TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl M12 (Mathematische Modellierung biologischer Systeme)

Integration of multiple omics levels for the analysis of adipocyte differentiation

Steffen Martin Sass

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Univ.-Prof. Dr. M. Hrabě de Angelis

Prüfer der Dissertation:

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

Die Dissertation wurde am 12.08.2014 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 09.02.2015 angenommen.

Danksagung

Zu allererst gilt mein ganz besonderer Dank natürlich Nikola. Dafür, dass du mir immer mit Rat und Tat zur Seite gestanden bist und auch jederzeit bei allen Problemchen ein offenes Ohr hattest.

Ebenso herzlich möchte ich mich bei Fabian bedanken. Vielen Dank, dass du mir die Möglichkeit gegeben hast in deiner Gruppe als Doktorand zu arbeiten. Es war toll, dass ich stets auf deine Unterstützung bauen konnte.

Weiterhin danke ich Hans-Werner Mewes für die Zweitbetreuung meiner Doktorabeit sowie Martin Hrabě de Angelis für den Vorsitz der Prüfungskommission.

Mein Dank geht auch an das ganze ICB für die tolle Atmosphäre und speziell natürlich an das ReNe team. Danke auch an das CoPhe team für die Bereicherung der team meetings.

Bedanken möchte ich mich auch ganz besonders bei Flo. Nicht nur für die tolle Zusammenarbeit, sondern auch dafür, dass du auch sonst stets gute Tipps parat hattest. Mein ganz spezieller Dank geht auch an Ivan, Michl und Dom für eure Unterstützung bei diversen Projekten.

Vielen Dank auch an alle Kollaborationspartner innerhalb und außerhalb vom HMGU. Hierbei gilt mein besonderer Dank Johannes Beckers für die wertvollen Ratschläge im Rahmen meines Thesis Committees.

Dankeschön auch an Jan und Michi. Auch wenn unsere langjährige Bürogemeinschaft bereits an akutem Zerfall leidet, wird sie mir immer in allerbester Erinnerung bleiben.

An dieser Stelle möchte ich auch auf besondere Weise meinen Eltern, meiner Oma und auch Sabine danken. Ohne eure unermüdliche Unterstützung wäre dieser Weg nicht möglich gewesen.

Abstract

Omics studies allow for a system-wide characterization of molecules on genomic, transcriptomic, proteomic or metabolomic level. Each of these levels only partly explain the molecular mechanisms that underlie these observations. Hence, the interplay between these levels is of key interest to better understand complex processes. The joint analysis of several omics levels enables us to capture the molecular changes more comprehensively. Typically, statistical and functional analysis of omics data only deals with a single molecular level. The integration of omics data from different molecular levels is, however, not straight-forward. To address this issue, we propose novel methods for the joint analysis of omics data across different levels. Our goal is to identify functional and regulatory properties of the molecular interplay between omics levels.

The understanding of interactions across different omics levels provides broader insights into the mechanisms of complex diseases such as type 2 diabetes mellitus. This disease becomes more and clinically relevant due to its sharp rise in prevalence. In particular the process of adipogenesis is an important factor of diabetes onset as disorders of adipocyte differentiation may have strong impact on the insulin homeostasis. But even though the process of adipogenesis has been widely studied, the complex interplay of molecular mechanisms across different levels is still not elucidated. We therefore propose methods for integrating microRNA and mRNA expression as well as DNA methylation data to give novel insights into the system-wide molecular properties of adipogenesis.

We initially focus on the functional role of microRNAs by combining microRNA with gene expression data to identify regulatory relationships. We show that microRNAs with functional similarities also tend to be co-expressed and thus contribute to the research on microRNA regulation. Based on these findings, we provide a biologically-driven method, called miRlastic, which uses multiple linear regression with elastic net penalty to identify putative miRNA-target relationships. We validate our approach on synthetic and experimental data and show that it outperforms related methods. We use the resulting regulatory network to determine biological functions of microRNAs by performing a local gene set enrichment. By assigning locally overrepresented cellular processes to the corresponding microRNAs, we can assess functional roles to single

microRNAs and sets of cooperating microRNAs in adipogenesis.

