BRCA dataset.

FOXA1, THRA, BRCA1, BRCA2, FOXM1, NR3C1, ETS2, CCND1, HEY2, TEAD4, SREBF1, hsa-mir-21-5p, LMO2, FOS, ETV5, TFDP1, TAL1, KIT, IRF7, TFF3, CEBPD, hsa-mir-145-5p, SRF, LEF1, EGFR, GBX2, CYP11A1, hsa-mir-21, E2F1, STAT5A, TFAP2A, TCF7L2, STAT5B, EGR1, JUN, JUND, IFNB1, CAV1, TNFSF12, hsa-mir-34a-5p, TRERF1, ESR1, BMP6, USF1, ESR2, VEGFA, NR4A1, IL6, EFNA2, WT1, RPA3, TCF3

the same level of topological overlap compared to the topological overlap of DESeq with edgeR, voom and VST. Since none of the 100 random networks reached the values for the real networks, the significance is below $p = 0.01$.

6.4 Summary and Discussion

In this work, we showed that the sets of hubs, MDS and MCDS are well consistent in disease-specific networks constructed from different sets of DE genes identified by different analysis methods. For this purpose, we used regulatory networks involving TFs, microRNAs, and target genes that we predicted with our TFmiR web server from a set of DE genes and identified the hub-degree nodes. Although the overlap between the sets of significant DE genes was only 26% in liver cancer and 28% in breast cancer, we found that the topology of the regulatory networks constructed using TFmiR for the different sets of DE genes was highly consistent with respect to hub-degree nodes and MDS and MCDS (70-90%). This suggests that key genes identified in regulatory networks derived from DE genes may be a more robust basis for understanding diseases processes than simply inspecting the lists of DE genes.

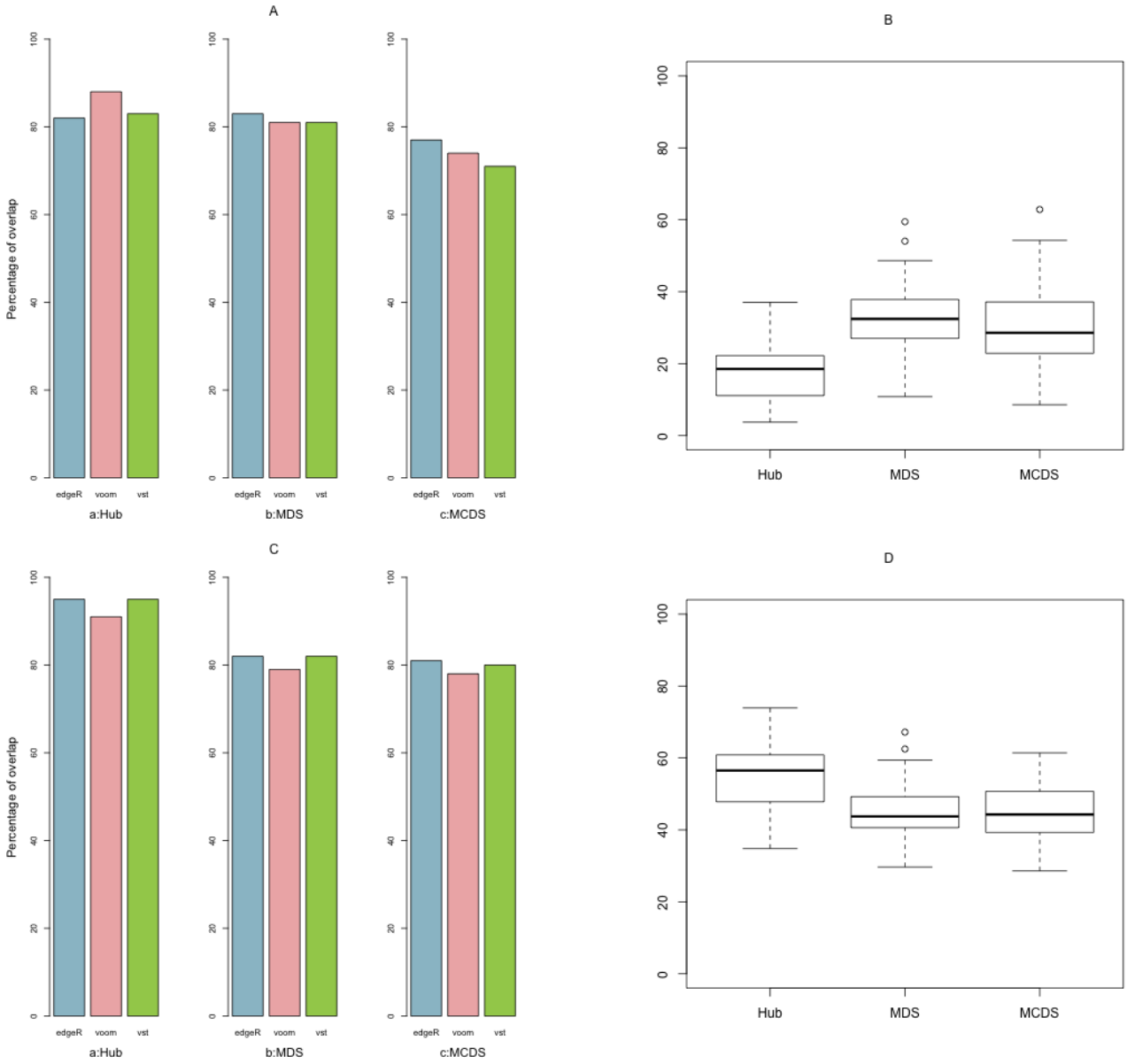

Figure 6.3: Panels (A) and (C) illustrate the percentage overlap of hubs, MDS and MCDS in the DESeq network with other three (edgeR, voom and VST) networks for the LIHC and BRCA datasets, respectively. Panels (B) and (D) stand for the overlap of the three mentioned topological features of DESeq with 100 disease-specific networks derived of 11000 and 14000 randomly selected genes from the LIHC and BRCA datasets, respectively.

Chapter 7

Randomization Strategies Affect Motif Significance Analysis in TF-miRNA-gene Regulatory Networks

This chapter is based on our paper entitled "Randomization strategies affect motif significance analysis in TF-miRNA-gene regulatory networks" by Sepideh Sadegh, Maryam Nazarieh, Christian Spaniol and Volkhard Helms that was published in the journal of Integrative Bioinformatics (Sadegh et al., 2017).
Sepideh Sadegh extended the Cytoscape plugin by adding cascade motifs, added conserved method to the plugin (in java), performed the experiments and wrote the manuscript. Maryam Nazarieh proposed the computational motivation of the paper (the number of required iterations for edge switching in co-regulatory networks) and suggested to add cascade motifs, implemented conserved method (in R) and integrated it to TFmiR2 web server, analyzed the results and wrote the manuscript (I was advisor of Sepideh Sadegh under supervision of Prof. Helms). We used the Cytoscape plugin that was developed by Christian Spaniol as part of his PhD thesis. Prof. Volkhard Helms designed the study, wrote and edited the manuscript.
Gene regulatory networks constitute an abstract way of capturing the regulatory connectivity between TFs, microRNAs and target genes in biological cells. Here, we address the problem of identifying enriched co-regulatory three-node motifs that are found significantly more often in the real network than in randomized networks. First, we compare two randomization strategies, that either only conserve the global degree distribution of the entire network, or that also conserve the degree distributions of different regulatory edge types. We argue that the edge-type preserving method leads to biologically more meaningful results. Then, we address the issue how convergence of randomization can be measured. We show that after $3 * |E|$ edge swappings, converged motif counts are obtained and the memory of initial edge identities is lost.

7.1 Introduction

GRNs are typically formulated as directed graphs whereby nodes stand for target genes, TFs, and microRNAs and arcs stand for activating or repressing regulatory interactions. TFs either activate or repress the transcription of target genes. MicroRNAs typ-

ically induce the degradation of messenger RNAs of their target genes. Hence, modern GRNs address the regulation of messenger RNA levels at the transcriptional and post-transcriptional levels (Hamed et al., 2015a; Zhang et al., 2015a). Our group recently introduced a web server termed TFmiR (Hamed et al., 2015a) that enables users to construct and analyze disease-specific TF and miRNA co-regulatory networks. Please see the chapter 4 for more details on TFmiR.
Shen-Orr and Alon were the first to identify regulatory motifs in a GRN of E.coli that only consisted of TFs and target genes (Shen-Orr et al., 2002). They discovered that feed-forward loops (FFLs) involving two TFs whereby TF1 regulates TF2 and both TFs jointly regulate a target gene are statistically significantly enriched in real GRNs with respect to randomized GRNs. Besides, they also discovered that single-input modules and densely overlapping regions are enriched too, but we will focus on FFL-type motifs here. Recently, several authors have expanded the concept of FFL-motifs to GRNs with TFs, miRNAs and target genes (Megraw et al., 2013; Zhang et al., 2015a; Hamed et al., 2015a). In this context, proper randomization of GRNs becomes even more important for determining which FFL motifs are enriched in the real GRN. In our original TFmiR paper, we did not distinguish between the three possible types of regulatory links, TF $\rightarrow$ gene, TF $\rightarrow$ miRNA, and miRNA $\rightarrow$ gene, during randomization. However, Ohler and co-workers recently pointed out that an edge-type preserving randomization strategy may be beneficial whereby switching of arc endpoints takes place only between two arcs that both of which belong to either one of the three groups of regulatory links (Megraw et al., 2013).
Another important technical question is how to properly quantify proper randomization. In our original TFmiR paper, we randomized $2 * |E|$ times, whereby E is the number of links in the GRN. It was argued that $100 * |E|$ switches of edge end points ensure proper randomization (Milo et al., 2004). Based on two GRNs with different link densities, we present here a thorough analysis of which motifs are statistically enriched in these GRNs under the edge-type conserving and non-conserving randomization strategies and how proper randomization can be quantified. For comparison, we also used the established motif-discovery tool FANMOD (Wernicke and Rasche, 2006).

7.2 Related Works

There exist many motif finding tools including the well-known tools mfinder (Kashtan et al., 2004) and FANMOD (Wernicke and Rasche, 2006). mfinder detects network motifs either by full enumeration of subgraphs, or by sampling of subgraphs for estimation of subgraph concentrations. The latter method is faster but has a bias in favor of sampling certain subgraphs more frequently than others (Wernicke, 2005). mfinder provides several methods to generate random networks including the switching method, the stub method, and the "go with the winners" algorithm (Milo et al., 2004). FANMOD uses an algorithm called RAND-ESU (Wernicke, 2005) that enables quick and accurate estimation of the total number of size-k subgraphs in a given network. A new randomization algorithm named WaRSwap (Megraw et al., 2013) provides a practical network motif discovery method for large multi-layer networks such as co-regulatory networks. However, this technique must be used together with a motif discovery tool such as FANMOD, which limits its applicability. WaRSwap generates randomized networks by preserving the indegree distribution of target nodes with respect to each source-target type rather than the exact indegrees.

Table 7.1: Density of BC-complete, BC-disease and GBM networks

	$\|E\|$	$\|V\|$	density
BC-complete	378	258	0.0057
BC-disease	297	206	0.0070
GBM	4248	408	0.0256

This randomization method seems to be more compatible with multi-layer networks than the universal method where only the in and outdegree of nodes are conserved.

7.3 Materials and Methods

7.3.1 Types of 3-node Motifs in miRNA-TF Synergistic Regulatory Networks

miRNA and TF co-regulatory networks contain four types of regulations, TF → Gene, TF → miRNA, miRNA → Gene, and miRNA → TF, that can be combined in ten different ways as 3-node motifs, see Figure 2.1. Eight of these are synergistic motifs consisting of two different types of regulators (miRNA and TF), and their directly/indirectly synergistically regulated target gene (first two rows of Figure 2.1). The last two motifs, where the target gene is not cooperatively regulated, are not studied here.

7.3.2 Datasets

We used miRNA and TF co-regulatory networks for two different complex diseases as input to our motif finding tool. The first network is associated with breast cancer (BC) (Hamed et al., 2015a) and the second network with glioblastoma multiforme brain tumor (GBM) (Sun et al., 2012). Table 7.1 lists topological properties of the two networks. The GBM network is about four times denser than the BC networks. In a study on breast cancer using gene and miRNA expression data from the The Cancer Genome Atlas (TCGA) portal, Hamed et al. (Hamed et al., 2015b) identified 1262 genes and 121 miRNAs that are deregulated in cancer tissue with respect to matched normal tissue. With the TFmiR web server (Hamed et al., 2015a) we identified regulatory interactions for the provided lists of up- and down-regulated genes and miRNAs using data from established and curated regulatory databases of both predicted and experimentally validated interactions. The resulting network is termed BC-complete in table 7.1. Then we used TFmiR to intersect this global network with genes associated with breast neoplasms based on the human miRNA disease database (HMDD) (Lu et al., 2008) and DisGeNET, a database for gene-disease association (Bauer-Mehren et al., 2010). This gave the breast cancer-specific subnetwork that we termed BC-disease.
A co-regulatory network for GBM with 415 genes and 124 mature GBM-related miRNAs was retrieved from Sun et al. (Sun et al., 2012), who used a similar approach for constructing GRN to the approach used in TFmiR. They retrieved 428 human TFs from the TRANSFAC database (Matys et al., 2006) and predicted the regulatory interactions between a TF, an miRNA and a gene using computational approaches.
The main difference between the construction of the two networks considered here is in the last step. In the GBM network, the authors included only miRNA-TF co-occurring pairs

that are significant based on the hypergeometric test. In contrast, TFmiR does not check for significance here. Another difference is that in building the GBM-specific co-regulatory network only predicted interactions were utilized, while in the BC-complete/disease co-regulatory networks both predicted and experimentally validated interactions were taken into account.

7.3.3 Motif Discovery Process

The steps used for the motif discovery are as follows: 1) A subgraph census is conducted for the types of desired motifs on the original network. 2) An ensemble of N similar random networks is generated and subgraph enumeration is applied to each of these networks. 3) Finally, after calculating the frequency of each type of subgraph in all networks (original and randomized), its significance metrics are calculated, with the over-represented subgraphs being reported as motifs. We implemented the entire process of motif finding as an in-house Cytoscape App (Shannon et al., 2003), which is an OSGi Bundle style App. This functionality will be made publicly available in the next release of TFmiR.

Enumeration of desired subgraphs

Typical algorithms for enumeration of subgraphs work on a connectivity matrix C, whose elements (C_{ij}) are equal to 1 if regulator i regulates target j and 0 otherwise. Then, they scan all n by n submatrices of C, that represent topologies of each desired type of size n motif. We modified this typical subgraph enumeration algorithm by using the data models in Cytoscape (namely CyNetwork and CyTable).

Generating Random Networks

Randomization of networks must be conducted such that sampling is performed as uniformly as possible from the collection of all obtainable random networks. (Megraw et al., 2013) suggested that low-variance distributions of motif counts in randomized networks are a sign of inadequate randomization, and that they can happen due to edge switching in large multi-layer networks. To evaluate the adequacy of sampling and uniformity of randomly generated networks, variances of the subgraph counts of all types of possible motifs in the randomized networks should be considered (see section 7.4.3).
The key aspect in assessing the statistical over-representation of motifs is to generate the random networks such that their characteristics are as similar as possible to the original network. The method using swapping of endpoints ensures that each node in the randomized networks has the same number of incoming and outgoing edges (in and outdegree) as the corresponding node in the real network. The universal method used for this purpose is the so-called switching method, employed for the first time in the field of motif detection by (Shen-Orr et al., 2002). By construction, this method strictly conserves the degree distribution of the graph and even of each node. The algorithm generates a Markov chain of states by randomly selecting a pair of edges ($A \rightarrow B$, $C \rightarrow D$) and swapping their endpoints to create the new edges ($A \rightarrow D$, $C \rightarrow B$). Creation of self-edges and multiple edges are not allowed and considered as failed attempts of switching. This process is repeated $Q * |E|$ times, where $|E|$ is the number of edges in the graph and Q is chosen large enough so that sthe Markov chain shows good mixing. (Milo et al., 2004) found

that for many networks, values of around $Q = 100$ appear to be more than adequate. In our approach failed attempts are not counted, i.e. we repeat swapping as many times as needed to reach $Q * |E|$ times successful attempts. This algorithm returns a shuffled version of the original network as a randomized network. We suggest some measures to assess proper mixing in section 7.3.4.

Different Strategies

We modified the current switching method to consider additional features of miRNA-TF synergistic regulatory networks (different node and edge types). To deal with networks with multiple types of connections, we use the terminology introduced by (Yeger-Lotem et al., 2004) where the extended degree of a node stands for the number of edges per type that point to/from a node. Two nodes have the same extended degree if they have the same number of incoming and outgoing edges for each edge type.

Based on this definition, one can develop a new switching strategy, which allows only swapping endpoints of edges with the same regulatory relationship among miRNAs, TFs and target genes. Hence, we distinguish a conserving method that conserves the extended degree of nodes, i.e. edges are switched only between edges of the same type, and a non-conserving method that does not conserve the extended degree of nodes, i.e. switching is performed without considering the edge type, consequently the frequency of each edge type is not conserved. This method is equivalent to the original switching method. Note that the non-conserving method can also create new edge types, which did not exist in the original network, unless this is prevented (such as TF $\rightarrow$ TF, or miRNA $\rightarrow$ miRNA edges).

An efficient algorithm for the conserving method can be implemented by grouping network edges of different edge types into different lists and then randomly selecting the second edge from the edge list of the first selected edge type. This helps to improve the efficiency of the randomization algorithm in terms of runtime.

Comparison of Real and Randomized Networks by Significance Metrics

The goal of network motif discovery is to determine which subgraph types occur in the original network at significantly higher frequencies than in random networks. For this purpose, the occurrence of a particular subgraph in the network of interest is compared to the distribution of counts for the same subgraph over a set of randomized networks using p-value and z-score. The p-value represents the probability of a motif to appear an equal or greater number of times in a random network than in the original network (Milo et al., 2002). This probability should be smaller than a determined probability threshold to reject the null hypothesis. This can be empirically determined using a large number of randomized networks:

$$p\text{-value} = \frac{N_{rh}}{N_t}$$

where N_{rh} is the number of random networks in which a certain motif type is acquired more than or equal to its number in the real network and N_t is the total number of randomized networks. It has been suggested (Hamed et al., 2015a) that 100 is sufficiently large. Alternatively, let f_{real} be the frequency in the real network and f_{rand} be the frequency in a random network. We can then define the z-score as follows (with σ being the standard deviation):

$$z\text{-score} = \frac{(f_{real}\text{-}\bar{f}_{rand})}{\sigma(f_{rand})},\ \sigma(f_{rand}) = \sqrt{\frac{\sum_i (f_{rand_i}\text{-}\bar{f}_{rand})^2}{N_t}}$$

Subgraphs with z-score ≥ 2 and p-value ≤ 0.05 are considered significant motifs as was previously performed (Shen-Orr et al., 2002; Wernicke and Rasche, 2006; Megraw et al., 2013).

7.3.4 Measures for Proper Mixing of Randomized Networks

One general drawback of randomizing networks by the switching method is that there is no measure of how long one needs to iterate over the select two edges and swap their endpoints routine to attain well randomized networks. Here we propose two measures to characterize whether randomized networks are properly mixed.
First, we measure the similarity of networks before and after randomization. Ideally, edges should be switched until there are no common edges between the original network and each randomized network. In other words one should search for the maximum difference between the given network and each randomized network to avoid situations of under-shuffling (Liang et al., 2015). Under-shuffling means that only a small fraction of the switchable edges were swapped. We defined a similarity metric to measure how similar is the ensemble of randomized networks to the original network in terms of common edges:

$$Similarity = \frac{<Sim>}{|E|}$$

Here, *Sim* is the number of common edges between the original and a particular randomized network, $<Sim>$ is its average in all randomized networks, and $|E|$ is the total number of edges in the original network. Lower *Similarity* values indicate better randomization. The lowest possible value of zero happens in case of no common edges. This definition considers the size of the network as well as the number of randomized networks. This enables comparison of the similarity metrics of randomization approaches applied to different given networks. A value close to zero indicates that under-shuffling is avoided.
Another measure is the convergence of subgraph counts during randomization. For $0.01*|E|$ to $100*|E|$ randomization iterations, we recorded how often the investigated subgraph types occurred in the random networks and checked whether this number converged to a specific value or whether it did not follow any pattern and changed erratically.

7.4 Results

7.4.1 Synergistic 3-node Motifs

Table 7.2 shows which co-regulatory 3-node motifs were significantly enriched in the real GRN vs. randomized GRNs when either the edge-type conserving randomization strategy was applied or the non-conserving one. $100 * |E|$ iterations were used for this part of our study. In the BC-complete network, no significant motif is found by the conserving method. In contrast, the composite-miRNA-mediated and cascade-miRNA- mediated motifs are reported as significant by the non-conserving method. In the BC-disease network, only the co-regulation type is identified as significant by the conserving method, whereas the non-conserving method gives the same significant motifs as for the BC-complete network. In the GBM network, TF-FFL and miRNA-FFL are reported as significant by

Table 7.2: p-values for different 3-node subgraphs in the considered networks when either the non-conserving or the conserving randomization strategy is used.

	BC-complete		**BC-disease**		**GBM**	
Subgraph type	Non-cons.	Cons.	Non-cons.	Cons.	Non-cons.	Cons.
Co- regulation	0.77	0.33	0.96	**0.04**	1.00	1.00
TF-FFL	1.00	0.77	0.98	0.61	**0.00**	**0.00**
miRNA- FFL	0.77	0.86	0.69	1.00	**0.00**	**0.00**
Composite- FFL	0.29	0.12	0.45	0.50	**0.00**	0.68
Composite- TF-Med.	0.62	0.26	0.66	0.42	**0.00**	0.55
Composite- miRNA-Med.	**0.00**	0.50	**0.01**	0.53	**0.00**	0.62
Cascade- TF-Med.	0.47	0.69	0.16	0.26	**0.00**	0.84
Cascade- miRNA-Med.	**0.00**	0.54	**0.01**	0.49	**0.00**	1.00

the conserving method; whereas by the non-conserving method all types of subgraphs except co-regulation are identified as significant. In all three networks, subgraphs of types composite-miRNA-mediated and cascade-miRNA-mediated are identified as statistically significant by the non-conserving method. All subgraphs meeting the p-value criterion in Table 7.2 also met the z-score criterion. Note that p-value is a probability and can be slightly different in each run of the algorithm due to the generation of different randomized networks.

The non-conserving method leads to detecting more subgraph types as significant compared with the conserving method. In randomization by the conserving method, swapping happens between all edges of the same type, hence the chance of having the same types of subgraphs in randomized networks compared to the original network is not decreased that much. This could result in higher p-values and consequently fewer subgraphs will show significant differences.

In the GBM network, we found 3-node FFLs to be significant while in BC networks they are not. One reason for this could be the higher density of the GBM network than the BC networks.

7.4.2 Motif Finding with FANMOD

FANMOD also employs the "switching method" for randomization of network. The randomization step can optionally keep the color degree of a vertex constant by exchanging edges only with edges of the same type. This is equal to randomization by the conserving method in our approach. The same number of swappings per edge ($Q = 100$) and the same criteria for p-value and z-score were chosen for both tools. FANMOD gave similar motif finding results for all three miRNA-TF co-regulatory networks (Figure 7.1) to those of our tool. The few dissimilarities can be due to slight differences of the randomization algorithms. In the routine of randomly selecting edges for swapping, we only count successful attempts until a pre-defined number of iterations is reached whereas FANMOD tries a limited pre-defined number of times to find an appropriate candidate for swapping irrespective of whether this is successful or not. By inspecting the output file of FANMOD we found that $\sim 80\%$ of the attempts were successful by randomization with the conserving method and $\sim 50\%$ with the non-conserving method.

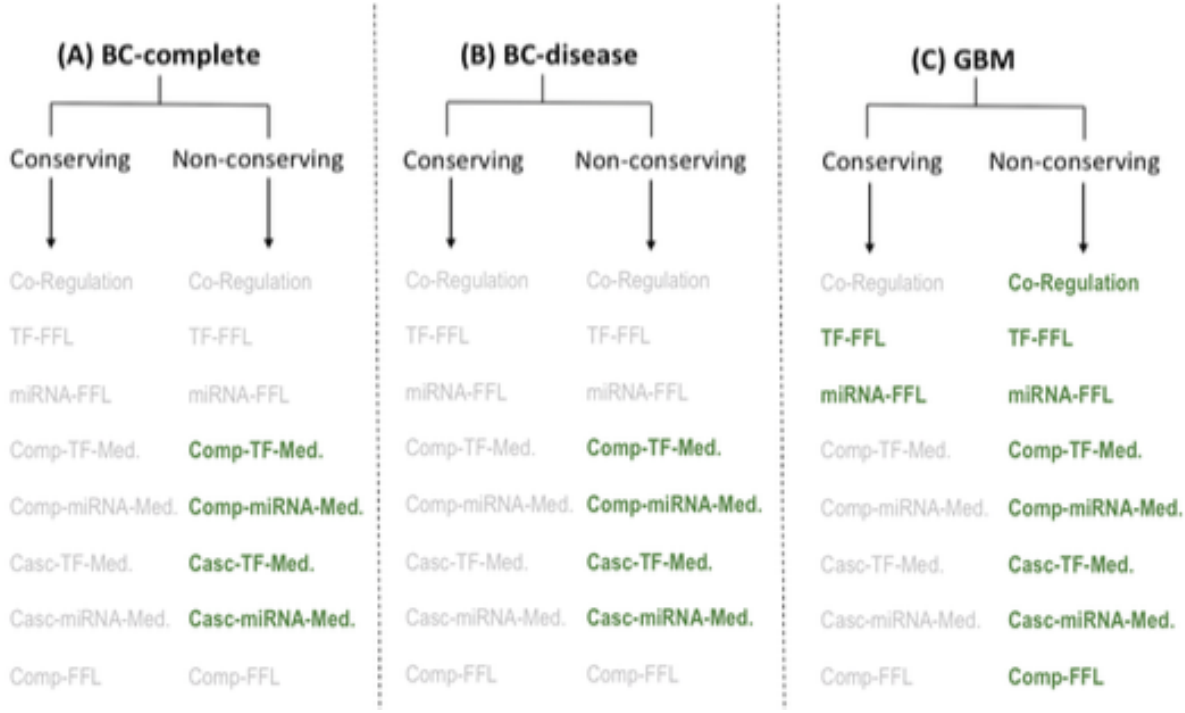

Figure 7.1: Significant 3-node motifs (highlighted in green) detected by the FANMOD tool with two different randomization strategies. (A) BC-complete. (B) BC-disease. (C) GBM.

7.4.3 Validation of Randomization

Uniform Sampling of Randomized Networks

(Megraw et al., 2013) observed many failed switches during the execution of FANMOD randomization, which was also the case here. As mentioned, our approach counts only the number of successful attempts until a pre-defined number of iterations is reached. By close inspection of the resulting background histogram of significant motifs, we observed in the miRNA-TF synergistic regulatory networks of BC and GBM a high variance of count distributions of subgraphs in randomized networks for all significant 3-node motifs, see Table 7.3 and Table 7.4. This indicates an adequate randomization of networks in our approach. For the GBM network, much higher variances were obtained than for the BC networks. It is suggestive to attribute this to the higher density of the network.

Measures for Proper Mixing of Randomized Networks

Similarity Metric

Two sets of 100 randomized networks were generated from the BC-complete network using in one case the edge-type conserving strategy and in the other case the non-conserving randomization strategy. Between $0.01 * |E|$ and $100 * |E|$ iterations of edge swapping ($Q*|E|$) were carried out. Figure 7.2 shows the similarity between original and randomized GRNs for varying Q. The number of iterations required to reach values close to zero depends on the randomization strategy. For the non-conserving method, the similarity metric reaches zero at fewer iterations ($Q = 7$ for BC-complete and $Q = 8$ for GBM) than

Table 7.3: Variance of count distributions of subgraphs in randomized networks for the BC-complete/-disease networks. Significant motifs are marked in bold.

	Non-conserving method		**Conserving method**	
Subgraph type	Variance		Variance	
	BC-disease	BC-complete	BC-disease	BC-complete
Co- regulation	14.9	32.0	**6.7**	10.3
TF-FFL	5.5	7.5	2.4	2.6
miRNA- FFL	4.2	4.9	7.0	8.4
Composite- FFL	1.7	1.5	2.0	1.7
Composite- TF-Med.	298.8	339.1	177.6	130.0
Composite- miRNA-Med.	**85.1**	**74.0**	391.9	458.5
Cascade- TF-Med.	592.3	880.8	174.8	134.4
Cascade- miRNA-Med.	**394.6**	**433.4**	435.9	522.6

Table 7.4: Variance of count distributions of subgraphs in randomized networks for GBM network. Significant motifs are marked in bold.

	Non-conserving method	**Conserving method**
Subgraph type	Variance	Variance
Co- regulation	18,335.7	4729.6
TF-FFL	**4340.9**	**3298.5**
miRNA- FFL	**1526.1**	**2344.3**
Composite- FFL	**277.0**	890.4
Composite- TF-Med.	**10,282.2**	38,169.3
Composite- miRNA-Med.	**1690.0**	6153.5
Cascade- TF-Med.	**53,127.1**	37,486.1
Cascade- miRNA-Med.	**35,798.9**	7742.2

the conserving method ($Q = 14$ for BC- complete and $Q = 15$ for GBM). Results for the GBM network are very similar to those for the BC-complete network, only slightly more iterations are needed to reach zero. Both methods of randomization for both networks reach similarities below 0.01 after $Q = 3$ iterations. This means that after $3*|E|$ iterations, less than 1 % of the edges in the ensemble of randomized networks are in common with the original network. This low percentage of similarity seems to be a good threshold for choosing a proper Q for our randomization method.

Convergence of Subgraph Counts

Figure 7.3 shows how often subgraph types occurred in the set of randomized BC-complete networks after randomization when Q was varied between 0.01 and 100. With the non-conserving method (Figure 7.3), the total subgraph count converged to a fixed value after $Q = 10$ iterations and did not change erratically thereafter. With the conserving method (Figure 7.3), the total number of subgraphs found in the randomized networks was quite stable over the whole range of $0.1 < Q < 100$.

Our empirical findings for both BC and GBM networks suggest that $Q = 1$ is adequate to obtain properly mixed randomized networks by the conserving method; whereas for the non-conserving method $Q = 10$ appears suitable for ensuring good mixing of the randomized networks. Evaluation of network similarity for the BC-complete network suggests $Q \sim 3$ as a good balance for both conserving and non-conserving methods of randomization.

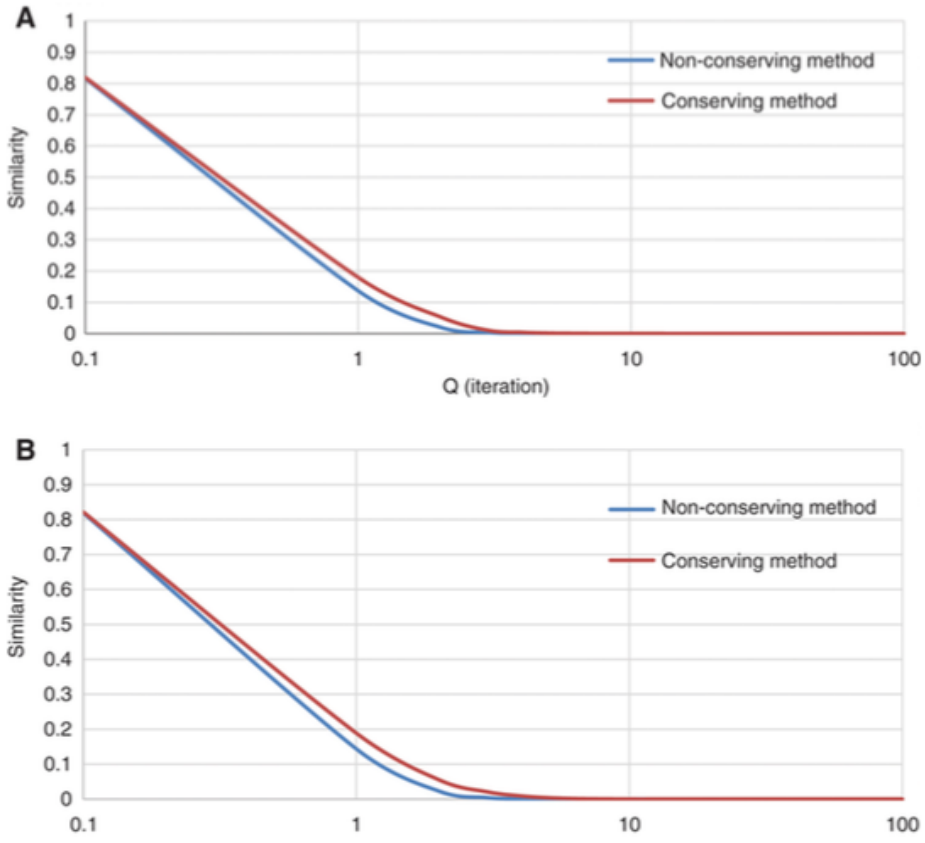

Figure 7.2: Similarity metric vs. number of iterations for (A) the BC-complete and (B) the GBM networks.

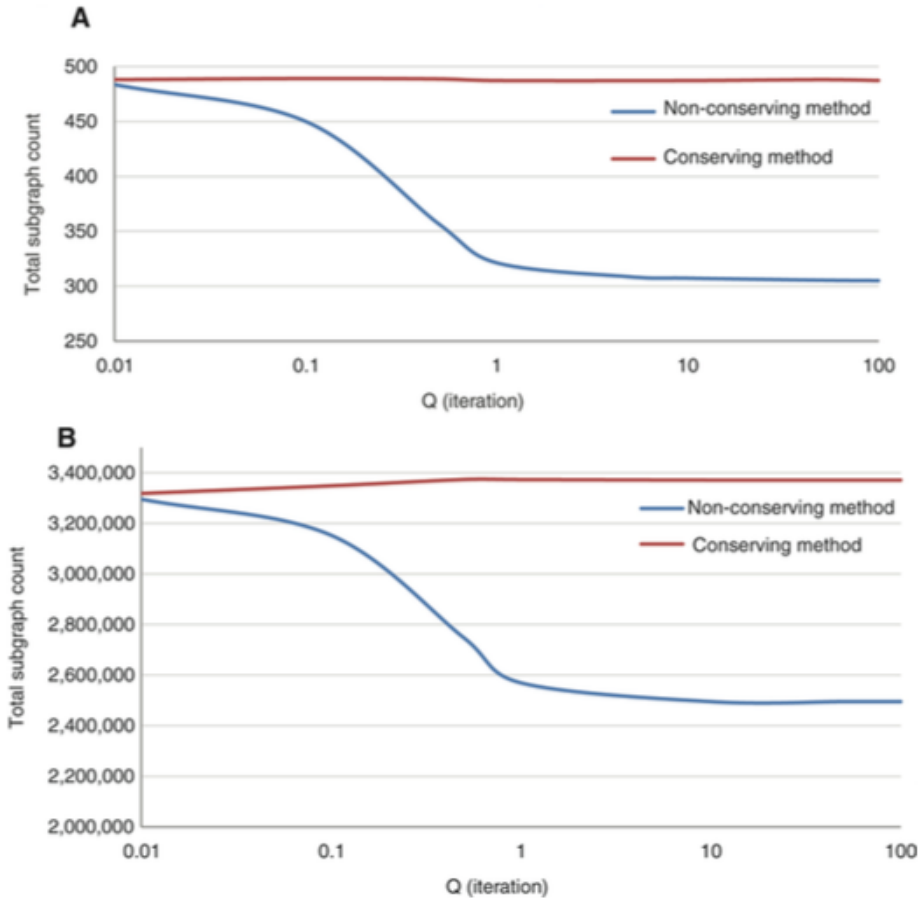

Figure 7.3: Total number of subgraphs vs. number of iterations for (A) the BC-complete and (B) the GBM networks.

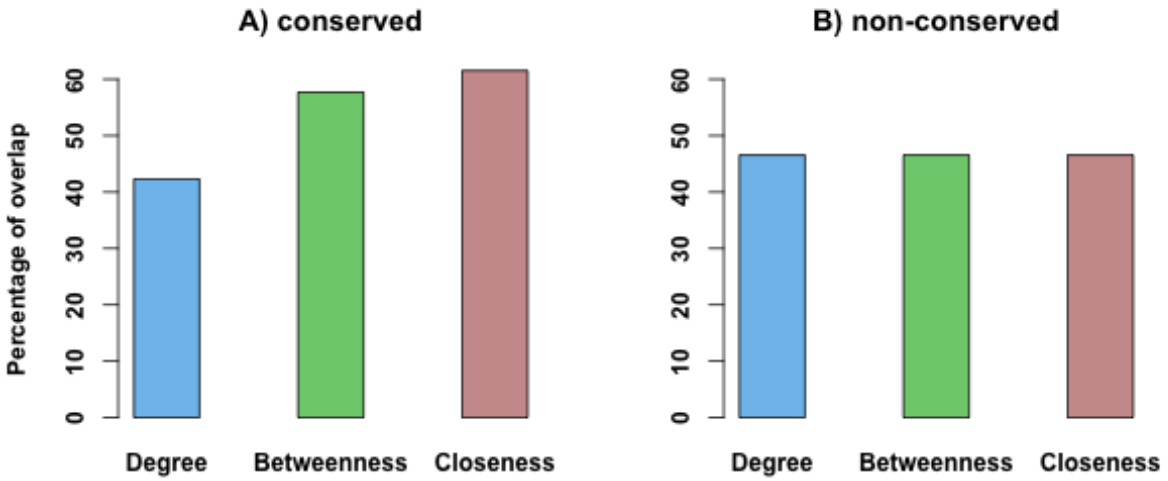

Figure 7.4: Overlap of most central nodes (according to three different centralities) with the set of genes and miRNAs in the statistically enriched motifs. (A) Conserving. (B) Non-conserving.

7.4.4 Network Centrality of Gene and miRNA Sets

Here, we analyzed the overlap between the genes and miRNAs participating in the enriched 3-node motifs (here termed motif nodes) and the most central genes and miRNAs with respect to degree, betweenness, and closeness centralities (here termed central nodes). Using either edge-type conserving or non-conserving randomization gave 26 and 130 genes and miRNAs in enriched motifs, respectively. These sets were compared to sets with the same number of most central genes and miRNAs. The centralities were measured using the igraph package (Csardi and Nepusz, 2006), considering only outdegree of nodes in the directed network. The motif nodes identified by the conserving method had the highest overlap with the central nodes defined according to closeness and betweenness centrality, respectively, see Figure 7.4 panel (A). In contrast, the motif nodes defined with the non-conserving method showed a similar overlap with the central nodes identified by all three centralities, see Figure 7.4 panel (B). This latter observation can be explained by noting that only 57 genes and miRNAs have outdegree greater than or equal to 1 in the BC disease networks. The overlap of around 45 % with the central nodes means that essentially all these 57 motif nodes are hub nodes in this network. A larger fraction of hub nodes (up to around 60 %) exists in the smaller set of 26 motif nodes defined by the conserving method.

Next, we analyzed the overlap of the 26 motif nodes identified by the conserving method with a MCDS of key regulatory genes and miRNAs that dominates the network. For this, we solved the ILP formulation of the respective MCDS in the LSCC of this network (Nazarieh et al., 2016) and obtained an MCDS of seven genes and miRNAs. Among these, TGFB1, TP53, ESR1, and hsa-mir-22 belong to the motif nodes.

7.4.5 Biological Relevance of the detected Motifs

The biological relevance of the genes among the motif nodes obtained by the conserving and non-conserving randomization methods was evaluated based on the functional categories in GO Direct using the enrichment analysis via DAVID (version 6.8) (Huang et al., 2009). p-values below the threshold 0.05 obtained by the hypergeometric test were adjusted for multiple testing using the BH procedure (Benjamini and Hochberg, 1995). Both methods returned almost the same number of significant GO terms, mostly involving transcription and apoptotic processes, although the non-conserving method considered 104 genes versus 14 genes considered by the conserving method, see Tables S28 and S29.

7.5 Summary and Discussion

If the network of interest contains more than one node or edge type, different randomization strategies can be applied for motif discovery. In this study, different strategies led to quite different enriched 3-node motif types.
The reason why FFLs were statistically significantly enriched only in the GBM network could originate from the difference in constructing the GBM and BC networks, where only significant TF-miRNA co-occurring pairs were considered in the regulatory network of GBM. This means that the TF $\rightarrow$ gene $\leftarrow$ miRNA triad is enriched a priori in this network. Our study suggests that the way of network construction and also the density of the network may affect the results of motif finding. For the considered BC-networks, only subgraphs of types other than FFLs were found to be significantly enriched. Our motif finding tool identified composite-miRNA-mediated and cascade-miRNA-mediated as statistically significant motifs (by the non-conserving method). Although the results are similar in BC-networks, the conserving method identified the co-regulation motif type to be significant in the filtered BC-disease network that was not found significant in the BC-complete network. We thus speculate that motif searches in filtered (i.e. more specific) networks may identify biologically more meaningful motifs.
We suggest variance of motif counts and similarity of original and randomized networks as suitable auxiliary measures to judge whether randomization generates properly mixed networks. Our study suggests that the density of networks does not affect the minimum required Q to obtain properly mixed randomized networks. In conclusion, the non-conserving method leads to detecting more subgraph types as being statistically significant compared with the conserving method. For the 2.5 networks studied here, we noticed that (a) the conserving randomization method identified significant motifs containing a larger fraction of the most central nodes (Figure 7.4) than the non-conserving method, and (b) both methods gave the same number of significant GO terms, although the conserving method considered much fewer genes for this than the non-conserving method. Certainly, the same analysis should be extended to a representative number of comparable GRNs. So far, it seems that the conserving method gives biologically more meaningful results.

Chapter 8

TFmiR2: Constructing and Analyzing Disease-, Tissue- and Process-specific TF and miRNA co-regulatory Networks

This chapter is based on the manuscript entitled "TFmiR2: Constructing and analyzing disease-, tissue- and process-specific transcription factor and miRNA co-regulatory networks" by Nazarieh et al. (2017). The abstract of the work was published earlier as mentioned in (Nazarieh et al., 2017).
Maryam Nazarieh developed both frontend and backend of the web server and extended the TFmiR databases and wrote the manuscript. Mohamed Hamed provided the mouse database. Christian Spaniol edited the frontend programs. Thorsten Will provided the tissue data for human and mouse and disease-associated genes for mouse. All authors including Prof. Volkhard Helms designed the study and edited the manuscript.
TFmiR2 is a freely available web server for constructing and analyzing integrated TF and miRNA co-regulatory networks in human and mouse. Due to the availability of genome scale data sets, the challenge is no longer to generate large regulatory networks, but rather to determine the parts that are essential to a scientific question. TFmiR2 helps to solve this issue by generating tissue- and biological process-specific networks as well as networks with multiple specificity for the set of deregulated genes and miRNAs provided by the user. Furthermore, the service can now aid the user to identify key driver genes and miRNAs in the constructed networks by utilizing the graph theoretical concept of a MCDS. Especially when combined with experimental validation, these putative key players as well as the newly implemented 4-node TF-miRNA motifs can potentially promote novel insights that may assist in developing new therapeutic approaches through identification of significantly enriched patterns of interactions between the components.
Availability: The TFmiR2 web server is available at `http://service.bioinformatik.uni-saarland.de/tfmir2`.

8.1 Introduction

The regulatory networks involving TFs, genes and miRNAs control all cellular processes and define tissue specificity (Nazarov et al., 2013). They are also tightly associated with

cellular malfunctions and disease pathways (Lee and Young, 2013). Therefore, construction and analysis of the corresponding regulatory networks are central issues in systems biology (Guzzi et al., 2015; Georgakilas et al., 2016). The predecessor of this new service, the TFmiR web server (Hamed et al., 2015a) enables integrative analysis of combinatorial regulatory interactions between TFs, miRNAs and deregulated target genes. TFmiR has been successfully used in diverse areas of biological research, e.g. in hematopoiesis (Hamed et al., 2017). We present here the significantly expanded new version TFmiR2. Added core features comprise the support for mouse data, confinement to tissue- and process-specific subnetworks, detection of 4-node motifs, and driver gene detection based on the topology of the deregulated networks by, for example, the MCDS algorithm (Nazarieh et al., 2016). Besides the disease-specific features of the original service, TFmiR2 now enables the user to also select multiple gene ontology terms related to biological processes as filters to generate process-specific networks around the set of user-specified deregulated genes and miRNAs. Such networks can then be even further filtered to specific tissues to arrive at tissue- and process-specific networks.
These new additions enable users to contextualize the data in a more specific setting than before. This could, for example, reveal the regulatory dependencies among the user-specified deregulated genes and/or miRNAs in a tissue of choice, or alternatively, to consider the mutual effect of this deregulated set in a confined and potentially aberrant biological process which ultimately causes a disease. Furthermore, the service can now aid the user to identify key driver genes and miRNAs in the constructed networks by utilizing the graph theoretical concept of the MCDS (Nazarieh et al., 2016). Such drivers could possibly drive the pathogenic processes of diseases and thus appear potential drug targets. TFmiR2 detects 4-node TF-miRNA motifs described in (Sun et al., 2012) by considering different randomization methods as discussed in (Sadegh et al., 2017).

8.2 Methods

The basic workflow in TFmiR2 consists of three steps: 1) data upload and processing, 2) construction of the networks, 3) network visualization and downstream analysis. Figure 8.1 depicts the overall workflow of TFmiR2. For the set of data provided by a user, a variety of options exist to generate networks related to a user-specified biological process, tissue and disease as well as combinations of them. In TFmiR2, the user can select a tissue, disease and related process and also restrict the regulatory interactions that are included to either experimental, predicted or both data sources. In addition to the disease-specific network and the full interaction network which were already featured in TFmiR, TFmiR2 constructs tissue-, process-, disease-tissue-, disease-process, and tissue-process specific networks.

8.2.1 Functionality of TFmiR2

TFmiR2 combines the databases for human previously used in TFmiR with the regNetwork database (Liu et al., 2015) which integrates gene regulatory networks from 25 selected databases as described in (Liu et al., 2015). If a tissue is selected, the full interaction network is filtered to a tissue-specific subnetwork by confining the interactions to those where both interaction partners are expressed in the selected tissue. This restriction is relaxed for miRNAs since we did not include evidence on tissue-specificity of

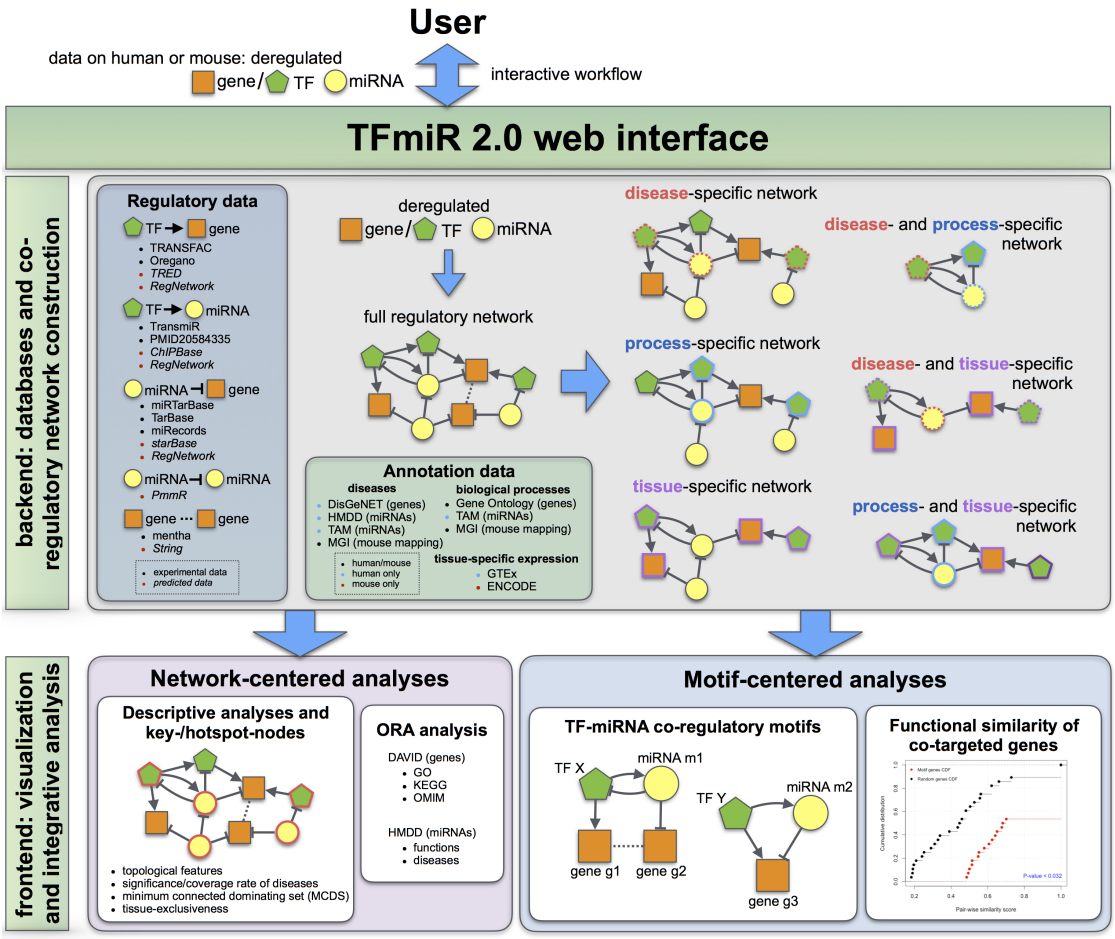

Figure 8.1: A system level overview of the TFmiR2 architecture describing the incorporated databases (top left), workflow and output downstream analysis.

miRNAs. In Process-specific networks that are derived from the full interaction network, we demand that at least one interactor is related to the selected biological process(es). When multiple processes are selected, the constructed network is even capable to demonstrate the cross-process interactions. The process of deriving a disease-specific network was described in (Hamed et al., 2015a), see chapter 4. Moreover, TFmiR2 can generate a disease-process specific network that corresponds to the set of mutual interactions between a disease gene and a potentially aberrant biological process. The interactions in the network are selected such that either regulator nodes are known to be associated with the selected disease found in (Bauer-Mehren et al., 2010) and target nodes are known to be related to the user-specified biological process(es) or vice versa. To limit the effect of disease genes to a specific tissue, disease-tissue specific networks can be constructed. Then the interactions in the network are selected such that either regulator or target nodes are known to be associated with the selected disease, but both regulator and target nodes are known to be expressed in the selected tissue. The process-specific network may be further confined to a tissue-process specific network by considering the interactions whose regulators and targets correspond to the user-specified tissue and biological process.