Finally, we extend the basic idea of integrative data analysis towards a modular framework for functional analyses on multiple molecular levels. We are the first to introduce a model-based enrichment analysis, called MONA, for joint analysis of multiple omics levels. We show that our multilevel approach provides better insights into processes, which play a role in adipogenesis. We implemented a web application to make this approach available to applied researchers, which is easy to use and provides an enhanced output.

In summary, the analysis of multi-level omics data using our novel methods allows us to determine interactions across DNA methylation, mRNA expression and microRNA expression in adipocyte differentiation and to characterize them with regard to their functional properties.

Zusammenfassung

Systemweite Studien, wie beispielsweise Genomik, Transkriptomik oder Proteomik, erlauben es uns Moleküle in großem Maßstab auf unterschiedlichen Ebenen zu charakterisieren. Jedes dieser Ebenen kann jedoch nur teilweise die molekularen Mechanismen erklären, die den beobachteten Daten zugrundeliegen. Das Zusammenspiel dieser Ebenen ist daher von zentralem Interesse, um komplexe Prozesse besser verstehen zu können. Die gemeinsame Analyse von mehreren Omik-Ebenen bietet uns die Möglichkeit, die molekularen Veränderungen umfassender nachzuvollziehen. Üblicherweise befassen sich statistische und funktionelle Analysen lediglich mit einer einzelnen molekularen Ebene, wohingegen die Integration von Omik-Daten aus unterschiedlichen molekularen Ebenen sehr komplex ist. Um diesen Ansatz zu realisieren, führen wir neue Methoden ein für die gemeinsame Analyse von Omik-Daten über verschiedene Ebenen. Unser Ziel hierbei ist die Identifikation von funktionellen und regulatorischen Eigenschaften des molekularen Zusammenspiels zwischen unterschiedlichen Omik-Ebenen.

Das Verständnis von Interatktionen über verschiedene Omik-Ebenen kann bessere Einblicke in die Mechanismen von komplexen Krankheiten wie Typ 2 Diabetes bieten. Aufgrund der weltweit dramatisch steigenden Häufigkeit dieser Krankheit, hat sie sich zu einem zentralen Forschungsthema entwickelt. Speziell der Adipogenese-Prozess ist ein wichtiger Faktor für den Ausbruch von Typ 2 Diabetes, da Störungen im Verlauf der Adipozytendifferenzierung großen Einfluss auf den Insulihaushalt haben können. Obwohl der Adipogenese-Prozess bereits intensiv untersucht wurde, ist das komplexe Zusammenspiel der molekularen Mechanismen über verschiedene Ebenen nach wie vor nicht aufgeklärt. Daher wenden wir unsere Methoden zur Multi-Omik Datenintegration auf einem Datensatz an, der sowohl aus microRNA- und mRNA-Expressionsdaten als auch aus DNA Methylierungsdaten besteht, um neue Einblicke in die systemweiten molekularen Eigenschaften der Adipogenese zu gewinnen.

Wir beschäftigen uns zunächst mit der funktionellen Rolle von microRNAs durch die Kombination von microRNA- mit Genexpressionsdaten, um regulatorische Beziehungen zu identifizieren. Wir können zeigen, dass funktionell ähnliche microRNAs dazu neigen, auch ko-exprimiert zu sein. Hierdurch tragen wir zu der Erforschung von microRNA-Regulation bei. Basierend auf diesen Er-

kenntnissen präsentieren wir eine Methode, genannt miRlastic, die multiple Regression mit Elastic-Net Penalisierung verwendet um mögliche miRNA-target-Beziehungen zu identifizieren. Wir validieren unseren Ansatz auf synthetischen und experimentellen Daten und können zeigen, dass er im Vergleich zu verwandten Methoden bessere Ergebnisse liefern kann. Wir verwenden das mit dieser Methode generierte regulatorische Netzwerk um durch die Anwendung eines lokalen Genset-Enrichments biologische Funktionen von microRNAs zu bestimmen. Durch das Zuweisen von lokal überrepräsentierten zellulären Prozessen zu den entsprechenden microRNAs können wir funktionelle Rollen einzelner oder mehrerer microRNAs beurteilen.