8.2.2 TFmiR2 user input Scenarios

TFmiR2 can be started from a set of deregulated miRNAs. In such a case, TFmiR2 identifies the set of genes whose target miRNAs as well as regulating miRNAs are significantly enriched within the input miRNAs using a hypergeometric test followed by BH adjustment (Benjamini and Hochberg, 1995). The user may also provide deregulated miRNAs and mRNAs or just a list of deregulated mRNAs as described in TFmiR (Hamed et al., 2015a). TFmiR2 enables users to determine the level of significance through adjusting the enrichment threshold for the set of genes and miRNAs.
Furthermore, the user may optionally set the p-value cutoff (default is 0.05) for ORA on the resulting network nodes (genes/miRNAs). Finally, the user can control the evidence level (experimentally validated, predicted, or both) for the constructed regulatory interactions that will be used in the subsequent network analysis.
TFmiR2 now enables the user to also select multiple GO terms (The-Gene-Ontology-Consortium, 2017) related to biological processes as filters to generate process-specific networks around the set of user-specified deregulated genes and miRNAs.

8.2.3 Identification of Network Key Nodes

In TFmiR2, a new method was added to find the key driver genes on the concept of based on network controllability (Nazarieh et al., 2016), see chapter 3. If the network includes a connected component, then the MCDS algorithm returns a set of genes and miRNAs that control the whole network, see Figure 8.2. Due to the large overlap between MDS and MCDS on the LCC as described in (Nazarieh et al., 2016), we focused here on the essential component of the network, the LSCC and applied MCDS on it. In a strongly

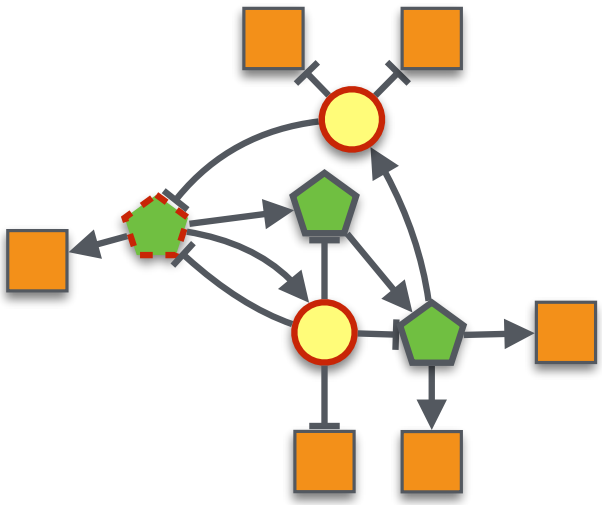

Figure 8.2: A graphical representation that illustrates the MCDS solution in the LSCC of an example TF and miRNA co-regulatory network. The red circle borders mark the one TF and two miRNAs belonging to the solution.

connected component of a GRN, there is a path from each node in the component to all other nodes. TFs and miRNAs constitute the dominators and connectors of the MCDS solution.

8.2.4 Tissue-exclusive Genes

Tissue-exclusive genes are the set of genes solely expressed in one tissue. TFmiR2 enables the user to study the interacting neighbours of tissue-exclusive genes to better understand the underlying regulatory network.

8.2.5 Identification of TF-miRNA 4-node Motifs

Integration of gene-gene and protein-protein interaction data from databases such as Mentha (Calderone et al., 2013) and String (Szklarczyk et al., 2015) enables TFmiR2 to detect motifs of size 4 in addition to study further types of co-regulatory interactions. Additionally, TFmiR2 allows a user to set a confidence threshold to retrieve the set of gene-gene and protein-protein interactions which exceed the threshold. TFmiR2 detects 12 different motifs of size 4 as shown in Figure 8.3 based on (Sun et al., 2012) where the gene-gene/protein-protein interactions are regulated by a TF and a miRNA. The co-regulation is defined by a regulatory pair of TF and miRNA (either in both directions or just in one direction) that co-regulate at most two target interacting genes.
In Figure 8.3 part a) Com/TF/miRNA-FFL, one of the target genes is regulated by a TF and just one of the target genes is repressed by a miRNA. If the TF regulates the miRNA and the miRNA represses the TF, then it is a Com-FFL. If just the TF regulates the miRNA, it is a TF-FFL and if just the miRNA represses the TF, it is a miRNA-FFL.
In Figure 8.3 part b) Com/TF/miRNA-FFL-miRNA, both target genes are regulated by a TF and just one of the target genes is repressed by a miRNA. If the TF regulates the miRNA and the miRNA repressed the TF, then it is a Com-FFL-TF. If just the TF regulates the miRNA, it is a TF-FFL-TF and if just the miRNA represses the TF, it is a miRNA-FFL-TF.
In Figure 8.3 part c) Com/TF/miRNA-FFL-miRNA, both target genes are repressed by a miRNA and just one of the target genes is regulated by a TF. If the TF regulates the miRNA and the miRNA represses the TF, then it is a Com-FFL-miRNA. If just the TF regulates the miRNA, it is a TF-FFL-miRNA and if just the miRNA represses the TF, it is a miRNA-FFL-miRNA.
In Figure 8.3 part d) Com/TF/miRNA-FFL-Full, both target genes are repressed by a miRNA and regulated by a TF. If the TF regulates the miRNA and the miRNA represses the TF, then it is a Com-FFL-Full. If just the TF regulates the miRNA, it is a TF-FFL-Full and if just the miRNA represses the TF, it is a miRNA-FFL-Full.

8.2.6 Conserved Randomization Strategy

In TFmiR2, the user can select a randomization strategy which can be either edge-type conserving or non-conserving, see chapter 7.3.3. The non-conserving method was realized in the first version of TFmiR. If the conserved method is selected, then swapping the edges is limited to the edge type. To do the swapping scheme, first all the interaction types (TF-gene, TF-miRNA, miRNA-gene, miRNA-miRNA and gene-gene) are grouped respectively. Then the endpoints are swapped inside each group. The number of swapping events for each group depends on the group size, e.g. if it is supposed to have totally N swappings in the network, then each group contributes proportionally.

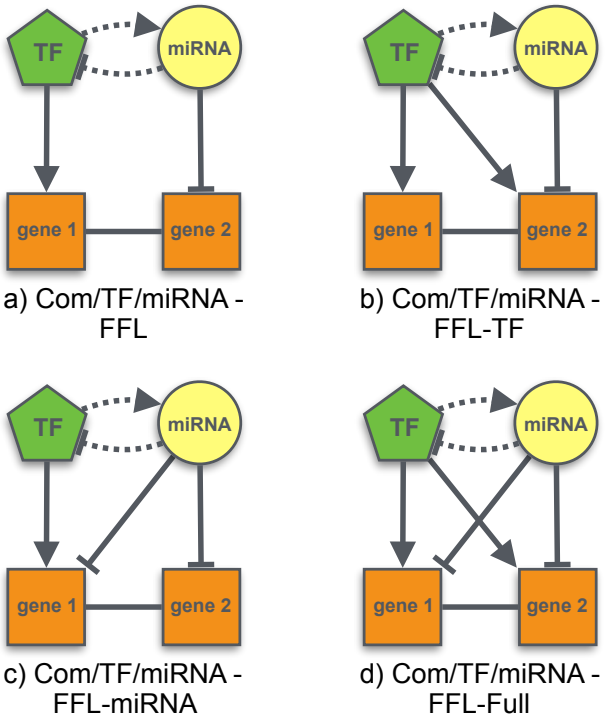

Figure 8.3: Schematic illustration of the four motif types detected in TFmiR2. All motifs contain a TF, a miRNA, and at least one common target gene.

8.2.7 Data Retrieval and Processing

The miRNA-process association files were downloaded from the TAM web service (The tool for annotations of human miRNAs) (Lu et al., 2010) and were matched literally to the textual IDs of the general GO terms. The gene-process association files were retrieved from UniProtKB (Ashburner et al., 2000; Barrell et al., 2009; The-Gene-Ontology-Consortium, 2017) with GOC validation date of 21.10.2017 for Go terms. Mapping of GO terms to UniProtKnowledgebase keywords was performed on 26.10.2017. Here we considered 42 biological processes that are commonly annotated in both miRNA and gene association files.
Tissue-specific genes in human were derived from the expression data of the GTEx project (release V6p) on 53 tissues (Mele et al., 2015). As in (Mele et al., 2015), a gene was considered as abundant in a sample when its RPKM value was above 0.1 and a gene was considered as abundant in a tissue if it was abundant in most of the samples of the respective tissue. Mouse data on tissue expression was inferred from ENCODE expression data on 30 tissues (release 3 of Sept 2012) (Pervouchine et al., 2015). Here, abundance of genes was determined according to the protocol of the identically quantified data in (Djebali et al., 2012). Thus, genes considered abundant in a tissue needed to be expressed (RPKM > 0) in both biological replicates of the corresponding tissue and have an npIDR value below 0.1. Several databases were used to generate networks for mouse. OregAnno (Griffith et al., 2008) was included for TFs regulating genes. TransmiR (Wang et al., 2010) and ChipBase (Yang et al., 2013) provided information on which TFs regulate miRNAs. miRTarBase (Hsu et al., 2010), TarBase (Sethupathy et al., 2006), miRecords (Xiao et al., 2009) and starBase (Yang et al., 2011) collect target genes of miRNAs. Gene-gene inter-

actions were adopted from mentha (Calderone et al., 2013) and String (Szklarczyk et al., 2015) databases. Moreover, our database for mouse was integrated with regNetwork (Liu et al., 2015) which includes interaction data from many databases. RegNetwork is an integrative database of transcriptional and post-transcriptional regulatory networks for human and mouse from 25 selected databases. It covers three types of experimentally validated and computationally predicted regulatory interactions including TF → gene, TF → miRNA and miRNA → gene. (Liu et al., 2015). The mouse genome database (MGD) is an integrative database of mouse genes and its features and functions. It is the core component of consortium of mouse genome informatics (MGI) (Smith et al., 2018). Mouse disease genes were mapped from human homologs using data of the MGI database.

8.3 Results

8.3.1 Case Study

To demonstrate the capabilities of the new features of TFmiR2, we considered the same dataset that was used in the earlier version of TFmiR (Hamed et al., 2015b). The PPI threshold was set to 80%. Cell proliferation, cell differentiation and apoptosis were selected from the list of processes. Tissue was set to Breast-Mammary Tissue. The MCDS algorithm was applied on the LSCC of the process-specific, tissue-specific and disease-specific networks. Out of 5 MCDS nodes, 4 nodes (ESR1, hsa-mir-221, PDGFA, TGFB1) were common among all the results. This indicates that these four nodes have significant roles in causing and maintaining the disease. Moreover, TFmiR2 detected multiple FFL motifs of size 4 in the disease-specific network. AAKT1, TP53, JAG1, PDGFA, NDRG3, FLI1, BRCA2, GNA13, ELK3 and COL15A1 were the dominators of the disease-process specific networks. These genes are associated mainly with the cell cycle and cell differentiation.

8.4 Summary and Discussion

TFmiR2 reinforces the earlier version, TFmiR, through involving tissue-and process-specific networks with whereby a disease-specific network can be restricted to a related tissue and multiple biological processes. These tailored new networks enable a user to analyse the mutual interaction between a disease gene and malfunctioned biological process, where the MCDS method identifies the putative key driver genes and miRNAs that control the LSCC of the network. Moreover, the co-regulatory PPI motifs represent the most significant genes, miRNAs and proteins with respect to the predefined pattern of interactions between them.

Chapter 9

Conclusions

The dissertation addressed several open problems in the research field of gene regulatory networks. Master regulatory genes are genes which either directly govern the particular cellular identity or are at the inception of developmental lineages and regulate a cascade of gene expressions to form specific lineages. The network in ESCs is spanned up by few connected master regulatory TFs which share many target genes. Slight changes in the expression levels of such a tightly interwoven network of transcription factors lead the ESCs into differentiation. Cancer stem cells show a very similar behavior as stem cells. They have the capability of generating an indefinite number of cells of the same type and cause different types of tumors. In this dissertation, we exploited the regulatory mechanisms underlying ESCs to identify cancer biomarkers. We addressed the problem by formulations of two combinatorial optimization problems which are based on the concept of network controllability. In the MDS approach, we search for a minimum number of dominators in directed graphs which control the whole network, whereas a MCDS returns a dominating pathway in the connected component of regulatory networks. Based on the success of these methods in achieving the master regulators which govern the global gene regulatory network of *E. coli* and the cell cycle genes of *S. cerevisiae* in addition to the gene regulatory network of mouse ESCs, we applied the methods to cancer networks. A significant set of breast cancer drug targets were found among a dominating pathway of the breast cancer network.

With respect to the fact that optimization problems usually do not have unique solutions and we needed to use a heuristic approximate approach of MCDS for solving some problem instances due to the long running time to achieve optimal solutions, a novel prioritization method was proposed. This method, which does not rely on any prior knowledge, was utilized to prioritize DE genes between two conditions. The proposed ranking system gives priority to the DE genes which interact with each other and comprise a network. The priority of the nodes in a network increases by the number of roles the respective genes take either as dominators or when taking part in a dominating pathway or as hub-degree nodes. DE genes with the same priority are sorted in descending order based on the absolute value of their fold changes. Highest priority is given to a gene with highest fold change with the same combination of topological scores.

The sets of DE genes identified by different bioinformatics tools do not agree with each other in terms of size and content. A comprehensive comparison between eleven different

tools took RNA-Seq read counts as input and showed that there is no general consensus among them (Soneson and Delorenzi, 2013). In this dissertation, we showed that there is a noticeable overlap between topological features of these methods. To explain this finding, we selected four differential gene expression analysis methods which take read counts as input and return p-values as output. We used our web server TFmiR to construct the disease-associated network.

TFmiR is a freely available web server which takes at least one set of DE genes or miRNAs as input. Then it constructs a network with all the interactions existing in selected regulatory databases for four types of interactions (TF $\rightarrow$ gene, TF $\rightarrow$ miRNA, miRNA $\rightarrow$ gene and miRNA $\rightarrow$ miRNA). If a disease is selected, then the network is pruned to a disease-specific network by selecting those edges for which at least one of their endpoints was reported in gene or miRNA disease databases.

TFmiR2 web server reinforces the functionality of TFmiR by pruning the complete network to process-specific and tissue-specific networks and also a combination of them such as disease-process, disease-tissue and process-tissue networks. Applying MCDS on the LSCC of the (disease, tissue, process) networks identified a set of genes and miRNAs of size five, where four of them were common in the dominating pathway of all of the networks.

FFL motifs of size three and four can be detected by TFmiR and TFmiR2. In TFmiR2, we offered two randomization strategies, an edge-type conserving method and a non-conserving method. In a separate work, we showed that the conserving method detected much fewer significant motifs compared to the non-conserving method in the breast cancer network, while carrying almost the same number of biological process GO terms. Moreover, we showed that a considerable fraction of co-regulation FFL motif nodes (genes and miRNAs) which are detected by the conserving method overlapped with the dominating pathway of the network.

To summarize, the methods that we developed in this thesis have a variety of applications. MDS is capable of identifying key regulators and master regulators in any network. Depends on the underlying network, the role of dominators are specified, e.g, if the network is a differential network, the dominators are responsible for the transition. The key dominating pathway in the network is detected by MCDS. Like MDS, the role of dominators and connectors in the MCDS are defined by the underlying networks. TFmiR2 constructs disease-, tissue-, process-specific networks or a combination of these networks by receiving at least one set of dysregulated genes or miRNAs. The set of hotspot nodes and network motifs can be detected by the web server tools.

Bibliography

Alcaraz, N., Kücük, H., Weile, J., Wipat, A., and Baumbach, J. (2011). Keypathwayminer: Detecting case-specific biological pathways using expression data. Internet Mathematics, 7(4):299–313.

Alon, U. (2007). Network motifs: theory and experimental approaches. Nat Rev Genet, 8(6):450–461.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology, 11(10):R106+.

Anders, S. and Huber, W. (2016). Differential expression of RNA-Seq data at the gene level - the DESeq package.

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., and Robinson, M. D. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc, 8(9):1765–1786.

Artyomov, M. N., Meissner, A., and Chakraborty, A. K. (2010). A Model for Genetic and Epigenetic Regulatory Networks Identifies Rare Pathways for Transcription Factor Induced Pluripotency. PLoS Comput Biol, 6(5):e1000785+.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. Nat Genet, 25(1):25–29.

Barabasi, A.-L. and Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. Nat Rev Genet, 5(2):101–113.

Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C., and Apweiler, R. (2009). The GOA database in 2009–an integrated Gene Ontology Annotation resource. Nucleic Acids Res, 37(Database issue):D396–403.

Bartel, D. (2009). Microrna: target recognition and regulatory functions. Cell, 136(2):215–233.

Bauer-Mehren, A., Rautschka, M., Sanz, F., and Furlong, L. I. (2010). DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. Bioinformatics (Oxford, England), 26(22):2924–2926.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Roy Statist Soc. Series B (Methodological), 57(1):289–300.

Bertucci, F., Salas, S., Eysteries, S., Nasser, V., Finetti, P., Ginestier, C., Charafe-Jauffret, E., Loriod, B., Bachelart, L., Montfort, J., Victorero, G., Viret, F., Ollendorff, V., Fert, V., Giovaninni, M., Delpero, J. R., Nguyen, C., Viens, P., Monges, G., Birnbaum, D., and Houlgatte, R. (2004). Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. Oncogene, 23(7):1377–1391.

Bossi, A. and Lehner, B. (2009). Tissue specificity and the human protein interaction network. Mol Syst Biol, 5(1):260.

Bouafia, A., Corre, S., Gilot, D., Mouchet, N., Prince, S., and Galibert, M.-D. (2014). p53 requires the stress sensor usf1 to direct appropriate cell fate decision. PLOS Genetics, 10(5):1–13.

Calderone, A., Castagnoli, L., and Cesareni, G. (2013). mentha: a resource for browsing integrated protein-interaction networks. Nat Meth, 10:690–691.

Cameron, A. and Trivedi, P. (1998). Regression Analysis of Count Data. Cambridge University Press.

Cameron, A. and Trivedi, P. (2007). Essentials of Count Data Regression. Blackwell Publishing Ltd.

Chang, D. T., Huang, C.-Y., Wu, C.-Y., and Wu, W.-S. (2011). YPA: An integrated repository of promoter features in Saccharomyces cerevisiae. Nucleic Acids Res, 39:D647–D652.

Chang, J. C., Wooten, E. C., Tsimelzon, A., Hilsenbeck, S. G., Gutierrez, M. C., Elledge, R., Mohsin, S., Osborne, C. K., Chamness, G. C., Allred, D. C., and O'Connell, P. (2003). Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. Lancet, 362(9381):362–369.

Chen, H. and Boutros, P. C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics, 12(1):35+.

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res, 37(suppl 2):W305–W311.

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

Chen, T., Li, M., He, Q., Zou, L., Li, Y., Chang, C., Zhao, D., and Zhu, Y. (2017). Liverwiki: a wiki-based database for human liver. BMC Bioinformatics, 18(1):452.

Ching, T., Huang, S., and Garmire, L. X. (2014). Power analysis and sample size estimation for RNA-Seq differential expression. RNA, 20(11):1684–1696.

Chlebík, M. and Chlebíková, J. (2008). Approximation hardness of dominating set problems in bounded degree graphs. Information and Computation, 206(11):1264–1275.

Chou, C.-H., Shrestha, S., Yang, C.-D., Chang, N.-W., Lin, Y.-L., Liao, K.-W., Huang, W.-C., Sun, T.-H., Tu, S.-J., Lee, W.-H., Chiew, M.-Y., Tai, C.-S., Wei, T.-Y., Tsai, T.-R., Huang, H.-T., Wang, C.-Y., Wu, H.-Y., Ho, S.-Y., Chen, P.-R., Chuang, C.-H., Hsieh, P.-J., Wu, Y.-S., Chen, W.-L., Li, M.-J., Wu, Y.-C., Huang, X.-Y., Ng, F. L., Buddhakosai, W., Huang, P.-C., Lan, K.-C., Huang, C.-Y., Weng, S.-L., Cheng, Y.-N., Liang, C., Hsu, W.-L., and Huang, H.-D. (2018). mirtarbase update 2018: a resource for experimentally validated microrna-target interactions. Nucleic Acids Research, 46(D1):D296–D302.

Cohnheim, J. (1867). Ueber entzundung und eiterung. Path Anat Physiol Klin Med, 40:1–79.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W. W., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. Genome biology, 17(1):1–19.

Cook, S. A., Salmon, P., Hayes, G., Byrne, A., and Fisher, P. L. (2018). Predictors of emotional distress a year or more after diagnosis of cancer: A systematic review of the literature. Psycho-Oncology.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2003). Introduction to Algorithms. McGraw-Hill Science / Engineering / Math, 2nd edition.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, Complex Systems:1695.

de Leeuw, J. R. J., de Graeff, A., Ros, W. J. G., Blijham, G. H., Hordijk, G.-J., and Winnubst, J. A. M. (2000). Prediction of depressive symptomatology after treatment of head and neck cancer: The influence of pre-treatment physical and depressive symptoms, coping, and social support. Head & Neck, 22(8):799–807.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Roder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L. H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigo, R., and Gingeras, T. R. (2012). Landscape of transcription in human cells. Nature, 489(7414):101–108.

Ernst, M., Du, Y., Warsow, G., Hamed, M., Endlich, N., Endlich, K., Murua Escobar, H., Sklarz, L.-M., Sender, S., Junghanß, C., Möller, S., Fuellen, G., and Struckmann, S. (2017). FocusHeuristics-expression-data-driven network optimization and disease gene prediction. Scientific Reports, 7:42638.

Esquela-Kerscher, A. and Slack, F. J. (2006). Oncomirs - microRNAs with a role in cancer. Nat Rev Cancer, 6(4):259–269.

Evan, G. I. and Vousden, K. H. (2001). Proliferation, cell cycle and apoptosis in cancer. Nature, 411(6835):342–348.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. Social Networks, 1:215–239.

Friard, O., Re, A., Taverna, D., De Bortoli, M., and Corá, D. (2010). CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse. BMC bioinformatics, 11(1):435+.

Garey, M. R. and Johnson, D. S. (1979). Computers and intractability: A guide to the theory of NP-Completeness. W.H. Freeman, first edition.

Georgakilas, G., Vlachos, I. S., Zagganas, K., Vergoulis, T., Paraskevopoulou, M. D., Kanellos, I., Tsanakas, P., Dellis, D., Fevgas, A., Dalamagas, T., and Hatzigeorgiou, A. G. (2016). DIANA-miRGen v3.0: accurate characterization of microRNA promoters and their regulators. Nucleic Acids Res, 44(D1).

Goh, K.-I. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. L. (2007). The human disease network. Proceedings of the National Academy of Sciences of the United States of America, 104(21):8685–8690.

Golipour, A., David, L., Liu, Y., Jayakumaran, G., Hirsch, C. L., Trcka, D., and Wrana, J. L. (2012). A late transition in somatic cell reprogramming requires regulators distinct from the pluripotency network. Cell Stem Cell, 11(6):769 – 782.

Griffith, O. L., Montgomery, S. B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., Mahony, S., Sleumer, M. C., Bilenky, M., Haeussler, M., Griffith, M., Gallo, S. M., Giardine, B., Hooghe, B., Van Loo, P., Blanco, E., Ticoll, A., Lithwick, S., Portales-Casamar, E., Donaldson, I. J., Robertson, G., Wadelius, C., De Bleser, P., Vlieghe, D., Halfon, M. S., Wasserman, W., Hardison, R., Bergman, C. M., Jones, S. J., and Open Regulatory Annotation Consortium (2008). ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res, 36(Database issue):D107–D113.