Die grundlegende Idee der integrativen Datenanalyse erweitern wir schließlich zu einer modularen Methode für die funktionelle Analyse von multiplen molekularen Ebenen. Wir sind hierbei die ersten, die eine modellbasierte Enrichmentmethode für die gemeinsame Analyse von multiplen Omik-Ebenen präsentieren. Wir können zeigen, dass unser multi-Ebenen-Ansatz einen besseren Einblick in Prozesse bietet, die eine Rolle in der Adipogenese spielen. Durch die Implementierung in Form einer Webapplikation machen wir diese Methode anwendungsorientierten Wissenschaftlern zugänglich. Diese Applikation ist einfach zu bedienen und bietet eine umfassende Aufbereitung der Ergebnisse.

Zusammenfassend können wir durch die Analyse von Omik-Daten multipler Ebenen unter Verwendung unsere neuen Methoden Interaktionen zwischen DNA Methylierung, mRNA-Expression und microRNA-Expression in der Adipozytendifferenzierung bestimmen und im Hinblick auf ihre funktionellen Eigenschaften charakterisieren.

Scientific Publications

The results of this thesis are partly based on previously published papers and papers, which are currently within the publication process. These and further publications are listed below:

- **S. Sass***, S. Dietmann*, U. C. Burk, S. Brabletz, D. Lutter, A. Kowarsch, K. F. Mayer, T. Brabletz, A. Ruepp, F. J. Theis, and Y. Wang. MicroRNAs coordinately regulate protein complexes. *BMC Systems Biology*, 5(1):136, August 2011.
- **S. Sass***, F. Buettner*, N. S. Mueller, and F. J. Theis. A modular framework for gene set analysis integrating multilevel omics data. *Nucleic Acids Res*, 41(21):9622–9633, Nov 2013.
- S. U. Meyer, K. Stoecker, **S. Sass**, F. J. Theis, and M. W. Pfaffl. Posttranscriptional regulatory networks: from expression profiling to integrative analysis of mRNA and microRNA data. *Methods Mol Biol*, 1160:165–188, 2014.
- **S. Sass**, F. Buettner, N. S. Mueller, and F. J. Theis. RAMONA: a web application for gene set analysis on multilevel omics data. Under review at *Bioinformatics*.
- K. Roeck, J. Tigges, **S. Sass**, A. Schuetze, A.-M. Florea, A. Fender, F. J. Theis, J. Krutmann, F. Boege, E. Fritsche, G. Reifenberger, and J. W. Fischer. miR-23a promotes dermal aging and cellular senescence by targeting hyaluronan synthase2. Under review at *Journal of Investigative Dermatology*.
- D. M. Waldera-Lupa, A.-M. Florea, F. Kalfalah, **S. Sass**, F. Kruse, J. Tigges, E. Fritsche, J. Krutmann, H. Busch, H. E. Meyer, F. Boege, F. J. Theis, G. Reifenberger, and K. Stuehler. Proteome-wide analysis of primary cultures of in situ aged human fibroblasts reveals a moderate age-associated cellular phenotype. Submitted to *Aging Cell*.
- S. U. Meyer, **S. Sass**, N. S. Mueller, S. Krebs, S. Bauersachs, S. Kaiser, H. Blum, C. Thirion, S. Krause, F. J. Theis, and M. W. Pfaffl. Integrative analysis of microRNA and mRNA data reveals an orchestrated function of microRNAs in skeletal myocyte differentiation in response to TNF- or IGF1. Submitted to *PLOS ONE*.

  $*$ = equal contributions

Contents

Nomenclature

Ago Argonaut, page 44

AIC Akaike information criterion, page 24

BMP bone morphogenetic protein, page 37

C/EBP CCAAT-enhancer-binding protein, page 37

cDNA complementary DNA, page 10

ChIP-seq Chromatin ImmunoPrecipitation DNA-Sequencing, page 69

DAG Directed acyclic graph, page 29

DAG directed acyclic graph, page 30

EP Expectation propagation, page 34

FDR False Discovery Rate, page 20

GO Gene Ontology, page 7

HITS-CLIP high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation, page 47