Gross, J. L. and Yellen, J. (2005). Graph Theory and Its Applications, Second Edition (Discrete Mathematics and Its Applications). Chapman & Hall/CRC.

Guha, S. and Khuller, S. (1998). Approximation algorithms for connected dominating sets. Algorithmica, 20:374–387.

Guo, Z., Naik, A., O'Malley, M., Popovic, P., Demarco, R., Hu, Y., Yin, X., Yang, S., Zeh, H., and Moss, B. (2005). The enhanced tumor selectivity of an oncolytic vaccinia lacking the host range and antiapoptosis genes SPI-1 and SPI-2. . Cancer Res, 65:9991–9998.

Guzzi, P. H., Di Martino, M. T., Tagliaferri, P., Tassone, P., and Cannataro, M. (2015). Analysis of mirna, mrna, and tf interactions through network-based methods. EURASIP Journal on Bioinformatics and Systems Biology, 2015(1):4.

Halbritter, F., Vaidya, H. J., and Tomlinson, S. R. (2011). GeneProf: Analysis of high-throughput sequencing experiments. Nature Methods, 9(1):7–8.

Hamed, M., Ismael, S., Paulsen, M., and Helms, V. (2012). Cellular functions of genetically imprinted genes in human and mouse as annotated in the gene ontology. PLoS ONE, 7:e50285.

Hamed, M., Spaniol, C., Nazarieh, M., and Helms, V. (2015a). TFmiR: A web server for constructing and analyzing disease-specific transcription factor and miRNA co-regulatory networks. Nucleic Acids Res, 43(W1):gkv418.

Hamed, M., Spaniol, C., Zapp, A., and Helms, V. (2015b). Integrative network based approach identifies key genetic elements in breast invasive carcinoma. BMC Genomics, 16(Suppl. 5):S2.

Hamed, M., Trumm, J., Spaniol, C., Sethi, R., Irhimeh, M., Fuellen, G., Paulsen, M., and Helms, V. (2017). Linking hematopoietic differentiation to co-expressed sets of pluripotency-associated and imprinted genes and to regulatory microrna-transcription factor motifs. PLOS ONE, 12.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. Cell, 144(5):646–674.

He, L., He, X., Lim, L., De Stanchina, E., Xuan, Z., Liang, Y., Xue, W., Zender, L., Magnus, J., and Ridzon, D. (2007). A microRNA component of the p53 tumour suppressor network. nature, 447:1130–1134.

Hix, L. M., Karavitis, J., Khan, M. W., Shi, Y. H., Khazaie, K., and Zhang, M. (2013). Tumor stat1 transcription factor activity enhances breast tumor growth and immune suppression mediated by myeloid-derived suppressor cells. Journal of Biological Chemistry, 288(17):11676–11688.

Hobert, O. (2008). Gene regulation by transcription factors and microrna. Science, 319(5871):1785–1786.

Hopcroft, J. and Tarjan, R. (1973). Algorithm 447: Efficient algorithms for graph manipulation. Commun. ACM, 16(6):372–378.

Hsu, S.-D., Lin, F.-M., Wu, W.-Y., Liang, C., Huang, W.-C., Chan, W.-L., Tsai, W.-T.and Chen, G.-Z., Lee, C.-J., Chiu, C.-M., Chien, C.-H., Wu, M.-C., Huang, C.-Y., Tsou, A.-P., and Huang, H.-D. (2010). miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res, 39(Database issue):D163–D169.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat protocols, 4(1):44–57.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics (Oxford, England), 18 Suppl 1(suppl 1):S233–S240.

Jiang, C., Xuan, Z., Zhao, F., and Zhang, M. Q. (2007). TRED: a transcriptional regulatory element database, new entries and other development. Nucleic Acids Res, 35(suppl 1):D137–D140.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res, 37(suppl 1):D98–D104.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. (2004). Human MicroRNA Targets. PLoS Biol, 2(11):e363+.

Kacprowski, T., Doncheva, N. T., and Albrecht, M. (2013). NetworkPrioritizer: A versatile tool for network - based prioritization of candidate disease genes or other molecules. Bioinformatics, 29(11):1471–1473.

Kashtan, N., Itzkovitz, S., Milo, R., and Alon, U. (2004). Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics, 20(11):1746–1758.

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

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S. H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. Cell, 132:1049–1061.

Koboldt, D. C., Fulton, R. S., McLellan, M. D., and et al. (2012). Comprehensive molecular portraits of human breast tumours. Nature, 490(7418):61–70.

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M., and Rajewsky, N. (2005). Combinatorial microRNA target predictions. Nat genetics, 37(5):495–500.

Kusenda, B., Mraz, M., Mayer, J., and Pospisilova, S. (2006). MicroRNA biogenesis, functionality and cancer relevance. Biomedical papers of the Medical Faculty of the University Palacký, Olomouc, Czechoslovakia, 150(2):205–215.

Laczny, C., Leidinger, P., Haas, J., Ludwig, N., Backes, C., Gerasch, A., Kaufmann, M., Vogel, B., Katus, H. A., Meder, B., Stahler, C., Meese, E., Lenhof, H. P., and Keller, A. (2012). miRTrail - a comprehensive webserver for analyzing gene and miRNA patterns to enhance the understanding of regulatory mechanisms in diseases. BMC bioinformatics, 13(1):36+.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology, 15(2):1–17.

Le, T. D., Liu, L., Tsykin, A., Goodall, G. J., Liu, B., Sun, B.-Y., and Li, J. (2013). Inferring microRNA-mRNA causal regulatory relationships from expression data. Bioinformatics, 29(6):765–771.

Lee, T. and Young, R. A. (2013). Transcriptional Regulation and Its Misregulation in Disease. Cell, 152(6):1237–1251.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J., Volkert, T. L., Fraenkel, E., Gifford, D., and Young, R. A. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science, 298:799–804.

Lesurf, R., Cotto, K. C., Wang, G., Griffith, M., Kasaian, K., Jones, S. J. M., Montgomery, S. B., Griffith, O. L., and Consortium, T. O. R. A. (2016). ORegAnno 3.0: a community-driven resource for curated regulatory annotation. Nucleic Acids Research, 44(D1):D126–D132.

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., and Yang, J.-H. (2014a). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Research, 42(D1):D92–D97.

Li, X., Cassidy, J. J., Reinke, C. A., Fischboeck, S., and Carthew, R. W. (2009). A MicroRNA Imparts Robustness against Environmental Fluctuation during Development. Cell, 137(2):273–282.

Li, X.-Q., Lu, J.-T., Tan, C.-C., Wang, Q.-S., and Feng, Y.-M. (2016). Runx2 promotes breast cancer bone metastasis by increasing integrin alpha 5 - mediated colonization. Cancer Letters, 380(1):78 – 86.

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., and Cui, Q. (2014b). Hmdd v2.0: a database for experimentally supported human microrna and disease associations. Nucleic Acids Research, 42(D1):D1070–D1074.

Liang, C., Li, Y., Luo, J., and Zhang, Z. (2015). A novel motif-discovery algorithm to identify co-regulatory motifs in large transcription factor and microrna co-regulatory networks in human. Bioinformatics, 31(14):2348–2355.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics (Oxford, England), 27(12):1739–1740.

Liu, Y.-Y., Slotine, J.-J., and Barabasi, A.-L. (2011). Controllability of complex networks. Nature, 473(7346):167–173.

Liu, Z., Wu, C., Miao, H., and Wu, H. (2015). Regnetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database, 2015:bav095.

Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q., and Bader, G. D. (2010). Cytoscape Web: an interactive web-based network browser. Bioinformatics, 26(18):2347–2348.

Lu, M., Shi, B., Wang, J., Cao, Q., and Cui, Q. (2010). Tam: A method for enrichment and depletion analysis of a microrna category in a list of micrornas. BMC Bioinformatics, 11(1):419.

Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., and Cui, Q. (2008). An Analysis of Human MicroRNA and Disease Associations. PLoS ONE, 3(10):e3420+.

Ma, H.-W. W., Buer, J., and Zeng, A.-P. P. (2004). Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. BMC bioinformatics, 5(1):199.

Makhorin, A. (2008). GLPK (GNU linear programming kit).

Matkovich, S., Hu, Y., and Dorn, G. (2013). Regulation of cardiac microRNAs by cardiac microRNAs. Circulation Res, 113:62–71.

Matys, V., Fricke, E., Geffers, R., Gössling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D. U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res, 31(1):374–378.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res, 34(Database issue):D108–D110.

McGuire, A., Brown, J., Malone, C., McLaughlin, R., and Kerin, M. (2015). Effects of Age on the Detection and Management of Breast Cancer. Cancers, 7(2):908–929.

Megraw, M., Mukherjee, S., and Ohler, U. (2013). Sustained-input switches for transcription factors and microRNAs are central building blocks of eukaryotic gene circuits. Genome Biology, 14(8):R85+.

Mele, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segre, A. V., Djebali, S., Niarchou, A., Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E. T., Ardlie, K. G., and Guigo, R. (2015). Human genomics. The human transcriptome across tissues and individuals. Science, 348(6235):660–665.

Milenković, T., Memišević, V., Bonato, A., and Pržulj, N. (2011). Dominating biological networks. PLoS ONE, 6:e23016+.

Milo, R., Kashtan, N., Itzkovitz, S., Newman, M. E. J., and Alon, U. (2004). On the uniform generation of random graphs with prescribed degree sequences.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network Motifs: Simple Building Blocks of Complex Networks. Science, 298(5594):824–827.

Miranda, K., Huynh, T., Tay, Y., Ang, Y. S., Tam, W. L., Thomson, A. M., Lim, B., and Rigoutsos, I. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell, 126(6):1203–1217.

Moreau, Y. and Tranchevent, L.-C. C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature reviews. Genetics, 13(8):523–536.

Navlakha, S. and Bar-Joseph, Z. (2014). Distributed information processing in biological and computational systems. Commun. ACM, 58(1):94–102.

Nazari, F., Pearson, A. T., Nör, J. E., and Jackson, T. L. (2018). A mathematical model for il-6-mediated, stem cell driven tumor growth and targeted treatment. PLoS Comput Biol, 14(1):1–32.

Nazarieh, M., Wiese, A., Will, T., Hamed, M., and Helms, V. (2016). Identification of key player genes in gene regulatory networks. BMC Systems Biology, 10(1):88.

Nazarieh, M., Will, T., Hamed, M., Spaniol, C., and Helms, V. (2017). Constructing and analyzing disease-specific or developmental stage-specific transcription factor and mirna co-regulatory networks. F1000Research 2016, 5:2225 (poster).

Nazarov, P. V., Reinsbach, S. E., Muller, A., Nicot, N., Philippidou, D., Vallar, L., and Kreis, S. (2013). Interplay of microRNAs, transcription factors and target genes: linking dynamic expression changes to function. Nucleic Acids Res, 41(5):2817–2831.

Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. Cell, 150(6):1274–1286.

Ohno, S. (1979). Major sex-determining genes. Springer-Verlag; Berlin, Germany, 39(suppl 1).

Pervouchine, D. D., Djebali, S., Breschi, A., Davis, C. A., Barja, P. P., Dobin, A., Tanzer, A., Lagarde, J., Zaleski, C., See, L. H., Fastuca, M., Drenkow, J., Wang, H., Bussotti, G., Pei, B., Balasubramanian, S., Monlong, J., Harmanci, A., Gerstein, M., Beer, M. A., Notredame, C., Guigo, R., and Gingeras, T. R. (2015). Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat Commun, 6:5903.

Piñero, J., Bravo, A., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L. I. (2017). Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research, 45(D1):D833–D839.

Poos, K., Smida, J., Nathrath, M., Maugg, D., Baumhoer, D., and Korsching, E. (2013). How MicroRNA and Transcription Factor Co-regulatory Networks Affect Osteosarcoma Cell Proliferation. PLoS Comput. Biol., 9(8):e1003210+.

Qiu, C., Wang, J., Yao, P., Wang, E., and Cui, Q. (2010). microRNA evolution in a human transcription factor and microRNA regulatory network. BMC systems biology, 4(1):90+.

Rahman, M., Deleyrolle, L., Vedam-Mai, V., Azari, H., Abd-El-Barr, M., and Reynolds, B. A. (2011). The cancer stem cell hypothesis: Failures and pitfalls. Neurosurgery, 68(2):531–545.

Rai, M., Verma, S., and Tapaswi, S. (2009). A power aware minimum connected dominating set for wireless sensor networks. J. Networks, 4:511–519.

Reithmeier, A., Panizza, E., Krumpel, M., Orre, L. M., Branca, R. M., Lehtiö, J., Ek-Rylander, B., and Andersson, G. (2017). Tartrate-resistant acid phosphatase (TRAP/ACP5) promotes metastasis-related properties via TGFβ2/TβR and CD44 in MDA-MB-231 breast cancer cells. BMC Cancer, 17(1):1–19.

Reya, T., Morrison, S. J., Clarke, M. F., and Weissman, I. L. (2001). Stem cells, cancer, and cancer stem cells. Nature, 414:105.

Rimmelé, P., Komatsu, J., Hupé, P., Roulin, C., Barillot, E., Dutreix, M., Conseiller, E., Bensimon, A., Moreau-Gachelin, F., and Guillouf, C. (2010). Spi-1/PU. 1 oncogene accelerates DNA replication fork elongation and promotes genetic instability in the absence of DNA breakage. Cancer Res, 70:6757–6766.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England), 26(1):139–140.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome biology, 11(3):1–9.

Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. Bioinformatics (Oxford, England), 23(21):2881–2887.

Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics (Oxford, England), 9(2):321–332.

Robinton, D. A. and Daley, G. Q. (2012). The promise of induced pluripotent stem cells in research and therapy. Nature, 481(7381):295–305.

Sadegh, S., Nazarieh, M., Spaniol, C., and Helms, V. (2017). Randomization strategies affect motif significance analysis in TF-miRNA-gene regulatory networks. Journal of Integrative Bioinformatics, 14:378–396.

Sakurai, T., Kondoh, N., Arai, M., Hamada, J., Yamada, T., Kihara-Negishi, F., Izawa, T., Ohno, H., Yamamoto, M., and Oikawa, T. (1775-1784). Functional roles of Fli-1, a member of the Ets family of transcription factors, in human breast malignancy. Cancer science, 98:1360–1371.

Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muñiz Rascado, L., García-Sotelo, J. S., Weiss, V., Solano-Lira, H., Martínez-Flores, I., Medina-Rivera, A., Salgado-Osorio, G., Alquicira-Hernández, S., Alquicira-Hernández, K., López-Fuentes, A., Porrón-Sotelo, L., Huerta, A. M., Bonavides-Martínez, C., Balderas-Martínez, Y. I., Pannier, L., Olvera, M., Labastida, A., Jiménez-Jacinto, V., Vega-Alvarado, L., Del Moral-Chávez, V., Hernández-Alvarez, A., Morett, E., and Collado-Vides, J. (2013). RegulonDB v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res, 41:D203–D213.

Sant, P. M. (2004). Rooted tree. in Dictionary of Algorithms and Data Structures [online], Vreda Pieterse and Paul E. Black eds.

Schwartz-Roberts, J. L., Cook, K. L., Chen, C., Shajahan-Haq, A. N., Axelrod, M., Wärri, A., Riggins, R. B., Jin, L., Haddad, B. R., Kallakury, B. V., Baumann, W. T., and Clarke, R. (2015). Interferon regulatory factor-1 signaling regulates the switch between autophagy and apoptosis to determine breast cancer cell fate. Cancer Res, 75(6):1046–1055.

Sedgewick, R. and Wayne, K. (2015). Algorithms, (Deluxe): Book and 24-Part lecture series. Addison-Wesley Professional.

Sengupta, D. and Bandyopadhyay, S. (2011). Participation of microRNAs in human interactome: extraction of microRNA-microRNA regulations. Molecular bioSystems.

Sengupta, D. and Bandyopadhyay, S. (2013). Topological patterns in microRNA-gene regulatory network: studies in colorectal and breast cancer. Mol. BioSyst, 9:1360–1371.

Sethupathy, P., Corda, B., and Hatzigeorgiou, A. G. (2006). TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA (New York, N.Y.), 12(2):192–197.

Seton-Rogers, S. (2014). Gender differences. Nat Rev Cancer, 14(9):578–579.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res, 13(11):2498–2504.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genetics, 31(1):64–68.

Simonetti, L., Da Cunha, A. S., and Lucena, A. (2011). The minimum connected dominating set problem: Formulation, valid inequalities and a branch-and-cut algorithm. Network Optimization, pages 162–169.

Singh, A. H., Wolf, D. M., Wang, P., and Arkin, A. P. (2008). Modularity of stress response evolution. Proceedings of the National Academy of Sciences, 105(21):7500–7505.

Smith, C. L., Blake, J. A., Kadin, J. A., Richardson, J. E., Bult, C. J., and the Mouse Genome Database Group (2018). Mouse genome database (mgd)-2018: knowledgebase for the laboratory mouse. Nucleic Acids Research, 46(D1):D836–D842.

Som, A., Harder, C., Greber, B., Siatkowski, M., Paudel, Y., Warsow, G., Cap, C., Schöler, H., and Fuellen, G. (2010). The PluriNetWork: An electronic representation of the network underlying pluripotency in mouse, and its applications. PloS one, 5(12):e15165.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics, 14(1):91+.

Spellman, P., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle regulated genes of the Yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell, 9:3273–3297.

Stock, A., Robinson, V. L., and Goudreau, P. N. (2000). Two-compont signal transduction. Annual Review of Biochemistry, 69(1):183–215.

Sun, J., Gong, X., Purow, B., and Zhao, Z. (2012). Uncovering MicroRNA and Transcription Factor Mediated Regulatory Networks in Glioblastoma. PLoS Comput Biol, 8(7):e1002488+.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., and von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic acids research, 43(Database issue):D447–D452.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J., and Von Mering, C. (2017). The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Research, 45(D1):D362–D368.

Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell, 126(4):663 – 676.

Tan, B. T., Park, C. Y., Ailles, L. E., and Weissman, I. L. (2006). The cancer stem cell hypothesis: A work in progress. Laboratory Investigation, 86(12):1203–1207.

Tarjan, R. (1972). Depth-first search and linear graph algorithms. SIAM Journal on Computing, 1(2):146–160.

The-Gene-Ontology-Consortium (2017). Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res, 45(D1):gkw1108+.

The Sage Developers (2015). Sage Mathematics Software (Version 6.7).

Thomas, P. D. (2010). Giga: a simple, efficient algorithm for gene tree inference in the genomic age. BMC Bioinformatics, 11(1):312.

Tranchevent, L.-C., Ardeshirdavani, A., ElShal, S., Alcaide, D., Aerts, J., Auboeuf, D., and Moreau, Y. (2016). Candidate gene prioritization with Endeavour. Nucleic Acids Res, 44(W1):W117–W121.

Urruticoechea, A., Alemany, R., Balart, J., Villanueva, A., Vinals, F., and Capella, G. (2010). Recent advances in cancer therapy: An overview. Current Pharmaceutical Design, pages 3–10.

Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. Journal of computational biology : a journal of computational molecular cell biology, 18(3):507–522.

Vergoulis, T., Vlachos, I. S., Alexiou, P., Georgakilas, G., Maragkakis, M., Reczko, M., Gerangelos, S., Koziris, N., Dalamagas, T., and Hatzigeorgiou, A. G. (2012). TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic acids research, 40(Database issue):D222–D229.

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. Nucleic Acid Res., 31(1):258–261.

Wang, J., Lu, M., Qiu, C., and Cui, Q. (2010). TransmiR: a transcription factor-microRNA regulation database. Nucleic Acids Res, 38(Database issue):D119–D122.

Weiwei, T., Jun-Feng, S., Qian, Z., Bin, X., Yu-Jie, S., and Chao-Jun, L. (2013). Egr-1 enhances drug resistance of breast cancer by modulating mdr1 expression in a ggpps-independent manner. Biomedicine and Pharmacotherapy, 67(3):197 – 202.

Wernicke, S. (2005). A Faster Algorithm for Detecting Network Motifs. Algorithms in Bioinformatics, 3692:165–177.

Wernicke, S. and Rasche, F. (2006). FANMOD: a tool for fast network motif detection. Bioinformatics, 22(9):1152–1153.

Wightman, P., Fabregas, A., and Labrador, M. (2011). A mathematical solution to the mcds problem for topology construction in wireless sensor networks. Latin America Transactions, IEEE (Revista IEEE America Latina), 9(4):534–541.

Will, T. and Helms, V. (2014). Identifying transcription factor complexes and their roles. Bioinformatics (Oxford, England), 30(17).

Will, T. and Helms, V. (2017). Rewiring of the inferred protein interactome during blood development studied with the tool ppicompare. BMC Systems Biology, 11(1):44.

Xi, Y., Formentini, A., Chien, M., Weir, D. B. B., Russo, J. J. J., Ju, J., Kornmann, M., and Ju, J. (2006). Prognostic Values of microRNAs in Colorectal Cancer. Biomark Insights, 2:113–121.

Xiao, F., Zuo, Z., Cai, G., Kang, S.and Gao, X., and Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res, 37(Database issue):D105–110.

Xu, H., Ang, Y.-S., Sevilla, A., Lemischka, I. R., and Ma'ayan, A. (2014). Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells. PLoS Comput Biol, 10:e1003777+.

Xu, H., Yu, H., Tu, K., Shi, Q., Wei, C., Li, Y.-Y. Y., and Li, Y.-X. X. (2013). cGRNB: a web server for building combinatorial gene regulatory networks through integrated engineering of seed-matching sequence information and gene expression datasets. BMC systems biology, 7 Suppl 2.

Yan, Z., Shah, P. K., Amin, S. B., Samur, M. K., Huang, N., Wang, X., Misra, V., Ji, H., Gabuzda, D., and Li, C. (2012). Integrative analysis of gene and miRNA expression profiles with transcription factor-miRNA feed-forward loops identifies regulators in human cancers. Nucleic Acids Res, 40(17):e135.

Yang, J.-H., Li, J.-H., Jiang, S., Zhou, H., and Qu, L.-H. (2013). ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. Nucleic Acids Res, 41(D1):D177–D187.

Yang, J.-H., Li, J.-H., Shao, P., Zhou, H., Chen, Y.-Q., and Qu, L.-H. (2011). starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. Nucleic Acids Res, 39(suppl 1):D202–D209.

Yang, L., Belaguli, N., and Berger, D. (2009). MicroRNA and colorectal cancer. World journal of surgery, 33:638–646.

Yang, W., Rosenstiel, P. C., and Schulenburg, H. (2016). Absseq: a new rna-seq analysis method based on modelling absolute expression differences. BMC Genomics, 17(1):541.

Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R. Y., Alon, U., and Margalit, H. (2004). Network motifs in integrated cellular networks of transcription - regulation and protein - protein interaction. Proceedings of the National Academy of Sciences of the United States of America, 101(16):5934–5939.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics, 26(7):976–978.

Yu, H. and Gerstein, M. (2006). Genomic analysis of the hierarchical structure of regulatory networks. Proceedings of the National Academy of Sciences, 103(40):14724–14731.

Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. PLoS Comput Biol, 3(4):1–8.

Zhang, H.-M., Kuang, S., Xiong, X., Gao, T., Liu, C., and Guo, A.-Y. (2015a). Transcription factor and microrna co-regulatory loops: important regulatory motifs in biological processes and diseases. Briefings in Bioinformatics, 16(1):45–58.

Zhang, X.-F., Ou-Yang, L., Zhu, Y., Wu, M.-Y., and Dai, D.-Q. (2015b). Determining minimum set of driver nodes in protein-protein interaction networks. BMC Bioinformatics, 16(1):146+.

Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PloS one, 9(1):e78644+.

Zhou, K.-R., Liu, S., Sun, W.-J., Zheng, L.-L., Zhou, H., Yang, J.-H., and Qu, L.-H. (2017). Chipbase v2.0: decoding transcriptional regulatory networks of non-coding rnas and protein-coding genes from chip-seq data. Nucleic Acids Research, 45(D1):D43–D50.