IGF1 insulin growth factor-1, page 37

KEGG Kyoto Encyclopedia of Genes and Genomes, page 7

KL Kulback-Leibler, page 35

KLF Krüppel-like factor, page 37

KS Kolmogorov-Smirnov, page 18

LARS Least Angle Regression, page 25

lasso least absolute shrinkage and selection operator, page 25

LEA local enrichment analysis, page 75

limma Linear Models for Microarray Data, page 27

MAPK mitogen-activated protein kinase, page 37

MCMC Markov chain Monte Carlo, page 31

MeDIP Methylated DNA immunoprecipitation, page 14

miRNA microRNA, page 3

mRNA messenger RNA, page 3

NGS next-generation sequencing, page 10

PAR-CLIP photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation, page 47

PPARγ per oxisome proliferator-activated receptor γ, page 37

RISC RNA-induced silencing complex, page 44

RMA Robust multi-array average, page 39

RNA-seq RNA sequencing, page 10

rRNA ribosomal RNA, page 9

RUNX2 runt-related transcription factor 2, page 37

RUNX2 runt-related transcription factor 2, page 67

TF Transcription factor, page 69

tRNA transfer RNA, page 9

Chapter 1

Introduction

Since the discovery of regulatory mechanisms on molecular level within prokaryotes in the late 1960s [184], research in molecular Biology has more and more focused on the investigation of complex regulatory interactions [177]. Even in this early stages of research in that field, the importance of the interplay between several molecular levels has already been pointed out. At that time, regulatory mechanisms have been studied on a small, well-defined scale and primarily in prokaryotes, such as the *lac* operon in *Escherichia coli* [11]. The progress in experimental techniques [74] and new insights into molecular and cellular mechanisms [149] enabled the researchers to extend their focus from small distinct regulatory mechanisms to complex molecular interactions in biological systems [150]. Nowadays, research in molecular biology is focusing on the comprehensive characterization of molecular mechanisms also in higher eukaryotes by investigating gene activity and regulatory features in a system-wide fashion [35, 139]. It is supported by modern high-throughput technologies for the large-scale measurement of biological molecules. These are also referred to as "omics studies" in molecular biology. They deal with the characterization of a comprehensive set of molecules, which primarily include DNA, RNA, proteins and metabolites. Omics studies allow us get a broad and unbiased insight into changes of molecular activities, which arise due to certain environmental or cellular conditions. This information can then be used to globally infer affected cellular mechanisms.

The understanding of these system-wide molecular mechanisms is essential

to reveal the causes of complex diseases, which may depend on a multitude of genetic and environmental factors. For example, due to their sharp rise in prevalence, type 2 diabetes mellitus [54] as well as obesity and associated cardiovascular diseases [140] become more and clinically relevant. These diseases have in common that they are directly linked to an excessive accumulation of adipose tissue, which is often accompanied by decreased insulin sensitivity [62]. The adipose tissue comprises adipocytes, which react to insulin with the storage of lipids and which can also secret hormones to modulate insulin sensitivity in distant tissues, thereby controlling the energy homeostasis in the whole human body [65]. The disruption of one of these processes, which may arise during the development of these cells, can lead to metabolic disorders and to a dramatically increased risk for type 2 diabetes [165]. The process of adipocyte differentiation has thus become an important research subject over the past two decades [145]. Several omics studies have been conducted to investigate the adipocyte differentiation process in a system-wide fashion using large-scale molecular profiling techniques [147, 181, 189]. However, the underlying molecular mechanisms of adipogenesis are still not fully understood.

The goal of this thesis is the integration of data from different omics experiments in order to reveal molecular mechanisms that are involved in the differentiation of adipocytes. For this purpose, we introduce novel methods that are specifically designed to account for the molecular characteristics of the respective level.

1.1 Molecular interactions across multiple levels

Changes in cellular or environmental conditions may act as a stimulus for a cell to modulate certain cellular properties and processes, which can influence the genome, transcriptome, proteome or metabolome level (Fig. 1.1).