Appendix A

Supplementary materials

A.1 A Guide to use ILP formulations of MDS and MCDS

This guide contains instructions for users how to use the provided tools in sage software system, MDS and MCDS in a directed network. The ILP formulations were written in the Sage software system and use the glpk solver to find the solutions. The ILP formulation for the directed form of MDS finds a minimum number of dominators in the directed graph. The program is applied to the full network to find the optimal solution for the MDS problem. The program takes two arguments (input file, output file). The first argument is a tab-delimited file as an input network and the second argument is the output text file to which the results are saved in the user-specified path. The program can be executed in the terminal as follows:

```
./sage MDS_direct.sage [input_file] [output_file]
```

Suppose we have a network named black.csv, the following command outputs the MDS result for the mentioned input network. .

```
./sage MDS_direct.sage black.csv mds_result.txt
```

To find the optimal solution for the MCDS problem, we considered two types of components in the network. The LCC and the LSCC. Because no feasible solution for MCDS exists in graphs that are not connected. The ILP formulation for the directed form of MCDS finds a set of minimum connected dominators in the directed graph. The program takes three arguments (component type, input file, output file). The first argument determines the type of component which can be either LCC or LSCC. The component type needs to be specified explicitly in upper case or lower case as given in the example. The other two arguments are a tab-delimited file as input network as in section 2 and the output text file to which the results are saved in the user-specified path. The program can be executed in the terminal as follows:

```
./sage MCDS_direct.sage [Component] [input_file] [output_file]
```

Suppose we have a network named black.csv, the following command outputs the MCDS result for the mentioned input network.

```
./sage MCDS_direct.sage LCC black.csv mcds_result.txt
```

A.2 Figures

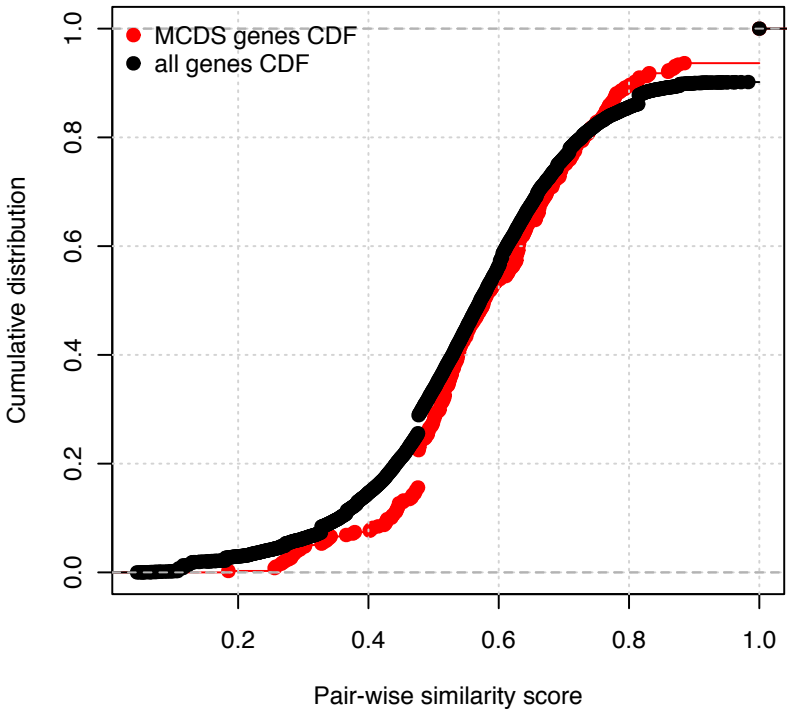

Figure S1: Cumulative distribution of the functional similarity scores between pairs of MCDS nodes of the mouse pluripotency network (in red) against the similarity between all pairs of genes in the pluripotency network (in black).

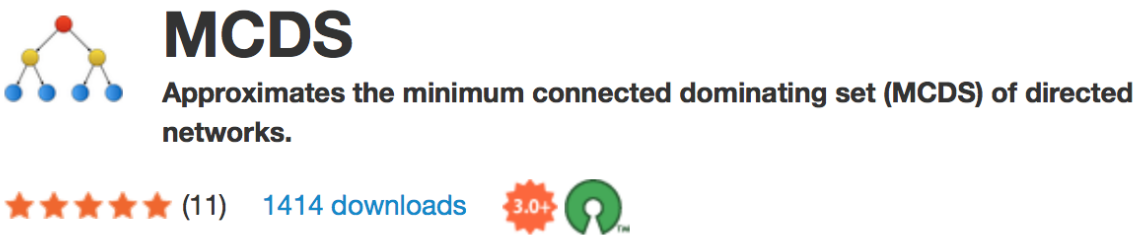

Figure S2: MCDS is a Cytoscape application that allows users to find an approximate MCDS solution in a directed graph. The user can choose to run the algorithm on the largest connected component underlying directed graph (LCCD) and undirected graph (LCC), as well as the LSCC of the network. The connectivity among the nodes in the MCDS can then be visualized via Cytoscape.

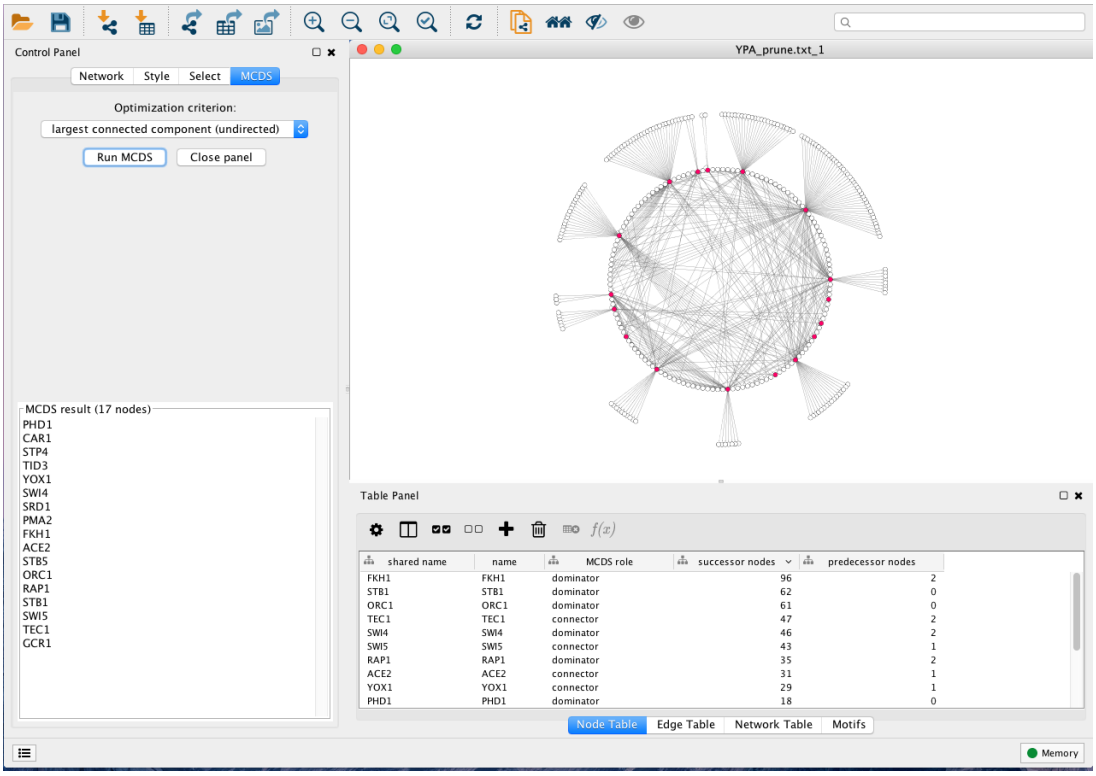

Figure S3: Example of MCDS on the LCC underlying the undirected graph of the cell cycle of *S. cerevisiae*. The associated table shows the role of the nodes in the solution with the number of successors and predecessors.

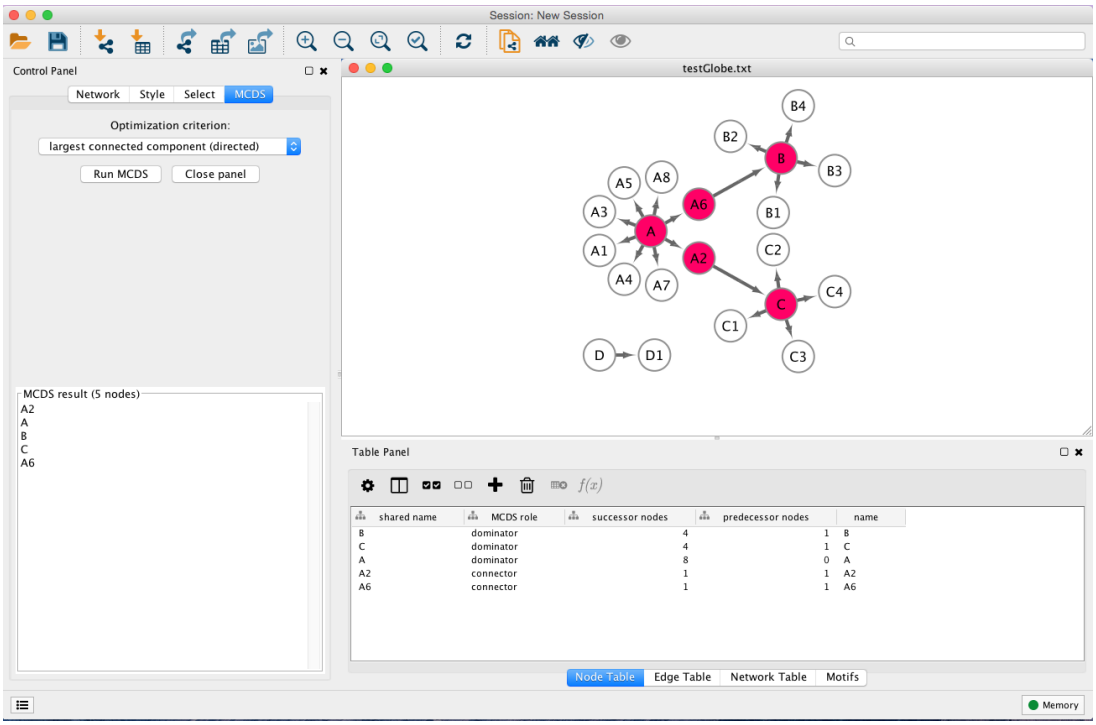

Figure S4: Example of MCDS on the LCC underlying directed graph (LCCD) of a toy network. The associated table shows the role of the nodes in the solution with the number of successors and predecessors.

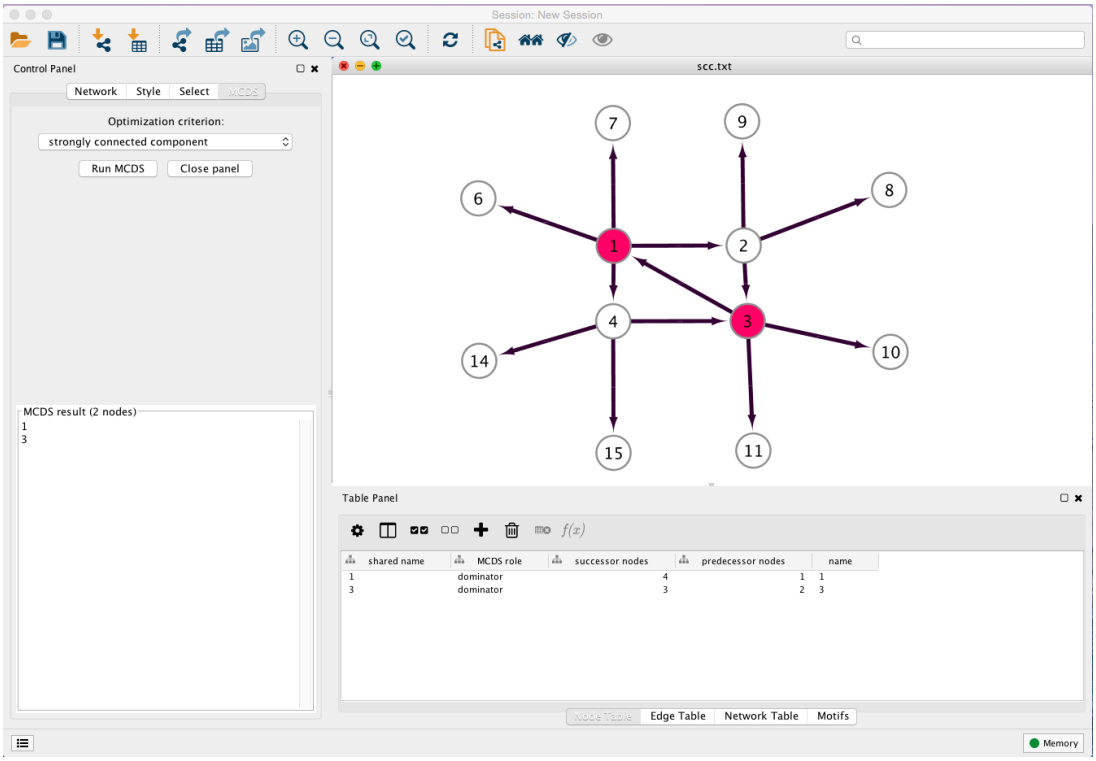

Figure S5: Example of MCDS on the LSCC of a toy network. The associated table shows the role of the nodes in the solution with the number of successors and predecessors.

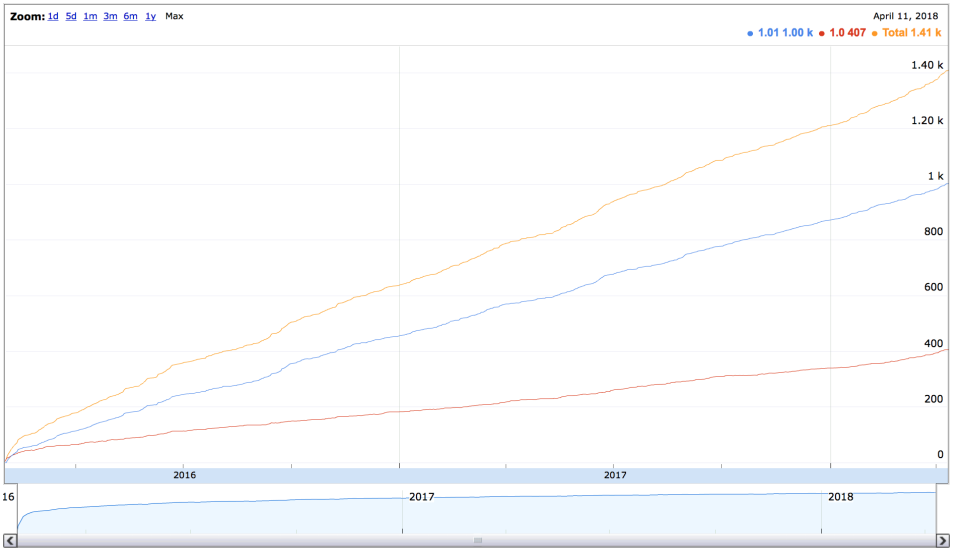

Figure S6: MCDS download statistics between February 2016 and April 2018.

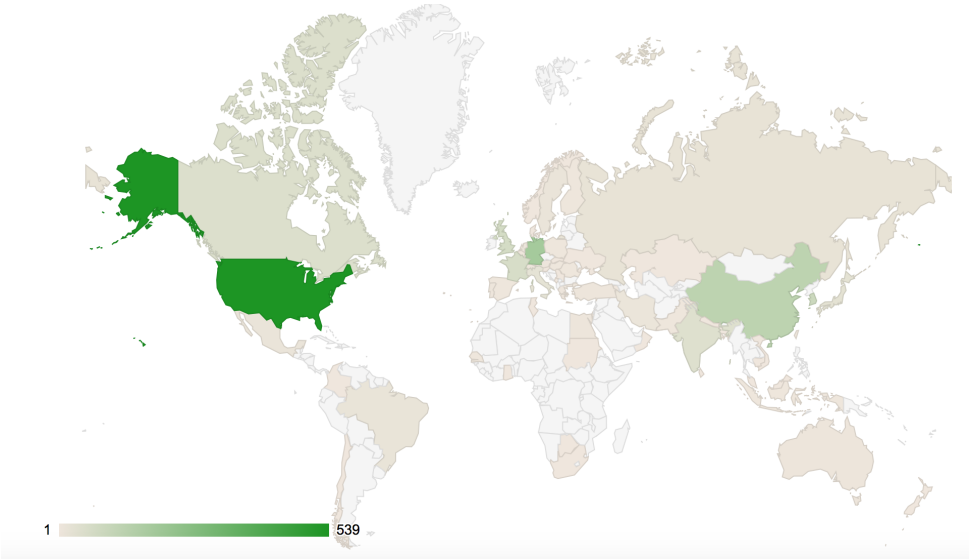

Figure S7: MCDS geographical usage distribution between February 2016 and April 2018.

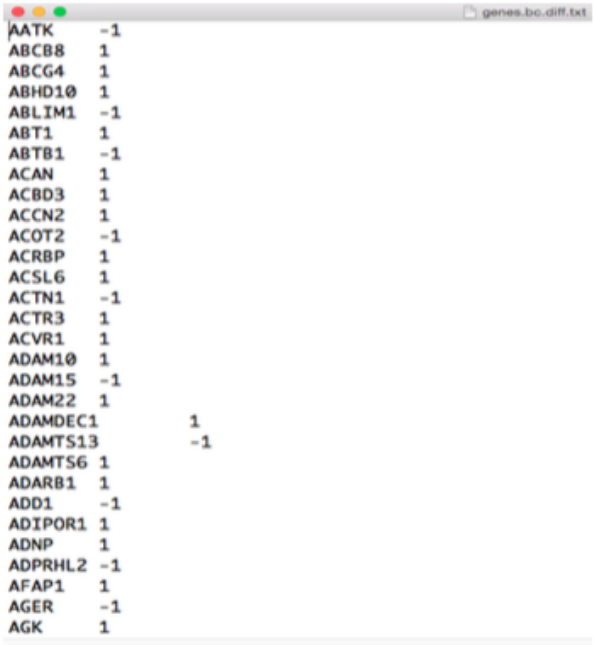

Figure S8: Sample input file of deregulated TFs/genes. (1), and (-1) refer to up- and down regulation, respectively.

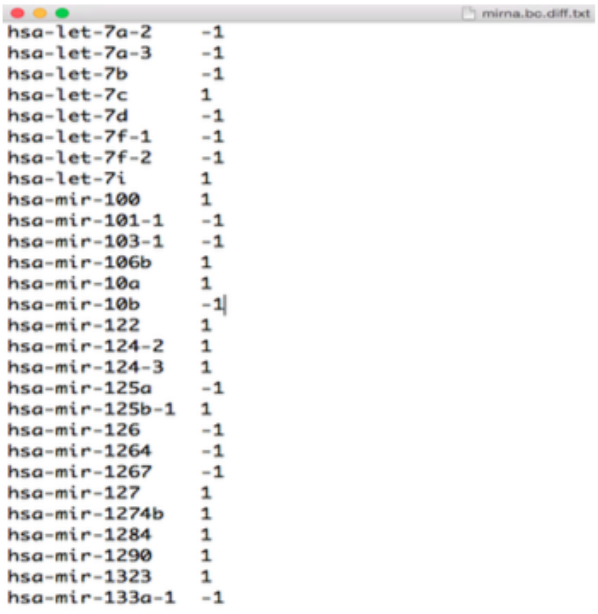

Figure S9: Sample input file of deregulated miRNAs. (1), and (-1) refer to up-and down regulation, respectively.

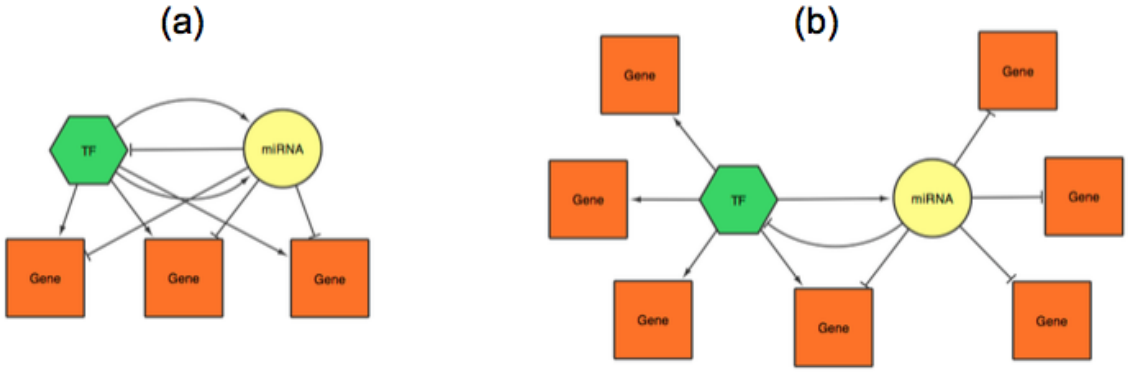

Figure S10: (a) Co-targeted genes defined as genes that are targeted by the same TF and miRNA pair. (b) Co-regulated genes defined as all genes regulated by the TF and the miRNA of this TF-miRNA pair.

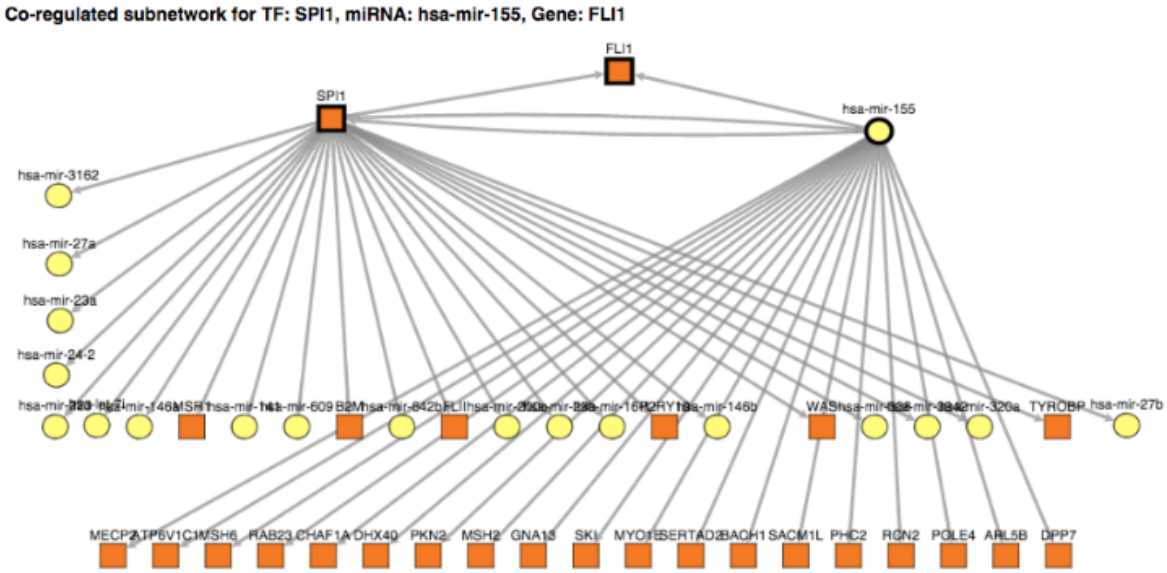

Figure S11: A composite FFL motif involves the TF SPI1, the miRNA has-mir-155, and the target gene FLI1. The co-regulated nodes are also visualized and are further tested whether they compose a cooperative functional module in breast cancerogenesis (see Figure S12).

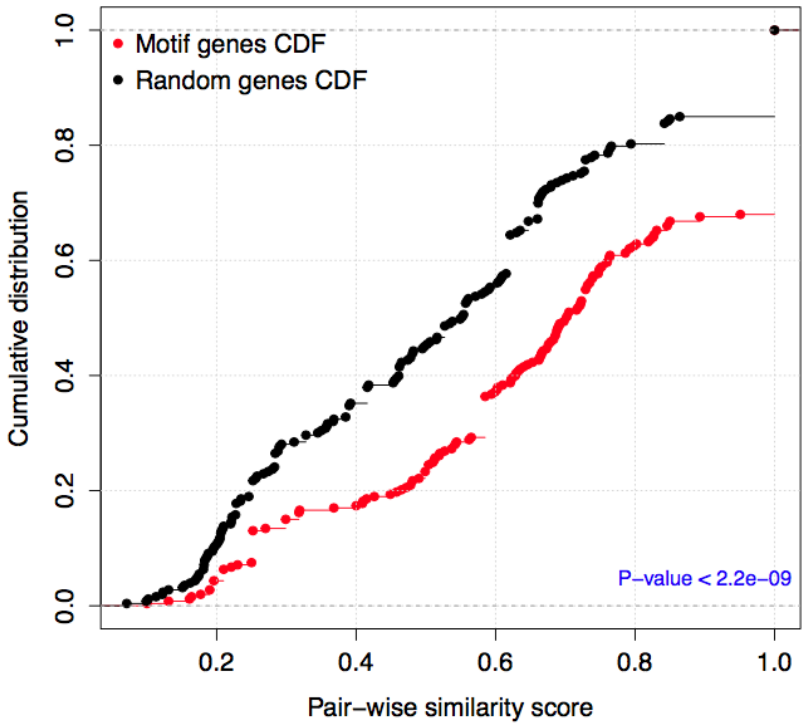

Figure S12: Cumulative distribution of GO functional semantic scores of gene pairs of co-regulated genes in the examined motif (red) versus randomly selected genes (black). The p-value was calculated using the Kolmogorov-Smirnov test.