On genome level, external factors usually do not alter the sequence of the DNA itself, but rather act *epigenetically* [79]. One of the most important epigenetic mechanisms is the methylation of DNA (Fig. 1.1b). The methylation of promoter regions prevents the attachment of transcription factors, thereby modulating the activity of the associated genes [83]. It has been shown that

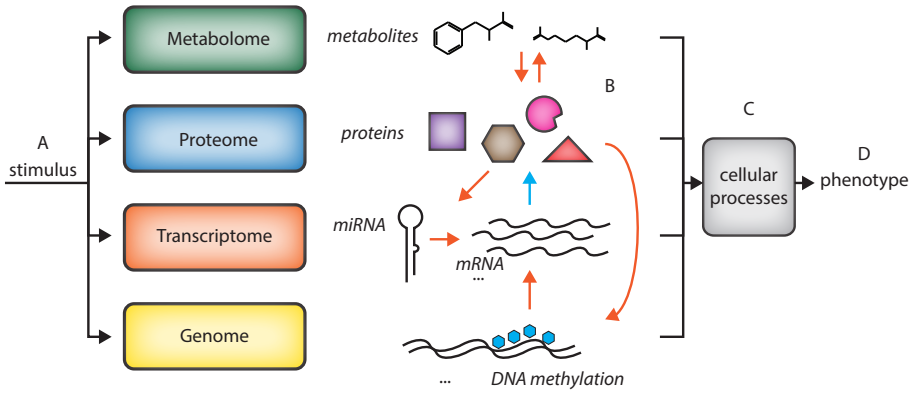


Figure 1.1: An external stimulus can affect several molecular levels (**A**). This leads to changes in the concentration or activity of molecules, which themselves affect other molecular levels via regulation (orange arrows) or translation (blue arrow) (**B**). The multi-level changes as a whole affect certain cellular processes (**C**), which then give rise to an altered phenotype (**D**).

the methylation state of a cell highly depends on environmental factors and is especially altered due to aging or dietary changes [79].

The transcriptome level comprises all RNA molecules, which are produced in a cell. The most-well studied response on transcriptional level is the modulation of *messenger RNA* (mRNA) expression. The mRNA is transcribed from protein-coding genes on the genome and then translated into proteins. It is thus regarded as a proxy for the abundance of proteins, which carry out the biochemical gene function. Another type of RNA molecules are *microRNAs* (miRNAs). miRNAs are very short and bind specifically to a target mRNA, which enables them to post-transcriptionally regulate the gene expression. Since the discovery of miRNAs in the early 90's [98], their important role in influencing biological processes has become more and more obvious [30]. But even though miRNAs have been studied intensively during the last two decades, their regulatory mechanisms are still not fully understood [86].

The proteome level describes the set of all proteins of a system. This level can be regarded as the "functional" level. Proteins can serve as enzymes for modulating the metabolome and metabolites can in turn influence the activity of proteins. The metabolome then provides a readout of the overall underlying molecular interactions

The interplay between all molecular levels contains the full regulatory infor-

mation of the cell, where one level alone might not be sufficient to understand the underlying biological processes. For example, we can not determine whether altered methylation patterns actually affect the expression of associated genes by only taking into account methylation data. Similarly, if we focus only on mRNA expression data, we do not know if the changes directly affect the protein level and to which extent mRNAs are regulated post-transcriptionally. Measurements on miRNA level do not provide any functional information if we do not take their target relationships into account. Equally, the proteome level alone cannot provide satisfactory functional insights, as the technologies for assessing large-scale protein concentrations are still laborious and less comprehensive than measurement techniques for mRNA expression [4]. Finally, metabolite concentrations only indicate the overall readout of cellular mechanisms and can not directly explain regulatory relationships [55]. In addition, all of these omics studies are prone to measurement errors arising from the individual high-throughput techniques. We thus benefit most from the joint analysis of different omics levels as we can combine the advantages of each of them.

1.2 Research questions

MiRNAs are known to fine-tune the expression of specific target genes post-transcriptionally but the mechanisms of target gene regulation are still not fully elucidated. However, to investigate miRNA influences on the adipocyte differentiation process, a clear understanding of miRNA-target relationships is necessary. A variety of target prediction algorithms has been proposed, which basically make use of sequence features. But since these approaches are prone to a large number of false positives [142], several methods have been proposed to include expression data to improve this prediction [153, 111, 122]. We asked whether we can improve these methods by taking into account the properties of miRNA regulation. Another question we want to answer is the specific role of miRNAs in the regulation of cellular processes.

Up to know, bioinformatic methods mainly focus on a single omics level. Hence, these methods are not able to fully capture molecular responses and can only partly reveal affected processes. We therefore asked whether we can make

use of the multilevel omics data to reliably identify molecular responses, which occur during adipogenesis.