A.3 Tables

Table S1: 34 MCDS genes of the *E.coli* LCC based taken from RegulonDB. 'D' stands for dominating node, 'C' for connecting node. Only the largest component of the network with 1198 genes was analyzed here.

Gene	Role	no. Target genes
feaR	C	2
BolA	D	2
zraR	C	3
AsnC	D	4
CreB	D	5
stpA	C	7
uxuR	C	8
rbsR	C	9
galS	C	10
malT	C	10
dpiA	C	11
metJ	D	15
ompR	C	17
leuO	D	20
nac	C	21
csgD	C	23
rob	D	26
gadX	C	28
gadE	C	28
fhlA	C	30
purR	D	31
oxyR	C	33
nagC	C	36
marA	C	38
soxS	C	40
pdhR	C	42
lrp	D	105
narL	C	121
fur	C	129
ArcA	C	173
IHF	D	219
fis	C	227
fnr	D	296
CRP	D	497

Table S2: Enriched GO terms (top) and KEGG pathways with adjusted p-values < 0.05 for the 34 genes in the MCDS for the *E.coli* GRN. p-values were adjusted for multiple testing using the BH procedure.

Enriched terms	count	*adj. p-values*
GO:0006350 ~ transcription	32	4.845E-28
GO:0045449 ~ regulation of transcription	33	4.043E-26
GO:0051171 ~ regulation of nitrogen compound metabolic process	33	3.204E-26
GO:0019219 ~ regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	33	3.204E-26
GO:0010556 ~ regulation of macromolecule biosynthetic process	33	3.108E-26
GO:0009889 ~ regulation of biosynthetic process	33	3.108E-26
GO:0031326 ~ regulation of cellular biosynthetic process	33	3.108E-26
GO:0080090 ~ regulation of primary metabolic process	33	2.843E-26
GO:0031323 ~ regulation of cellular metabolic process	33	2.610E-26
GO:0006355 ~ regulation of transcription, DNA-dependent	32	3.767E-26
GO:0051252 ~ regulation of RNA metabolic process	32	3.572E-26
GO:0010468 ~ regulation of gene expression	33	6.725E-26
GO:0060255 ~ regulation of macromolecule metabolic process	33	7.117E-26
GO:0019222 ~ regulation of metabolic process	33	7.779E-26
GO:0010467 ~ gene expression	32	2.898E-25
GO:0050794 ~ regulation of cellular process	33	1.280E-24
GO:0050789 ~ regulation of biological process	33	6.833E-24
GO:0065007 ~ biological regulation	33	1.549E-23
GO:0034645 ~ cellular macromolecule biosynthetic process	32	1.114E-22
GO:0009059 ~ macromolecule biosynthetic process	32	2.139E-22
GO:0044249 ~ cellular biosynthetic process	32	3.984E-17
GO:0009058 ~ biosynthetic process	32	3.175E-16
GO:0044260 ~ cellular macromolecule metabolic process	33	2.886E-14
GO:0006139 ~ nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	32	7.960E-14
GO:0034641 ~ cellular nitrogen compound metabolic process	33	4.637E-13
GO:0043170 ~ macromolecule metabolic process	33	1.015E-12
GO:0006807 ~ nitrogen compound metabolic process	33	1.708E-12
GO:0044238 ~ primary metabolic process	34	7.490E-10
GO:0044237 ~ cellular metabolic process	34	1.056E-9
GO:0009987 ~ cellular process	34	2.724E-7
GO:0000160 ~ two-component signal transduction system (phosphorelay)	10	1.913E-6
GO:0008152 ~ metabolic process	34	4.584E-5
ecd02020:Two-component system	6	3.558E-7
ect02020:Two-component system	6	1.853E-7
ecr02020:Two-component system	6	1.853E-7
ecg02020:Two-component system	6	1.340E-7
eck02020:Two-component system	6	1.340E-7
ecq02020:Two-component system	6	1.088E-7
eum02020:Two-component system	6	9.060E-8
ecz02020:Two-component system	6	9.060E-8
ecf02020:Two-component system	6	9.060E-8
ecx02020:Two-component system	6	8.481E-8
ecj02020:Two-component system	6	8.481E-8
eco02020:Two-component system	6	7.553E-8
ecm02020:Two-component system	6	7.127E-8
ecv02020:Two-component system	6	6.576E-8
ecw02020:Two-component system	5	5.241E-6
eci02020:Two-component system	5	5.952E-6
ece02020:Two-component system	5	5.627E-6
ecc02020:Two-component system	5	5.688E-6

Table S3: Dominators in the identified MDS for the cell-cycle specific GRN of *S. cerevisiae*.

Genes
FKH1, GCR1, ORC1, YOX1, PHD1, ACE2, STB1, SWI5, STB5, SWI4, TEC1, RAP1

Table S4: 14 TFs and 3 target genes in the identified MCDS for the cell-cycle specific GRN of *S. cerevisiae*. 'C' and 'D' stand for the roles of connector and dominating nodes.

Gene	Role	no. Target genes
PMA2	C	0
TID3	C	0
CAR1	C	0
STP4	C	1
SRD1	C	3
GCR1	D	11
STB5	D	12
PHD1	D	18
YOX1	C	29
ACE2	C	31
RAP1	D	35
SWI5	C	43
SWI4	D	46
TEC1	C	47
ORC1	D	61
STB1	D	62
FKH1	D	96

Table S5: Enriched GO terms (top) and KEGG pathways (bottom line) with adjusted p-values < 0.05 for the 17 genes in the MCDS for the cell-cycle specific GRN of *S. cerevisiae*. P-values were adjusted for multiple testing using the BH procedure.

Enriched terms	count	*adj. P-values*
GO:0006355 ~ regulation of transcription, DNA-dependent	12	2.952E-6
GO:0051252 ~ regulation of RNA metabolic process	12	1.813E-6
GO:0045449 ~ regulation of transcription	12	4.110E-5
GO:0019219 ~ regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	12	7.514E-5
GO:0051171 ~ regulation of nitrogen compound metabolic process	12	6.098E-5
GO:0034645 ~ cellular macromolecule biosynthetic process	15	1.015E-4
GO:0009059 ~ macromolecule biosynthetic process	15	9.269E-5
GO:0010468 ~ regulation of gene expression	12	2.122E-4
GO:0010556 ~ regulation of macromolecule biosynthetic process	12	2.201E-4
GO:0006357 ~ regulation of transcription from RNA polymerase II promoter	8	2.219E-4
GO:0010467 ~ gene expression	15	2.078E-4
GO:0031326 ~ regulation of cellular biosynthetic process	12	1.921E-4
GO:0009889 ~ regulation of biosynthetic process	12	1.878E-4
GO:0006350 ~ transcription	10	2.838E-4
GO:0060255 ~ regulation of macromolecule metabolic process	12	4.468E-4
GO:0031323 ~ regulation of cellular metabolic process	12	4.545E-4
GO:0080090 ~ regulation of primary metabolic process	12	4.592E-4
GO:0019222 ~ regulation of metabolic process	12	8.004E-4
GO:0048522 ~ positive regulation of cellular process	7	8.463E-4
GO:0044249 ~ cellular biosynthetic process	15	0.001
GO:0048518 ~ positive regulation of biological process	7	0.001
GO:0009058 ~ biosynthetic process	15	0.001
GO:0010628 ~ positive regulation of gene expression	6	0.001
GO:0045941 ~ positive regulation of transcription	6	0.001
GO:0051173 ~ positive regulation of nitrogen compound metabolic process	6	0.002
GO:0045935 ~ positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	6	0.002
GO:0010557 ~ positive regulation of macromolecule biosynthetic process 6 0.002		
GO:0009891 ~ positive regulation of biosynthetic process	6	0.002
GO:0031328 ~ positive regulation of cellular biosynthetic process	6	0.002
GO:0051329 ~ interphase of mitotic cell cycle	5	0.002
GO:0051325 ~ interphase	5	0.003
GO:0010604 ~ positive regulation of macromolecule metabolic process	6	0.003
GO:0009893 ~ positive regulation of metabolic process	6	0.003
GO:0031325 ~ positive regulation of cellular metabolic process	6	0.003
GO:0034641 ~ cellular nitrogen compound metabolic process	14	0.004
GO:0050794 ~ regulation of cellular process	12	0.005
GO:0006807 ~ nitrogen compound metabolic process	14	0.005
GO:0006139 ~ nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	13	0.006
GO:0044260 ~ cellular macromolecule metabolic process	16	0.006
GO:0065007 ~ biological regulation	13	0.008
GO:0043170 ~ macromolecule metabolic process	16	0.008
GO:0050789 ~ regulation of biological process	12	0.010
GO:0000278 ~ mitotic cell cycle	6	0.014
sce04111:Cell cycle	4	0.051

Table S6: 29 MCDS genes of the mouse ESC pluripotency network identified in the LSCC.

Gene	Role	out_degree
Wnt5a	C	1
Dvl1	D	1
Map2k1	D	1
Raf1	C	1
Dkk1	C	1
Smad3	D	1
Nedd4l	D	2
Akt1	C	2
Fgfr1	D	2
Grb2	D	2
Ptpn11	C	2
I tgb1	C	2
Tcl1	D	2
Mapk3	D	3
Lifr	D	3
Notch1	D	3
Hras1	D	3
Cdk2ap1	C	3
Cdx2	C	4
Gsk3b	D	5
Smad1	D	8
Sall4	C	9
Tcf3	D	12
Esrrb	D	14
Nr5a2	D	20
Stat3	D	20
Sox2	C	30
Nanog	C	46
Pou5f1	D	82

Table S7: Enriched GO terms and KEGG pathways with adjusted p-values < 0.05 for the 29 genes in the MCDS for the mouse pluripotency network. p-values were adjusted for multiple testing using the BH procedure.

Enriched terms	count	*adj. p-values*
GO:0009653 anatomical structure morphogenesis	19	5.814E-11
GO:0009790 embryonic development	16	9.890E-11
GO:0048513 organ development	21	7.347E-11
GO:0048731 system development	22	1.464E-10
GO:0032502 developmental process	24	2.106E-10
GO:0048519 negative regulation of biological process	19	1.957E-10
GO:0030154 cell differentiation	20	1.889E-10
GO:0048856 anatomical structure development	22	2.854E-10
GO:0048869 cellular developmental process	20	3.122E-10
GO:0048523 negative regulation of cellular process	18	3.431E-10
GO:0048518 positive regulation of biological process	19	8.095E-10
GO:0048522 positive regulation of cellular process	18	1.479E-9
GO:0048598 embryonic morphogenesis	12	2.233E-9
GO:0001701 in utero embryonic development	11	2.790E-9
GO:0007275 multicellular organismal development	22	2.957E-9
GO:0050793 regulation of developmental process	13	1.355E-8
GO:0051093 negative regulation of developmental process	10	1.501E-8
GO:0019827 stem cell maintenance	6	2.525E-8
GO:0048468 cell development	13	2.638E-8
GO:0048864 stem cell development	6	2.900E-8
GO:0009888 tissue development	13	3.913E-8
GO:0048646 anatomical structure formation involved in morphogenesis	11	3.778E-8
GO:0045596 negative regulation of cell differentiation	9	5.834E-8
GO:0045595 regulation of cell differentiation	11	5.952E-8
GO:0042127 regulation of cell proliferation	12	8.561E-8
GO:0048863 stem cell differentiation	6	1.114E-7
GO:0043009 chordate embryonic development	11	1.221E-7
GO:0009792 embryonic development ending in birth or egg hatching	11	1.289E-7
GO:0045165 cell fate commitment	8	3.238E-7
GO:0050789 regulation of biological process	28	3.465E-7
GO:0007167 enzyme linked receptor protein signaling pathway	9	1.042E-6
GO:0065007 biological regulation	28	1.328E-6
GO:0050794 regulation of cellular process	27	1.525E-6
GO:0009887 organ morphogenesis	11	2.550E-6
GO:0006357 regulation of transcription from RNA polymerase II promoter	11	3.481E-6
GO:0007243 protein kinase cascade	8	6.748E-6
GO:0045944 positive regulation of transcription from RNA polymerase II promoter	9	7.026E-6
GO:0045893 positive regulation of transcription, DNA-dependent	9	2.135E-5
GO:0008284 positive regulation of cell proliferation	8	2.171E-5
GO:0051254 positive regulation of RNA metabolic process	9	2.141E-5
GO:0032501 multicellular organismal process	22	3.182E-5
GO:0007169 transmembrane receptor protein tyrosine kinase signaling pathway	7	3.416E-5
GO:0010604 positive regulation of macromolecule metabolic process	10	4.234E-5
GO:0031325 positive regulation of cellular metabolic process	10	4.771E-5
GO:0045941 positive regulation of transcription	9	4.873E-5
GO:0010628 positive regulation of gene expression	9	5.828E-5
GO:0044260 cellular macromolecule metabolic process	22	6.809E-5
GO:0009893 positive regulation of metabolic process	10	7.095E-5
GO:0045935 positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	9	7.588E-5
GO:0007242 intracellular signaling cascade	11	9.113E-5
GO:0051173 positive regulation of nitrogen compound metabolic process	9	9.160E-5
GO:0010557 positive regulation of macromolecule biosynthetic process	9	9.500E-5
GO:0048568 embryonic organ development	7	1.007E-4
GO:0031328 positive regulation of cellular biosynthetic process	9	1.234E-4
GO:0009891 positive regulation of biosynthetic process	9	1.294E-4
GO:0051239 regulation of multicellular organismal process	10	1.755E-4
GO:0000003 reproduction	9	2.858E-4

Enriched terms	count	*adj. p-values*
GO:0007399 nervous system development	10	3.578E-4
GO:0001890 placenta development	5	3.527E-4
GO:0019222 regulation of metabolic process	17	3.675E-4
GO:0043170 macromolecule metabolic process	22	4.118E-4
GO:0040007 growth	6	4.679E-4
GO:0030326 embryonic limb morphogenesis	5	5.073E-4
GO:0035113 embryonic appendage morphogenesis	5	5.073E-4
GO:0048589 developmental growth	5	5.629E-4
GO:0001707 mesoderm formation	4	6.255E-4
GO:0006355 regulation of transcription, DNA-dependent	12	7.191E-4
GO:0051094 positive regulation of developmental process	6	7.089E-4
GO:0048332 mesoderm morphogenesis	4	7.045E-4
GO:0001710 mesodermal cell fate commitment	3	7.035E-4
GO:0051252 regulation of RNA metabolic process	12	7.834E-4
GO:0010468 regulation of gene expression	15	7.754E-4
GO:0031323 regulation of cellular metabolic process	16	7.767E-4
GO:0001704 formation of primary germ layer	4	7.666E-4
GO:0000165 MAPKKK cascade	5	8.128E-4
GO:0035107 appendage morphogenesis	5	8.297E-4
GO:0035108 limb morphogenesis	5	8.297E-4
GO:0048333 mesodermal cell differentiation	3	8.931E-4
GO:0000122 negative regulation of transcription from RNA polymerase II promoter	6	8.864E-4
GO:0048736 appendage development	5	9.114E-4
GO:0060173 limb development	5	9.114E-4
GO:0010646 regulation of cell communication	9	9.341E-4
GO:0007166 cell surface receptor linked signal transduction	15	9.336E-4
GO:0048729 tissue morphogenesis	6	9.703E-4
GO:0030182 neuron differentiation	7	0.001
GO:0022414 reproductive process	8	0.001
GO:0006468 protein amino acid phosphorylation	8	0.001
GO:0060255 regulation of macromolecule metabolic process	15	0.002
GO:0001708 cell fate specification	4	0.002
GO:0007498 mesoderm development	4	0.002
GO:0007420 brain development	6	0.002
GO:0009966 regulation of signal transduction	8	0.002
GO:0042221 response to chemical stimulus	9	0.002
GO:0001892 embryonic placenta development	4	0.002
GO:0010556 regulation of macromolecule biosynthetic process	14	0.002
GO:0044237 cellular metabolic process	22	0.002
GO:0045892 negative regulation of transcription, DNA-dependent	6	0.002
GO:0051253 negative regulation of RNA metabolic process	6	0.002
GO:0006350 transcription	12	0.002
GO:0007439 ectodermal gut development	3	0.003
GO:0048567 ectodermal gut morphogenesis	3	0.003
GO:0031326 regulation of cellular biosynthetic process	14	0.003
GO:0048699 generation of neurons	7	0.003
GO:0009889 regulation of biosynthetic process	14	0.003
GO:0045597 positive regulation of cell differentiation	5	0.003
GO:0016310 phosphorylation	8	0.003
GO:0007369 gastrulation	4	0.003
GO:0001829 trophectodermal cell differentiation	3	0.003
GO:0006464 protein modification process	10	0.003
GO:0045995 regulation of embryonic development	3	0.004
GO:0022008 neurogenesis	7	0.004
GO:0045449 regulation of transcription	13	0.004
GO:0048732 gland development	5	0.004
GO:0043412 biopolymer modification	10	0.005
GO:0007417 central nervous system development	6	0.005

Enriched terms	count	*adj. p-values*
GO:0016481 negative regulation of transcription	6	0.005
GO:0043687 post-translational protein modification	9	0.005
GO:0009987 cellular process	27	0.005
GO:0051726 regulation of cell cycle	5	0.005
GO:0080090 regulation of primary metabolic process	14	0.005
GO:0001825 blastocyst formation	3	0.006
GO:0019219 regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	13	0.006
GO:0045934 negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	6	0.007
GO:0051171 regulation of nitrogen compound metabolic process	13	0.007
GO:0051172 negative regulation of nitrogen compound metabolic process	6	0.007
GO:0034645 cellular macromolecule biosynthetic process	13	0.007
GO:0010629 negative regulation of gene expression	6	0.007
GO:0009059 macromolecule biosynthetic process	13	0.007
GO:0006793 phosphorus metabolic process	8	0.008
GO:0006796 phosphate metabolic process	8	0.008
GO:0043066 negative regulation of apoptosis	5	0.008
GO:0045667 regulation of osteoblast differentiation	3	0.008
GO:0048547 gut morphogenesis	3	0.008
GO:0010558 negative regulation of macromolecule biosynthetic process	6	0.008
GO:0001568 blood vessel development	5	0.008
GO:0043069 negative regulation of programmed cell death	5	0.008
GO:0060548 negative regulation of cell death	5	0.008
GO:0008283 cell proliferation	5	0.008
GO:0001944 vasculature development	5	0.008
GO:0031327 negative regulation of cellular biosynthetic process	6	0.009
GO:0009890 negative regulation of biosynthetic process	6	0.009
GO:0031667 response to nutrient levels	4	0.009
GO:0007492 endoderm development	3	0.009
GO:0032526 response to retinoic acid	3	0.009
GO:0033189 response to vitamin A	3	0.010
GO:0003006 reproductive developmental process	5	0.010
GO:0033273 response to vitamin	3	0.011
GO:0060541 respiratory system development	4	0.011
GO:0001763 morphogenesis of a branching structure	4	0.011
GO:0048546 digestive tract morphogenesis	3	0.011
GO:0055123 digestive system development	3	0.012
GO:0016055 Wnt receptor signaling pathway	4	0.012
GO:0060711 labyrinthine layer development	3	0.012
GO:0009880 embryonic pattern specification	3	0.012
GO:0001501 skeletal system development	5	0.013
GO:0007389 pattern specification process	5	0.013
GO:0009991 response to extracellular stimulus	4	0.013
GO:0007398 ectoderm development	4	0.013
GO:0051049 regulation of transport	5	0.013
GO:0016043 cellular component organization	11	0.013
GO:0031324 negative regulation of cellular metabolic process	6	0.013
GO:0048565 gut development	3	0.015
GO:0010033 response to organic substance	6	0.015
GO:0010605 negative regulation of macromolecule metabolic process	6	0.015
GO:0044238 primary metabolic process	21	0.015
GO:0030030 cell projection organization	5	0.018
GO:0007165 signal transduction	13	0.018
GO:0009892 negative regulation of metabolic process	6	0.019
GO:0009725 response to hormone stimulus	4	0.022
GO:0030900 forebrain development	4	0.022
GO:0051716 cellular response to stimulus	6	0.022
GO:0050896 response to stimulus	12	0.024
GO:0032989 cellular component morphogenesis	5	0.024
GO:0030278 regulation of ossification	3	0.025
GO:0001824 blastocyst development	3	0.026
GO:0008152 metabolic process	22	0.026
GO:0009719 response to endogenous stimulus	4	0.028

Enriched terms	count	*adj. p-values*
GO:0032870 cellular response to hormone stimulus	3	0.029
GO:0007049 cell cycle	6	0.031
GO:0006950 response to stress	8	0.034
GO:0050678 regulation of epithelial cell proliferation	3	0.034
GO:0010467 gene expression	12	0.034
GO:0044249 cellular biosynthetic process	13	0.036
GO:0009605 response to external stimulus	6	0.037
GO:0032879 regulation of localization	5	0.038
GO:0048852 diencephalon morphogenesis	2	0.038
GO:0000904 cell morphogenesis involved in differentiation	4	0.039
GO:0031175 neuron projection development	4	0.042
GO:0051051 negative regulation of transport	3	0.042
GO:0035270 endocrine system development	3	0.042
GO:0006366 transcription from RNA polymerase II promoter	3	0.044
GO:0007507 heart development	4	0.044
GO:0009058 biosynthetic process	13	0.044
GO:0007584 response to nutrient	3	0.045
GO:0051216 cartilage development	3	0.047
mmu05200:Pathways in cancer	13	1.902E-9
mmu05220:Chronic myeloid leukemia	8	7.787E-8
mmu05210:Colorectal cancer	8	1.252E-7
mmu05215:Prostate cancer	8	1.296E-7
mmu05213:Endometrial cancer	7	1.327E-7
mmu05221:Acute myeloid leukemia	7	1.952E-7
mmu05211:Renal cell carcinoma	7	5.896E-7
mmu04722:Neurotrophin signaling pathway	8	8.510E-7
mmu04662:B cell receptor signaling pathway	7	1.030E-6
mmu04012:ErbB signaling pathway	7	1.537E-6
mmu04916:Melanogenesis	7	3.216E-6
mmu05223:Non-small cell lung cancer	6	3.483E-6
mmu04062:Chemokine signaling pathway	8	5.253E-6
mmu04660:T cell receptor signaling pathway	7	6.742E-6
mmu05214:Glioma	6	6.581E-6
mmu04510:Focal adhesion	8	7.542E-6
mmu05218:Melanoma	6	9.775E-6
mmu05212:Pancreatic cancer	6	9.901E-6
mmu04910:Insulin signaling pathway	7	1.245E-5
mmu04664:Fc epsilon RI signaling pathway	6	1.702E-5
mmu04650:Natural killer cell mediated cytotoxicity	6	1.134E-4
mmu04320:Dorso-ventral axis formation	4	1.346E-4
mmu04370:VEGF signaling pathway	5	2.677E-4
mmu04010:MAPK signaling pathway	7	4.049E-4
mmu04540:Gap junction	5	3.992E-4
mmu04912:GnRH signaling pathway	5	6.119E-4
mmu05219:Bladder cancer	4	7.840E-4
mmu04810:Regulation of actin cytoskeleton	6	0.001
mmu04310:Wnt signaling pathway	5	0.002
mmu04630:Jak-STAT signaling pathway	5	0.002
mmu04720:Long-term potentiation	4	0.003
mmu04730:Long-term depression	4	0.003
mmu04914:Progesterone-mediated oocyte maturation	4	0.005
mmu04666:Fc gamma R-mediated phagocytosis	4	0.007
mmu05216:Thyroid cancer	3	0.007
mmu05020:Prion diseases	3	0.010
mmu04360:Axon guidance	4	0.015
mmu05217:Basal cell carcinoma	3	0.023
mmu04920:Adipocytokine signaling pathway	3	0.033
mmu04520:Adherens junction	3	0.041
mmu04350:TGF-beta signaling pathway	3	0.051

Table S8: Runtime to determine an optimal solution for individual modules in the BC network with differing number of nodes and edges. Listed is also the resulting component density. All computations were conducted on a single threaded Intel XEON machine running at 2.2 Ghz. The module (grey) could not be solved in appropriate time using ILP (not even after a month)

Module	Nodes	Edges	Density	Running Time (s)
black	41	233	0.13	1.91
brown	195	18843	0.49	52.71
green	110	595	0.04	94.60
magenta	26	123	0.18	1.91
pink	30	149	0.16	1.82
red	93	473	0.05	3.02
yellow	132	663	0.03	3499.55
blue	247	30239	0.49	14287.19
turquoise	295	43213	0.49	266.18
grey	148	723	0.03	NA

Table S9: Identified genes in the MCDS (heuristic approach) for 10 modules of the breast cancer network. The genes, whose protein products are known to be targeted by drugs, are marked in bold.