1.3 Overview of this thesis

In the following, we will give a brief outline of this thesis. A graphical representation of this outline is given in Figure 1.2.

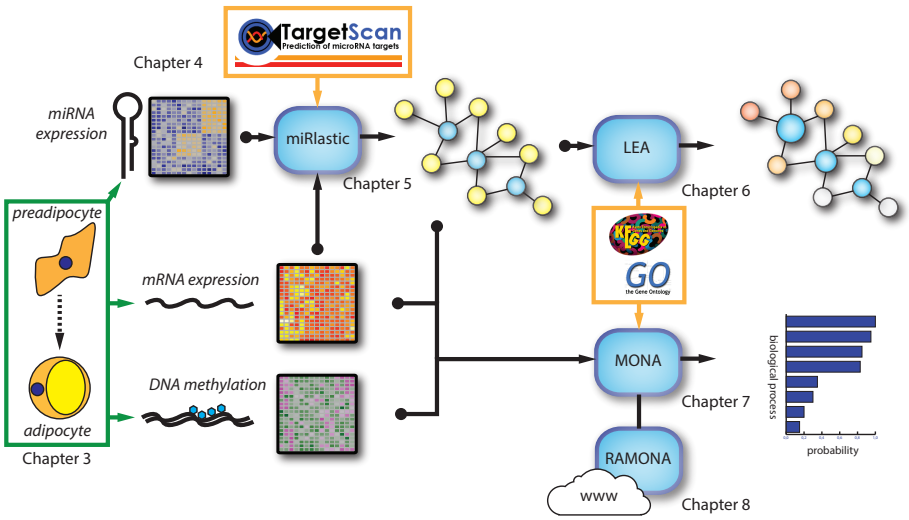


Figure 1.2: Overview of this thesis. In chapter 3, we will introduce the adipocyte differentiation dataset, which is analyzed throughout this thesis. Basic mechanisms of miRNA regulation are described in chapter 4. In chapter 5 and 6, we will integrate the miRNA and mRNA expression data on the basis of miRNA target predictions to generate a miRNA-mRNA regulatory network and to functionally characterize adipogenesis-associated miRNAs. In chapter 7, we will integrate the multiple molecular levels to perform a multilevel ontology analysis. Finally, we will introduce a web application for this analysis in chapter 8.

Chapter 2 introduces the biological and technical background, which is relevant throughout the thesis. This includes an overview of gene expression profiling, omics analysis and statistical methods. In addition, we will introduce the basic molecular mechanisms of adipocyte differentiation.

Chapter 3 introduces the adipogenesis dataset, which is used to investigate the molecular properties of adipocyte differentiation on multiple molecular levels in the subsequent chapters. We will show the results of statistical analyses for determining differentially expressed mRNAs and miRNAs as well as

differentially methylated CpG sites.

Chapter 4 consists of two parts. In the first part, we will introduce the mechanisms of post-transcriptional gene regulation through miRNAs as well as the properties of transcriptional regulation of miRNAs themselves. We will initially point out important steps in the biogenesis of mature miRNAs and introduce important bioinformatic resources, which are essential for the investigation of miRNAs. One of the most important questions in miRNA research is the prediction of miRNA-mRNA interactions to determine functional properties of miRNAs. We will outline common approaches, which deal with this target prediction, especially the *in silico* methods.

In the second part of this chapter, we will introduce our own bioinformatic analyses to show that co-expression among miRNAs, which arises due to the fact that several miRNAs are often under the control of a common promoter, is directly linked to a coordinated regulation of protein complexes.

In **Chapter 5** we will introduce a novel approach for the identification of miRNA-mRNA interactions (*miRlastic*), which is based on combined miRNA-mRNA expression data as well as on *in silico* target predictions. By integrating these different levels of information using a multiple regression approach, we aim to identify a reliable regulatory miRNA-mRNA network, which is specific for the given experimental setup. We will show that co-expression among miRNAs is an important aspect in this analysis of combined expression data. We thus account for the correlation between expression profiles in our method, which we will then use for building up a miRNA-mRNA network that is relevant for the regulatory processes in adipocyte differentiation. Finally, we will show that our method outperforms other common approaches both for simulated and real data.