Method	Module	Network Size	Result size	Key driver genes
MCDS	black	41	8	ZNF254, SEC24B, ZNF681, ZZZ3, WNT5B, **CEP350**, ZNF426, ZNF137
MCDS	blue	247	3	FAM54A,**ACAN**, PDGFA
MCDS	brown	195	1	**AATK**
MCDS	green	110	29	GCDH, HDAC10, FLII, PIGQ, CXXC1, CNDP2, DPP7, **OS9**, MAGOH, **AKT1**, E4F1, UTP14A, SH3GLB2, ATPBD4, C9orf7, **AP1B1**, TIMM44, WDR13, USF2, USF1, CDK9, UCK1, **CDC34**, PQLC1, KIAA0664, **CDC37**, C7orf27, **CDK10**, LTBR
MCDS	grey	148	38	FAM59A, FBN2, COLEC10, FRMPD1, IL5RA, RORA, GEFT, RNF2, **CACNA1H**, CLGN, **CAMK2N1**, CEP72, **CA6**, PRND, ZC3H14, OR7A17, PRPH, **BRD1**, TAF15, ZRANB2, ZNF480, **ANXA13**, UPF3A , SNCAIP, HPCAL4, MR1, POU4F2, SYT6, DHDDS, CXorf26, GRHPR, PHF16, CNTN4, LHX4, **CCDC130**, PPARA , **ABCG4**, SPRR1B
MCDS	magenta	26	7	VPS72, **BGLAP**, SESN2, TAF7, MED26, ILF2, **ATF6**
MCDS	pink	30	6	ZNF706 , TCEB1, CHRAC1, DHX35, ZNF250, TMEM70
MCDS	red	93	22	TGIF1, ZNF485, ZNF691, **TGFB1**, DHX8, USP21, PHF20, GTF3A, FHL3, RPS3A , JOSD2 , **ATP1B1**, SUMF2, EPN3, HTR6, PPP4C, **CCDC9**, **MYC**, UBAP1, PCGF1, C6orf134, **TP53**
MCDS	turquoise	295	1	**ABHD10**
MCDS	yellow	132	26	SPI1, TRAF3IP3, HDAC11, HTRA4, CXCR4, IL2RG, ETS1, FUT4, FAM129C, FAM124B , **CASP10**, RASSF5, PHACTR2, TSPAN2, PAG1, SLAMF1, SLC31A2, **DAPK1**, TNFRSF9, **NFKB1**, FLI1, GZMK, SLFN11, PRKD3, LCP2, DFNA5

Table S10: Enriched GO terms and KEGG pathways with adjusted p-values < 0.05 for the 141 genes in the aggregated MCDS for the modules of the breast cancer network. p-values were adjusted for multiple testing using the BH procedure.

Enriched terms	count	*adj. p-values*
GO:0009059~macromolecule biosynthetic process	43	0.001
GO:0010468~regulation of gene expression	43	0.001
GO:0006357~regulation of transcription from RNA polymerase II promoter	20	0.001
GO:0032774~RNA biosynthetic process	12	0.001
GO:0051254~positive regulation of RNA metabolic process	15	0.001
GO:0006350~transcription	37	0.001
GO:0010556~regulation of macromolecule biosynthetic process	43	0.001
GO:0032583~regulation of gene-specific transcription	9	0.001
GO:0009889~regulation of biosynthetic process	43	0.001
GO:0009891~positive regulation of biosynthetic process	18	0.001
GO:0045449~regulation of transcription	40	0.001
GO:0010467~gene expression	43	0.001
GO:0006351~transcription, DNA-dependent	12	0.001
GO:0045893~positive regulation of transcription, DNA-dependent	15	0.001
GO:0034645~cellular macromolecule biosynthetic process	43	0.001
GO:0031326~regulation of cellular biosynthetic process	43	0.001
GO:0051171~regulation of nitrogen compound metabolic process	42	0.002
GO:0019219~regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	41	0.002
GO:0060255~regulation of macromolecule metabolic process	45	0.002
GO:0051173~positive regulation of nitrogen compound metabolic process	17	0.002
GO:0000122~negative regulation of transcription from RNA polymerase II promoter	11	0.002
GO:0019222~regulation of metabolic process	48	0.002
GO:0010557~positive regulation of macromolecule biosynthetic process	17	0.002
GO:0006139~nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	46	0.002
GO:0043170~macromolecule metabolic process	65	0.003
GO:0034641~cellular nitrogen compound metabolic process	48	0.003
GO:0031323~regulation of cellular metabolic process	46	0.003
GO:0031328~positive regulation of cellular biosynthetic process	17	0.003
GO:0045935~positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	16	0.003
GO:0010551~regulation of specific transcription from RNA polymerase II promoter	7	0.003
GO:0080090~regulation of primary metabolic process	44	0.004
GO:0045941~positive regulation of transcription	15	0.004
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	12	0.004
GO:0044260~cellular macromolecule metabolic process	60	0.004
GO:0044249~cellular biosynthetic process	45	0.004
GO:0006807~nitrogen compound metabolic process	48	0.004
GO:0010628~positive regulation of gene expression	15	0.004
GO:0006355~regulation of transcription, DNA-dependent	29	0.005
GO:0051252~regulation of RNA metabolic process	29	0.006
GO:0009058~biosynthetic process	45	0.008
GO:0045892~negative regulation of transcription, DNA-dependent	11	0.013
GO:0006366~transcription from RNA polymerase II promoter	9	0.014
GO:0051253~negative regulation of RNA metabolic process	11	0.014
GO:0010629~negative regulation of gene expression	13	0.014
GO:0043193~positive regulation of gene-specific transcription	6	0.016
GO:0009893~positive regulation of metabolic process	18	0.018
GO:0016070~RNA metabolic process	18	0.021
GO:0031667~response to nutrient levels	8	0.021
GO:0010604~positive regulation of macromolecule metabolic process	17	0.022
GO:0016481~negative regulation of transcription	12	0.022
GO:0031325~positive regulation of cellular metabolic process	17	0.027
GO:0031327~negative regulation of cellular biosynthetic process	13	0.031
GO:0045595~regulation of cell differentiation	12	0.034
GO:0009890~negative regulation of biosynthetic process	13	0.036
GO:0009991~response to extracellular stimulus	8	0.037
GO:0019216~regulation of lipid metabolic process	6	0.040
GO:0045934~negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	12	0.044
GO:0051172~negative regulation of nitrogen compound metabolic process	12	0.048
GO:0009892~negative regulation of metabolic process	15	0.055
hsa05200:Pathways in cancer	12	0.004

Table S11: The integrated databases and interaction types in TFmiR. (P) means predicted interactions and (E) means experimentally validated interactions.

Interaction	Databases $(P/E)^*$	Genes	miRNAs	Regulatory links	Version /frozen data
$TF \rightarrow gene$	TRANSFAC (E)	1279	–	2943	V11.4
$TF \rightarrow gene$	OregAnno (E)	1132	–	1083	Nov 2010
$TF \rightarrow gene$	TRED (P)	3038	–	6462	2007
$TF \rightarrow miRNA$	TransmiR (E)	158	175	567	V1.2, Jan 2013
$TF \rightarrow miRNA$	PMID20584335 (E)	58	56	102	Apr 2009
$TF \rightarrow miRNA$	ChipBase (P)	119	1380	33087	V1.1, Nov 2012
$miRNA \rightarrow gene$	miRTarBase (E)	2244	551	5640	V4.5, Nov 2013
$miRNA \rightarrow gene$	TarBase (E)	422	79	492	V7.0
$miRNA \rightarrow gene$	miRecords (E)	543	157	780	Mar 2009
$miRNA \rightarrow gene$	starBase (P)	5720	249	56051	V2.0, Sep 2013
$miRNA \rightarrow miRNA$	PmmR (P)	–	312	3846	Mar 2011

Table S12: The most significant functions and diseases enriched in the miRNA nodes of the breast cancer disease network (Hamed et al., 2015b).

Category	Term	miRNAs Count	P-value
Function	Epithelial-mesenchymal transition	17	0.022
Function	glucose metabolism	4	0.048
Disease	Breast Neoplasms	67	1.43E-25
Disease	Lung Neoplasms	50	4.33E-17
Disease	Neoplasms	44	3.15E-15
Disease	Ovarian Neoplasms	43	1.30E-14
Disease	Adenocarcinoma	27	2.59E-13
Disease	Pancreatic Neoplasms	39	7.30E-13
Disease	Prostatic Neoplasms	41	3.49E-12
Disease	Melanoma	45	1.25E-11
Disease	Colonic Neoplasms	32	4.6E-11
Disease	Colorectal Neoplasms	45	5.69E-11

Table S13: Key genes and miRNAs in the breast cancer network (Hamed et al., 2015b).

Key genes	E2F6, TP53, SPI1, TGFB1, SMAD4, ESR1, TERT, E2F3, BRCA2, AKT1
Key miRNAs	hsa-mir-148a, hsa-mir-21, hsa-mir-93, hsa-mir-152, hsa-mir-106b, hsa-mir-143, hsa-mir-200c, hsa-mir-27a, hsa-mir-23a, hsa-mir-22, , hsa-mir-146a, hsa-mir-335

Table S14: The identified key gene nodes in the breast cancer network (Hamed et al., 2015b) whose protein products are targeted by anti-cancer drugs. (1) means that at least one drug that targets this gene product is reported in this database, and (0) means no drugs are reported for the respective gene in this database. Not included are substances that are known to be cancerogenous or mutagenic.

Target gene	Drug and antineoplastic agents	CTD	PharmGKB	Cancer Resource
AKT1	U 0126;tyrphostin AG 1478; Ursodeoxycholic Acid;Valproic Acid;tyrphostin AG 1024; trametinib; Tretinoin	1	0	1
BRCA2	Tretinoin; trichostatin A; Estradiol; transplatin; troglitazone; Tunicamycin; fulvestrant	1	0	1
ESR1	exemestane;tamoxifen	0	1	1
TGFB1	Doxorubicin; Fluorouracil; Thalidomide; Entinostat; Hyaluronidase	0	0	1
TP53	4-biphenylmine; alliin; Apigenin; Atropine;bicalutamide;butylidenephthalide	0	0	1

Table S15: 82 Candidates for hepatocelular carcinoma in the fourth layer identified by TopControl. They were sorted initially by their scores, then by LFC. D stands for degree of the node and LFC for $\log_2$(fold change).

gene	D	hub	mds	mcds	score	LFC
E2F1	25	1	1	1	3	3.76
EGR1	18	1	1	1	3	2.33
ESR1	9	1	1	1	3	2.19
JUN	33	1	1	1	3	1.39
NR1I2	9	1	1	1	3	1.37
MYC	18	1	1	1	3	1.07
JUND	6	1	1	1	3	0.99
STAT3	10	1	1	1	3	0.81
USF1	15	1	1	1	3	0.73
NR1H4	7	1	1	1	3	0.63
ETS1	9	1	1	1	3	0.61
IRF1	5	1	1	1	3	0.61
NR1I3	8	1	1	1	3	0.6
hsa-let-7b	47	1	1	1	3	-
hsa-mir-26a-5p	16	1	1	1	3	-
hsa-mir-29a	21	1	1	1	3	-
hsa-mir-29a-3p	18	1	1	1	3	-
hsa-mir-34a-5p	28	1	1	1	3	-
POU3F2	2	0	1	1	2	7.44
TP73	4	0	1	1	2	3.92
FOXM1	3	0	1	1	2	3.69
MYCN	3	0	1	1	2	3.25
ETV4	4	0	1	1	2	3.02
FOS	23	1	0	1	2	2.93
IL1B	1	0	1	1	2	2.25
NOS2	5	1	0	1	2	2.2
NR4A1	3	0	1	1	2	2.13
HBB	6	1	0	1	2	2.05
FOXO1	3	0	1	1	2	1.84
CYP3A4	6	1	0	1	2	1.84
IRF8	2	0	1	1	2	1.76
KCNIP3	1	0	1	1	2	1.76
ETS2	5	0	1	1	2	1.68
JUNB	2	0	1	1	2	1.55
FOSL1	10	1	0	1	2	1.37
SATB1	1	0	1	1	2	1.32
CEBPD	1	0	1	1	2	1.3
KLF6	2	0	1	1	2	1.28
KLF4	2	0	1	1	2	1.27
MAFG	3	0	1	1	2	1.21
KLF11	3	0	1	1	2	1.15
PDGFB	1	0	1	1	2	1.11
HIVEP1	1	0	1	1	2	1.04
NME2	1	0	1	1	2	0.95
CYBB	5	1	0	1	2	0.95
TCF3	2	0	1	1	2	0.92
ZBTB7B	1	0	1	1	2	0.88
TNFRSF1A	1	0	1	1	2	0.82
MAZ	4	0	1	1	2	0.8
GATA4	2	0	1	1	2	0.76
F12	1	0	1	1	2	0.75
CREM	3	0	1	1	2	0.74
CNBP	1	0	1	1	2	0.73
FOXA3	1	0	1	1	2	0.7
HLTF	1	0	1	1	2	0.7
SREBF2	3	0	1	1	2	0.68
NFE2L2	4	0	1	1	2	0.66
COL2A1	2	0	0	1	1	10.97
TERT	6	1	0	0	1	9.17
HOXD10	1	0	1	0	1	5.57
OTX1	1	0	1	0	1	5.53
CDK1	2	0	0	1	1	3.41
RRM2	2	0	0	1	1	2.98
SERPINE1	4	0	0	1	1	1.76
HMGA1	5	0	0	1	1	1.65
IGFBP1	2	0	0	1	1	1.45
GTF2IRD1	1	0	1	0	1	1.3
WEE1	1	0	1	0	1	1.25
COL1A2	4	0	0	1	1	1.18
PLAU	8	1	0	0	1	1.1
APOH	6	1	0	0	1	1.09
APOA5	2	0	0	1	1	1.05
ABCA1	2	0	0	1	1	1.02
AR	3	0	1	0	1	1
GATA6	1	0	1	0	1	0.9
HSF1	1	0	1	0	1	0.85
RORA	1	0	1	0	1	0.81
OAS1	1	0	1	0	1	0.8
NFKB2	2	0	1	0	1	0.76
HTATIP2	2	0	0	1	1	0.68
CCND1	12	1	0	0	1	0.67
ALDOC	2	0	0	1	1	0.67

Table S16: Enriched GO terms (top) and KEGG pathways (bottom line) with adjusted p-values < 0.05 for the hubs in the hepatocellular carcinoma network. P-values were adjusted for multiple testing using the BH procedure.

Enriched terms	count	*adj. p-values*
GO:0045944 positive regulation of transcription from RNA polymerase II promoter	15	5.185E-10
GO:0006366 transcription from RNA polymerase II promoter	10	1.908E-6
GO:0045893 positive regulation of transcription, DNA-templated	10	1.316E-6
GO:0042493 response to drug	8	1.016E-5
GO:1902895 positive regulation of pri-miRNA transcription from RNA polymerase II promoter	4	2.219E-4
GO:0032355 response to estradiol	5	4.619E-4
GO:0000122 negative regulation of transcription from RNA polymerase II promoter	8	0.001
GO:0032870 cellular response to hormone stimulus	4	0.001
GO:0051591 response to cAMP	4	0.001
GO:0034097 response to cytokine	4	0.002
GO:0048146 positive regulation of fibroblast proliferation	4	0.002
GO:0043401 steroid hormone mediated signaling pathway	4	0.002
GO:0009612 response to mechanical stimulus	4	0.002
GO:0001666 response to hypoxia	5	0.002
GO:0051412 response to corticosterone	3	0.008
GO:0051726 regulation of cell cycle	4	0.017
GO:0045787 positive regulation of cell cycle	3	0.026
GO:0006367 transcription initiation from RNA polymerase II promoter	4	0.027
GO:0030522 intracellular receptor signaling pathway	3	0.029
GO:0007568 aging	4	0.031
GO:0008285 negative regulation of cell proliferation	5	0.037
GO:0042127 regulation of cell proliferation	4	0.039
GO:0042542 response to hydrogen peroxide	3	0.043
GO:0071277 cellular response to calcium ion	3	0.043
GO:0006357 regulation of transcription from RNA polymerase II promoter	5	0.048
hsa05166:HTLV-I infection	9	9.692E-6
hsa05161:Hepatitis B	6	0.001
hsa04917:Prolactin signaling pathway	5	8.985E-4
hsa04380:Osteoclast differentiation	5	0.007
hsa05200:Pathways in cancer	7	0.006
hsa05210:Colorectal cancer	4	0.008
hsa05133:Pertussis	4	0.012
hsa05222:Small cell lung cancer	4	0.015
hsa05205:Proteoglycans in cancer	5	0.016
hsa05219:Bladder cancer	3	0.046
hsa04310:Wnt signaling pathway	4	0.043
hsa05206:MicroRNAs in cancer	5	0.043

Table S17: Enriched GO terms (top) and KEGG pathways (bottom line) with adjusted p-values < 0.05 for the MDS in the hepatocellular carcinoma network. P-values were adjusted for multiple testing using the BH procedure.

Enriched terms	count	*adj. p-values*
GO:0045944 positive regulation of transcription from RNA polymerase II promoter	43	3.867E-38
GO:0006366 transcription from RNA polymerase II promoter	25	1.062E-19
GO:0045893 positive regulation of transcription, DNA-templated	24	1.923E-18
GO:0000122 negative regulation of transcription from RNA polymerase II promoter	22	6.235E-13
GO:0006351 transcription, DNA-templated	30	1.904E-11
GO:0010628 positive regulation of gene expression	12	9.784E-8
GO:0045892 negative regulation of transcription, DNA-templated	13	6.778E-6
GO:0030522 intracellular receptor signaling pathway	6	1.504E-5
GO:0006367 transcription initiation from RNA polymerase II promoter	8	5.651E-5
GO:0043401 steroid hormone mediated signaling pathway	6	9.578E-5
GO:0051726 regulation of cell cycle	7	2.341E-4
GO:0008285 negative regulation of cell proliferation	10	3.478E-4
GO:0006357 regulation of transcription from RNA polymerase II promoter	10	7.629E-4
GO:0006355 regulation of transcription, DNA-templated	17	9.599E-4
GO:0048146 positive regulation of fibroblast proliferation	5	0.001
GO:0070301 cellular response to hydrogen peroxide	5	0.001
GO:0009612 response to mechanical stimulus	5	0.002
GO:0042493 response to drug	8	0.002
GO:0050728 negative regulation of inflammatory response	5	0.005
GO:0032873 negative regulation of stress-activated MAPK cascade	3	0.006
GO:0045597 positive regulation of cell differentiation	4	0.009
GO:0071499 cellular response to laminar fluid shear stress	3	0.010
GO:0042127 regulation of cell proliferation	6	0.012
GO:2000188 regulation of cholesterol homeostasis	3	0.012
GO:0032870 cellular response to hormone stimulus	4	0.014
GO:0051591 response to cAMP	4	0.014
GO:0045766 positive regulation of angiogenesis	5	0.016
GO:0010941 regulation of cell death	3	0.016
GO:0034097 response to cytokine	4	0.018
GO:0008284 positive regulation of cell proliferation	8	0.021
GO:0010629 negative regulation of gene expression	5	0.027
GO:0007165 signal transduction	12	0.030
GO:0060337 type I interferon signaling pathway	4	0.029
GO:0006006 glucose metabolic process	4	0.033
GO:0045444 fat cell differentiation	4	0.040
GO:1902895 positive regulation of pri-miRNA transcription from RNA polymerase II promoter	3	0.043
GO:0032496 response to lipopolysaccharide	5	0.044
GO:0007568 aging	5	0.044
hsa05166:HTLV-I infection	11	6.294E-5
hsa04010:MAPK signaling pathway	8	0.0181
hsa04380:Osteoclast differentiation	6	0.021

Table S18: Enriched GO terms (top) and KEGG pathways (bottom line) with adjusted p-values < 0.05 for the MCDS in the hepatocellular carcinoma network. P-values were adjusted for multiple testing using the BH procedure.

Enriched terms	count	*adj. p-values*
GO:0045944 positive regulation of transcription from RNA polymerase II promoter	40	5.582E-30
GO:0006366 transcription from RNA polymerase II promoter	25	3.745E-18
GO:0045893 positive regulation of transcription, DNA-templated	23	1.051E-15
GO:0000122 negative regulation of transcription from RNA polymerase II promoter	19	1.414E-8
GO:0010628 positive regulation of gene expression	12	5.040E-7
GO:0042493 response to drug	12	1.973E-6
GO:0006351 transcription, DNA-templated	25	5.148E-6
GO:0045892 negative regulation of transcription, DNA-templated	13	2.690E-5
GO:0051591 response to cAMP	6	7.307E-5
GO:0009612 response to mechanical stimulus	6	2.315E-4
GO:0006357 regulation of transcription from RNA polymerase II promoter	11	3.759E-4
GO:0008285 negative regulation of cell proliferation	10	0.001
GO:0032870 cellular response to hormone stimulus	5	0.001
GO:0034097 response to cytokine	5	0.002
GO:0048146 positive regulation of fibroblast proliferation	5	0.002
GO:1902895 positive regulation of pri-miRNA transcription from RNA polymerase II promoter	4	0.003
GO:0043401 steroid hormone mediated signaling pathway	5	0.003
GO:0070301 cellular response to hydrogen peroxide	5	0.003
GO:0042127 regulation of cell proliferation	7	0.003
GO:0045766 positive regulation of angiogenesis	6	0.003
GO:0051726 regulation of cell cycle	6	0.004
GO:0010629 negative regulation of gene expression	6	0.006
GO:0032873 negative regulation of stress-activated MAPK cascade	3	0.008
GO:0006367 transcription initiation from RNA polymerase II promoter	6	0.009
GO:0007568 aging	6	0.013
GO:0030522 intracellular receptor signaling pathway	4	0.013
GO:0071499 cellular response to laminar fluid shear stress	3	0.013
GO:0045429 positive regulation of nitric oxide biosynthetic process	4	0.018
GO:0010941 regulation of cell death	3	0.023
GO:0071222 cellular response to lipopolysaccharide	5	0.025
GO:0071277 cellular response to calcium ion	4	0.026
GO:0006915 apoptotic process	9	0.032
GO:0030194 positive regulation of blood coagulation	3	0.033
GO:0008284 positive regulation of cell proliferation	8	0.043
hsa05166:HTLV-I infection	12	3.013E-5
hsa04380:Osteoclast differentiation	8	7.793E-4
hsa05133:Pertussis	6	0.003
hsa05142:Chagas disease (American trypanosomiasis)	6	0.012
hsa04010:MAPK signaling pathway	8	0.020

Table S19: 140 candidates for breast neoplasms in the fourth layer identified by TopControl. They were sorted initially by their scores, then by LFC. D stands for degree of the node and LFC for $\log_2$(fold change).

gene	D	hub	mds	mcds	score	LFC
EGR1	19	1	1	1	3	2.59
ESR2	7	1	1	1	3	2.58
FOS	20	1	1	1	3	2.47
E2F1	25	1	1	1	3	2.34
CEBPA	17	1	1	1	3	2.1
ESR1	19	1	1	1	3	1.79
JUN	45	1	1	1	3	1.6
STAT5A	7	1	1	1	3	1.6
RUNX2	5	1	1	1	3	1.4
STAT1	34	1	1	1	3	1.28
ETS2	6	1	1	1	3	1.15
MITF	8	1	1	1	3	0.94
NR1H3	8	1	1	1	3	0.9
TFAP2A	24	1	1	1	3	0.9
IRF1	16	1	1	1	3	0.65
ARHGEF7	11	1	1	1	3	0.62
USF1	16	1	1	1	3	0.61
SRF	5	1	1	1	3	0.58
TFDP1	10	1	1	1	3	0.58
hsa-mir-1	86	1	1	1	3	-
hsa-mir-145-5p	39	1	1	1	3	-
hsa-mir-146a	31	1	1	1	3	-
hsa-mir-21	44	1	1	1	3	-
hsa-mir-21-5p	32	1	1	1	3	-
hsa-mir-34a-5p	27	1	1	1	3	-
EFNA2	2	0	1	1	2	Inf
CGA	5	1	0	1	2	7.25
GBX2	1	0	1	1	2	5.7
GATA4	3	0	1	1	2	5.6
WT1	2	0	1	1	2	5.39
LHX2	1	0	1	1	2	4.88
HBB	5	1	0	1	2	4.73
BMPR1B	1	0	1	1	2	4.41
POU3F2	1	0	1	1	2	4.14
IFNB1	9	1	0	1	2	4.13
RRM2	2	0	1	1	2	3.69
FOXM1	1	0	1	1	2	3.54
KIT	6	1	0	1	2	2.79
IL6	7	1	0	1	2	2.76
HOXA5	2	0	1	1	2	2.55
OTX1	1	0	1	1	2	2.49
TFF3	1	0	1	1	2	2.35
NR4A1	4	0	1	1	2	2.17
TAL1	1	0	1	1	2	1.92
BMP6	1	0	1	1	2	1.87
FOXA1	3	0	1	1	2	1.87
GATA3	3	0	1	1	2	1.75
SOX10	1	0	1	1	2	1.66
IRF7	5	0	1	1	2	1.6
NR3C1	8	1	0	1	2	1.57
PLAU	8	1	0	1	2	1.49
THRB	3	0	1	1	2	1.33
NR5A2	1	0	1	1	2	1.29
LMO2	1	0	1	1	2	1.26
MAZ	4	0	1	1	2	1.24
STAT5B	4	0	1	1	2	1.24
SATB1	1	0	1	1	2	1.16
FLI1	5	0	1	1	2	1.14
JUNB	2	0	1	1	2	1.12
KLF8	1	0	1	1	2	1.11
SERPINE1	7	1	0	1	2	1.11
RARB	6	1	1	0	2	1.1
ETV5	4	0	1	1	2	1.09
TRERF1	1	0	1	1	2	1.08
MYC	21	1	0	1	2	1.07
THRA	2	0	1	1	2	1.07
HEY2	1	0	1	1	2	1.06
NFATC2	3	0	1	1	2	1.06

TNFSF12	3	0	1	1	2	1.05
IRF9	1	0	1	1	2	1.03
MECOM	1	0	1	1	2	1.03
FOXO4	1	0	1	1	2	1.02
SREBF1	7	1	0	1	2	1.02
MEIS1	1	0	1	1	2	1.01
TCF7L2	3	0	1	1	2	1.01
KLF11	4	0	1	1	2	1
NR2F6	2	0	1	1	2	1
CEBPD	4	0	1	1	2	0.97
HMGB2	1	0	1	1	2	0.97
PBX1	2	0	1	1	2	0.93
TEAD4	1	0	1	1	2	0.91
KLF13	2	0	1	1	2	0.88
JUND	5	0	1	1	2	0.87
KRAS	1	0	1	1	2	0.83
RPA3	1	0	1	1	2	0.82
MYEF2	1	0	1	1	2	0.8
TFCP2L1	1	0	1	1	2	0.76
TFE3	1	0	1	1	2	0.68
VEGFA	7	1	0	1	2	0.66
TCF3	3	0	1	1	2	0.65
AR	5	0	1	1	2	0.63
ICAM1	6	1	0	1	2	0.62
MMP1	5	1	0	0	1	6.92
HBG1	4	0	0	1	1	5.42
INSM1	2	0	1	0	1	4.93
APOB	4	0	0	1	1	4.73
ADIPOQ	3	0	0	1	1	4.71
SLC2A4	2	0	0	1	1	4.54
PF4	2	0	0	1	1	4.45
ACACB	1	0	1	0	1	3.71
TYRP1	2	0	0	1	1	3.65
ZBTB16	2	0	0	1	1	3.31
ISG15	4	0	0	1	1	2.95
MUC1	4	0	0	1	1	2.77
CDC25A	5	1	0	0	1	2.32
ALDOC	2	0	0	1	1	2.24
ATF3	6	1	0	0	1	2.21
LEF1	4	0	0	1	1	2.06
PIGR	4	0	0	1	1	2.05
ABCB1	3	0	0	1	1	1.94
ERBB2	6	1	0	0	1	1.89
COL1A2	5	1	0	0	1	1.88
CYP11A1	3	0	0	1	1	1.86
BRCA2	4	0	0	1	1	1.81
EPAS1	1	0	1	0	1	1.73
FOXO1	1	0	1	0	1	1.69
APOC2	2	0	0	1	1	1.5
RFX2	1	0	1	0	1	1.46
ZNF219	1	0	1	0	1	1.43
AFP	1	0	1	0	1	1.43
EGFR	8	1	0	0	1	1.4
PPARA	3	0	0	1	1	1.36
HMGA1	3	0	0	1	1	1.34
PDGFA	2	0	0	1	1	1.33
BRCA1	5	0	0	1	1	1.24
KLF6	1	0	1	0	1	1.24
HDGF	1	0	1	0	1	1.21
PARP1	1	0	1	0	1	1.2
RARA	3	0	1	0	1	1.13
BCL6	3	0	0	1	1	1.09
HOXD9	2	0	1	0	1	1.09
CCND1	15	1	0	0	1	1.01
MYB	6	1	0	0	1	0.99
CCL5	6	1	0	0	1	0.94
SOX4	1	0	1	0	1	0.89
PGR	5	0	0	1	1	0.87
ZEB1	1	0	1	0	1	0.72
ZNF444	1	0	1	0	1	0.65
FUS	2	0	1	0	1	0.64
NR1D1	3	0	1	0	1	0.62

Table S20: Enriched GO terms (top) and KEGG pathways (bottom line) with adjusted p-values < 0.05 for the hubs in the breast neoplasms network. P-values were adjusted for multiple testing using the BH procedure.

Enriched terms	count	*adj. p-values*
GO:0045944 positive regulation of transcription from RNA polymerase II promoter	27	5.900E-20
GO:0045893 positive regulation of transcription, DNA-templated	16	1.506E-10
GO:0000122 negative regulation of transcription from RNA polymerase II promoter	17	8.232E-10
GO:0006366 transcription from RNA polymerase II promoter	15	1.345E-9
GO:0048661 positive regulation of smooth muscle cell proliferation	6	5.457E-5
GO:0050679 positive regulation of epithelial cell proliferation	6	5.457E-5
GO:0010628 positive regulation of gene expression	8	3.862E-4
GO:0045429 positive regulation of nitric oxide biosynthetic process	5	4.226E-4
GO:0061029 eyelid development in camera-type eye	4	4.089E-4
GO:0042493 response to drug	8	6.854E-4
GO:0006357 regulation of transcription from RNA polymerase II promoter	9	7.250E-4
GO:0048146 positive regulation of fibroblast proliferation	5	6.762E-4
GO:0008284 positive regulation of cell proliferation	9	9.027E-4
GO:0070374 positive regulation of ERK1 and ERK2 cascade	6	0.004
GO:0030335 positive regulation of cell migration	6	0.004
GO:0051091 positive regulation of sequence-specific DNA binding transcription factor activity	5	0.006
GO:0001541 ovarian follicle development	4	0.007
GO:0051591 response to cAMP	4	0.009
GO:0034097 response to cytokine	4	0.013
GO:0008285 negative regulation of cell proliferation	7	0.015
GO:0043406 positive regulation of MAP kinase activity	4	0.017
GO:0006367 transcription initiation from RNA polymerase II promoter	5	0.020
GO:0060337 type I interferon signaling pathway	4	0.019
GO:0071347 cellular response to interleukin-1	4	0.025
GO:0035458 cellular response to interferon-beta	3	0.024
GO:0001666 response to hypoxia	5	0.026
GO:0060749 mammary gland alveolus development	3	0.025
GO:0007623 circadian rhythm	4	0.025
GO:0030324 lung development	4	0.025
GO:0007596 blood coagulation	5	0.029
GO:1902895 positive regulation of pri-miRNA transcription from RNA polymerase II promoter	3	0.031
GO:0045892 negative regulation of transcription, DNA-templated	7	0.032
GO:0046427 positive regulation of JAK-STAT cascade	3	0.035
GO:0007165 signal transduction	10	0.037
GO:0006351 transcription, DNA-templated	13	0.039
GO:0002053 positive regulation of mesenchymal cell proliferation	3	0.045
hsa05200:Pathways in cancer	16	8.489E-8
hsa05166:HTLV-I infection	13	3.355E-7
hsa05219:Bladder cancer	7	2.695E-6
hsa04917:Prolactin signaling pathway	8	2.475E-6
hsa05161:Hepatitis B	9	2.010E-5
hsa05323:Rheumatoid arthritis	7	1.346E-4
hsa05212:Pancreatic cancer	6	4.0173E-4
hsa05221:Acute myeloid leukemia	5	0.003
hsa05223:Non-small cell lung cancer	5	0.003
hsa05142:Chagas disease (American trypanosomiasis)	6	0.002
hsa04620:Toll-like receptor signaling pathway	6	0.002
hsa04151:PI3K-Akt signaling pathway	9	0.004
hsa05205:Proteoglycans in cancer	7	0.006
hsa05206:MicroRNAs in cancer	8	0.006
hsa04630:Jak-STAT signaling pathway	6	0.008
hsa04012:ErbB signaling pathway	5	0.009
hsa04066:HIF-1 signaling pathway	5	0.013
hsa04915:Estrogen signaling pathway	5	0.013
hsa05164:Influenza A	6	0.015
hsa04668:TNF signaling pathway	5	0.015
hsa05213:Endometrial cancer	4	0.015
hsa05168:Herpes simplex infection	6	0.016
hsa04110:Cell cycle	5	0.023
hsa05210:Colorectal cancer	4	0.022
hsa04510:Focal adhesion	6	0.023
hsa05230:Central carbon metabolism in cancer	4	0.022
hsa04380:Osteoclast differentiation	5	0.023
hsa05162:Measles	5	0.024
hsa05160:Hepatitis C	5	0.024
hsa05218:Melanoma	4	0.027
hsa05220:Chronic myeloid leukemia	4	0.027
hsa05133:Pertussis	4	0.029
hsa04060:Cytokine-cytokine receptor interaction	6	0.028
hsa04932:Non-alcoholic fatty liver disease (NAFLD)	5	0.031
hsa05222:Small cell lung cancer	4	0.037
hsa05215:Prostate cancer	4	0.040
hsa05202:Transcriptional misregulation in cancer	5	0.041
hsa05143:African trypanosomiasis	3	0.043
hsa05020:Prion diseases	3	0.043

Table S21: Enriched GO terms (top) and KEGG pathways (bottom line) with adjusted p-values < 0.05 for the MDS in the breast neoplasms network. P-values were adjusted for multiple testing using the BH procedure.

Enriched terms	count	*adj. p-values*
GO:0045944 positive regulation of transcription from RNA polymerase II promoter	60	7.919E-48
GO:0006366 transcription from RNA polymerase II promoter	39	3.038E-31
GO:0000122 negative regulation of transcription from RNA polymerase II promoter	38	1.036E-24
GO:0045893 positive regulation of transcription, DNA-templated	33	1.777E-23
GO:0006351 transcription, DNA-templated	44	3.246E-15
GO:0006357 regulation of transcription from RNA polymerase II promoter	22	2.381E-12
GO:0045892 negative regulation of transcription, DNA-templated	22	2.294E-11
GO:0006367 transcription initiation from RNA polymerase II promoter	14	2.058E-10
GO:0043401 steroid hormone mediated signaling pathway	10	2.195E-9
GO:0030522 intracellular receptor signaling pathway	8	1.236E-7
GO:0006355 regulation of transcription, DNA-templated	29	2.346E-7
GO:0008285 negative regulation of cell proliferation	16	2.543E-7
GO:0045597 positive regulation of cell differentiation	7	3.505E-6
GO:0034097 response to cytokine	7	2.672E-5
GO:0045666 positive regulation of neuron differentiation	7	2.814E-4
GO:0045669 positive regulation of osteoblast differentiation	6	0.001
GO:0001938 positive regulation of endothelial cell proliferation	6	0.002
GO:0042493 response to drug	10	0.002
GO:0048469 cell maturation	5	0.002
GO:0003215 cardiac right ventricle morphogenesis	4	0.002
GO:0035855 megakaryocyte development	4	0.004
GO:0032870 cellular response to hormone stimulus	5	0.004
GO:0051591 response to cAMP	5	0.005
GO:0071277 cellular response to calcium ion	5	0.007
GO:0009612 response to mechanical stimulus	5	0.012
GO:0060337 type I interferon signaling pathway	5	0.016
GO:0042127 regulation of cell proliferation	7	0.019
GO:0051726 regulation of cell cycle	6	0.021
GO:0071773 cellular response to BMP stimulus	4	0.020
GO:0045444 fat cell differentiation	5	0.023
GO:0045647 negative regulation of erythrocyte differentiation	3	0.035
GO:0043065 positive regulation of apoptotic process	8	0.039
GO:0033148 positive regulation of intracellular estrogen receptor signaling pathway	3	0.041
GO:0030218 erythrocyte differentiation	4	0.041
GO:0048646 anatomical structure formation involved in morphogenesis	3	0.047
GO:0010941 regulation of cell death	3	0.047
GO:0008584 male gonad development	5	0.049
hsa05202:Transcriptional misregulation in cancer	13	1.158E-6
hsa05200:Pathways in cancer	16	3.003E-5
hsa04917:Prolactin signaling pathway	8	7.036E-5
hsa05161:Hepatitis B	9	7.276E-4
hsa05166:HTLV-I infection	11	0.001
hsa05221:Acute myeloid leukemia	6	0.001
hsa04380:Osteoclast differentiation	8	0.001
hsa04919:Thyroid hormone signaling pathway	7	0.005
hsa05220:Chronic myeloid leukemia	5	0.039

Table S22: Enriched GO terms (top) and KEGG pathways (bottom line) with adjusted p-values < 0.05 for the MCDS in the breast neoplasms network. P-values were adjusted for multiple testing using the BH procedure.

Enriched terms	count	*adj. p-values*
GO:0045944 positive regulation of transcription from RNA polymerase II promoter	66	5.561E-50
GO:0006366 transcription from RNA polymerase II promoter	40	3.410E-29
GO:0045893 positive regulation of transcription, DNA-templated	36	2.942E-24
GO:0000122 negative regulation of transcription from RNA polymerase II promoter	35	1.772E-18
GO:0006357 regulation of transcription from RNA polymerase II promoter	22	1.264E-10
GO:0045892 negative regulation of transcription, DNA-templated	23	1.268E-10
GO:0006367 transcription initiation from RNA polymerase II promoter	14	2.797E-9
GO:0043401 steroid hormone mediated signaling pathway	10	1.531E-8
GO:0030522 intracellular receptor signaling pathway	8	6.067E-7
GO:0006351 transcription, DNA-templated	36	7.362E-7
GO:0042127 regulation of cell proliferation	12	3.413E-6
GO:0042493 response to drug	14	7.937E-6
GO:0010628 positive regulation of gene expression	13	1.161E-5
GO:0008285 negative regulation of cell proliferation	15	2.143E-5
GO:0006355 regulation of transcription, DNA-templated	28	4.631E-5
GO:0045669 positive regulation of osteoblast differentiation	7	2.024E-4
GO:0035855 megakaryocyte development	5	2.355E-4
GO:0060337 type I interferon signaling pathway	7	2.640E-4
GO:0045597 positive regulation of cell differentiation	6	2.768E-4
GO:0001938 positive regulation of endothelial cell proliferation	7	3.703E-4
GO:0032870 cellular response to hormone stimulus	6	6.743E-4
GO:0051591 response to cAMP	6	7.186E-4
GO:0034097 response to cytokine	6	0.001
GO:0008284 positive regulation of cell proliferation	13	0.002
GO:0042593 glucose homeostasis	7	0.002
GO:0051091 positive regulation of sequence-specific DNA binding transcription factor activity	7	0.003
GO:0071356 cellular response to tumor necrosis factor	7	0.003
GO:0048469 cell maturation	5	0.004
GO:0001666 response to hypoxia	8	0.005
GO:0035162 embryonic hemopoiesis	4	0.006
GO:0006978 DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator	4	0.006
GO:0060749 mammary gland alveolus development	4	0.007
GO:0001701 in utero embryonic development	8	0.008
GO:0032332 positive regulation of chondrocyte differentiation	4	0.010
GO:0030318 melanocyte differentiation	4	0.011
GO:0008584 male gonad development	6	0.014
GO:0048146 positive regulation of fibroblast proliferation	5	0.015
GO:0045648 positive regulation of erythrocyte differentiation	4	0.018
GO:0030878 thyroid gland development	4	0.020
GO:0048589 developmental growth	4	0.020
GO:0030097 hemopoiesis	5	0.020
GO:0009612 response to mechanical stimulus	5	0.020
GO:0002053 positive regulation of mesenchymal cell proliferation	4	0.021
GO:0032496 response to lipopolysaccharide	7	0.022
GO:0045931 positive regulation of mitotic cell cycle	4	0.026
GO:0001569 patterning of blood vessels	4	0.026
GO:0071773 cellular response to BMP stimulus	4	0.031
GO:0030855 epithelial cell differentiation	5	0.034
GO:0060333 interferon-gamma-mediated signaling pathway	5	0.035
GO:0007596 blood coagulation	7	0.036
GO:0051726 regulation of cell cycle	6	0.036
GO:1902042 negative regulation of extrinsic apoptotic signaling pathway via death domain receptors	4	0.036
GO:0045444 fat cell differentiation	5	0.036
GO:0007623 circadian rhythm	5	0.039
GO:0030509 BMP signaling pathway	5	0.040
GO:0045647 negative regulation of erythrocyte differentiation	3	0.040
GO:0030501 positive regulation of bone mineralization	4	0.039
GO:0032869 cellular response to insulin stimulus	5	0.040
GO:0045666 positive regulation of neuron differentiation	5	0.041
hsa05202:Transcriptional misregulation in cancer	16	5.609E-8
hsa05200:Pathways in cancer	20	3.179E-6
hsa05221:Acute myeloid leukemia	9	8.284E-6
hsa05161:Hepatitis B	12	2.050E-5
hsa05166:HTLV-I infection	15	2.333E-5
hsa04917:Prolactin signaling pathway	9	2.708E-5
hsa04380:Osteoclast differentiation	9	0.002
hsa05210:Colorectal cancer	6	0.011
hsa05160:Hepatitis C	8	0.011
hsa05220:Chronic myeloid leukemia	6	0.018
hsa04919:Thyroid hormone signaling pathway	7	0.023
hsa05215:Prostate cancer	6	0.036
hsa05162:Measles	7	0.041
hsa05216:Thyroid cancer	4	0.042

Table S23: Specifications of disease-specific networks for the LIHC dataset.

method	# nodes	# edges	# hubs	# MDS	# MCDS
DESeq	163	199	17	37	35
edgeR	454	579	46	87	98
voom	483	608	49	94	105
VST	475	586	48	93	99

Table S24: Consistent hub genes and miRNAs for the LIHC dataset.

hsa-let-7b, JUN, E2F1, FOS, MYC, CCND1, ESR1, TERT, STAT3, NFE2L2, HBB, APOH, MIER1

Table S25: Consistent MDS genes and miRNAs for the LIHC dataset.

NFE2L2, NME2, MAZ, MYCN, JUN, NR1H4, KCNIP3, NR4A1, TCF3, FOS, ETV4, ESR1, CREM, CNBP, FOXM1, hsa-let-7b, STAT3, USF1, LEF1, SREBF2, HIVEP1, MYC, JUND, CEBPD, ETS2, KLF6, AR, E2F1

Table S26: Consistent MCDS genes and miRNAs for the LIHC dataset.

LEF1, MAZ, CREM, FOXM1, ETS2, MYC, HIVEP1, E2F1, hsa-let-7b, EGR1, JUN, RRM2, JUND, KCNIP3, CNBP, STAT3, NME2, FOS, ETV4, ESR1, USF1, NR4A1, TCF3, NFE2L2

Table S27: Specifications of disease-specific networks for the BRCA dataset.

method	# nodes	# edges	# hubs	# MDS	# MCDS
DESeq	227	302	23	64	70
edgeR	864	1185	87	145	173
voom	756	1065	76	144	169
VST	851	1199	86	147	168

Table S28: Enriched GO terms (top) and KEGG pathways (bottom line) with adjusted p-values < 0.05 for the conserved method in breast cancer disease. P-values were adjusted for multiple testing using the BH procedure.

Enriched terms	count	*adj. p-values*
GO:0042771 intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator	4	7.463E-4
GO:0045944 positive regulation of transcription from RNA polymerase II promoter	8	6.630E-4
GO:0045893 positive regulation of transcription, DNA-templated	6	0.004
GO:0010165 response to X-ray	3	0.014
GO:0006366 transcription from RNA polymerase II promoter	5	0.044
GO:0042981 regulation of apoptotic process	4	0.039
hsa05212:Pancreatic cancer	6	2.201E-6
hsa05220:Chronic myeloid leukemia	6	1.855E-6
hsa05200:Pathways in cancer	8	1.301E-5
hsa05161:Hepatitis B	6	3.130E-5
hsa05166:HTLV-I infection	6	4.079E-4
hsa04110:Cell cycle	5	3.893E-4
hsa05210:Colorectal cancer	4	0.001

Table S29: Enriched GO terms (top) and KEGG pathways (bottom line) with adjusted p-values < 0.05 for the non-conserved method in breast cancer disease. P-values were adjusted for multiple testing using the BH procedure.

Enriched terms	count	*adj. p-values*
GO:0045892 negative regulation of transcription, DNA-templated	15	0.001
GO:0051092 positive regulation of NF-kappaB transcription factor activity	8	0.006
GO:0008340 determination of adult lifespan	4	0.018
GO:0006468 protein phosphorylation	12	0.018
GO:0045944 positive regulation of transcription from RNA polymerase II promoter	17	0.034
GO:0045893 positive regulation of transcription, DNA-templated	12	0.035
GO:0000122 negative regulation of transcription from RNA polymerase II promoter	14	0.040
GO:0016032 viral process	9	0.046
hsa05200:Pathways in cancer	14	2.495E-4
hsa05212:Pancreatic cancer	7	3.700E-4
hsa05166:HTLV-I infection	11	3.851E-4
hsa05210:Colorectal cancer	6	0.002
hsa05220:Chronic myeloid leukemia	6	0.003
hsa05145:Toxoplasmosis	6	0.031
hsa04110:Cell cycle	6	0.033
hsa05162:Measles	6	0.039
hsa04350:TGF-beta signaling pathway	5	0.042
hsa05222:Small cell lung cancer	5	0.040
hsa04064:NF-kappa B signaling pathway	5	0.040
hsa05215:Prostate cancer	5	0.038
hsa05161:Hepatitis B	6	0.035

Abbreviations:

- GRN: Gene Regulatory Network
- TF: Transcription Factor
- DE: Differentially Expressed
- MDS: Minimum Dominating Set
- MCDS: Minimum Connected Dominating Set
- LCC: Largest Connected Component
- LSCC: Largest Strongly Connected Component
- VST: Variance-Stabilizing Transformation
- GO: Gene Ontology
- ESC: Embryonic Stem Cell
- BH: Benjamini-Hochberg
- ILP: Integer Linear Programming
- DAVID: The Database for Annotation, Visualization and Integrated Discovery
- ORA: Over Representation Analysis
- HMDD: Human miRNA Disease Database
- DisGeNET: A Database for Gene-Disease Association
- FFL: Feed Forward Loop
- IPSC: Induced Pluripotent Stem Cell

List of Achievements of this dissertation including Publications, Contributions and Manuscripts:

1. Hamed, H., Spaniol, C., Zapp, A., Helms, V. (2015) Integrative network-based approach identifies key genetic elements in breast invasive carcinoma. BMC Genomics.

2. Hamed, M., Spaniol, C., Nazarieh, M., Helms, V. (2015) TFmiR: a web server for constructing and analyzing disease-specific transcription factor and miRNA co- regulatory networks. Nucleic Acids Res.
   Web server available at: http://service.bioinformatik.uni-saarland.de/tfmir/

3. Nazarieh, M., Wiese, A., Will, T., Hamed, M., Helms, V. (2016) Identification of key player genes in gene regulatory networks. I presented the abstract (talk) at 2nd International Annual Conference of the German Stem Cell Network (GSCN 2014), published in BMC Systems Biology.
   Software available at: http://apps.cytoscape.org/apps/mcds. Code Availability: https://github.com/maryamNazarieh

4. Sadegh, S., Nazarieh, M., Spaniol, C., Helms, V. (2017) Randomization Strategies Affect Motif Significance Analysis in TF-miRNA-Gene Regulatory Networks. Proceedings of 13th International Symposium on Integrative Bioinformatics, Journal of Integrative Bioinformatics.

5. Will, T., Helms, V. (2017) Rewiring of the inferred protein interactome during blood development studied with the tool PPICompare. BMC Systems Biology.

6. Nazarieh, M, Helms, V. (2018). Topology Consistency of Disease-specific Networks. (shortly to be submitted).

7. Nazarieh, M. Helms, V. (2018). Candidate Disease Gene Prioritization based on Topological Features. (shortly to be submitted).

8. Nazarieh, M., Hamed, H., Spaniol, C., Will, T., Helms, V. (2018) TFmiR2: Constructing and analyzing disease-,tissue- and process-specific transcription factor and miRNA co-regulatory networks. I presented the abstract (talk) at ECCB conference and received travel fellowship award. (shortly to be submitted).
   Web server available at: http://service.bioinformatik.uni-saarland.de/tfmir2/