Chapter 6 introduces a method, which can be applied on the previously generated miRNA-mRNA networks in order to gather functional information about miRNAs. We introduce a *local enrichment analysis* (LEA) in the network for the identification of miRNAs that are supposed to significantly contribute to a given process. We will furthermore use this method to determine processes out of a given set of functional groups, whose associated genes are targeted only by a specific set of miRNAs. By applying LEA to the data-driven miRNA-mRNA

network for adipocyte differentiation, we revealed miRNAs regulating processes and signaling cascades, which play crucial roles in this differentiation procedure. We furthermore observe that several miRNAs act together in regulating these processes even if they are apparently not directly related to each other.

In **chapter 7** we will extend the basic idea of integrative data analysis towards a modular framework for functional analyses on multiple molecular levels. We will introduce a method for *multi-level ontology analyis* (MONA), which integrates data resulting from various omics experiments. Using a model-based enrichment approach, we will summarize the diverse molecular levels to a functional unit, which is regarded as gene response. These gene responses then serve as a basis for the inference of potentially affected biological functions with regard to the experimental setup. The MONA framework can be easily extended to simultaneously account for any possible regulatory mechanism. We will use MONA for the integration of mRNA, methylation and miRNA data from the adipocyte differentiation study in order to determine associated biological processes. We will show that our approach is appropriate for the functional characterization of affected molecular changes during the formation of mature adipocytes.

An implementation of the MONA approach in form of a web application is introduced in **chapter 8**. This web application is called *remotely accessible multi-level ontology analysis* (RAMONA) and provides an easy-to-use interface for performing a model-based enrichment analysis on combined data from different molecular levels. It incorporates a database for processing several common gene identifiers and for mapping them to functional categories from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [84] and Gene Ontology (GO) [5]. RAMONA depicts the results from the underlying MONA algorithm in a comprehensive fashion to provide functional insight into the inferred processes. We will demonstrate the application of RAMONA by showing the multi-level effects of adipocyte differentiation on signaling pathways.

Chapter 2

Background

This chapter introduces the theoretical concepts and experimental techniques, which are used throughout this thesis. First, the experimental approaches section describes technologies for measuring transcriptomic and DNA methylation profiles. In addition, we will clarify the principles of hypothesis testing, correlation and regression analysis, omics analysis and Bayesian networks. Finally, we will briefly outline the basic mechanisms of adipogenesis.

2.1 The omics landscape

Large-scale measurements of biological molecules like DNA, RNA, proteins or metabolites are also commonly referred to as *omics studies*. Several experimental techniques have been developed for measuring these molecules. In the following, we will introduce the basic principles of the experimental techniques that can be used for transcriptional or DNA methylation profiling.

2.1.1 Transcriptomics

Transcriptomics refers to studies dealing with the genome-wide analysis of all RNA molecules produced in a certain organism. These RNA molecules include mRNAs, ribosomal RNAs (rRNAs), transfer RNAs (tRNAs) and all non-coding RNAs. The focus of transcriptomics is the large-scale measurement of RNA abundance in order to determine the global gene activity.

Currently, the most commonly used experimental method for transcriptional profiling are DNA microarrays. The great advantage of microarrays is that they are easy to handle and relatively cheap. However, it is necessary to define a set of transcripts in advance. This does not allow for the identification of unknown transcripts or alternative splice variants.

With the advent of next-generation sequencing (NGS) for RNA expression profiling, which is commonly referred to as RNA sequencing (RNA-seq), these issues could be overcome. Since almost all transcripts in a biological sample are measured, even unknown transcripts can be identified and quantified. Furthermore, the identification of alternative splicing is possible, allowing for a much deeper insight into transcriptional processes. However, the analysis of RNA-seq data is cumbersome and demands a large amount of computational power. In addition, the generation of RNA-seq data is more expensive than microarrays, even though the costs for NGS are constantly decreasing [115]. In addition, we have to note that coverage for identification as well as for quantification is not complete, especially for low abundant transcripts. Also the assembly of transcripts from sequenced reads and expression estimation is not yet standardized.

Transcriptional profiling using microarrays

The basic principle of a microarray experiment is to combine a set of DNA spots on a chip (probes) with the target sequences from the sample of interest by DNA strand hybridizations (Fig. 2.1). The probe sequences are specifically designed to serve as reporters for transcripts. The isolated RNA from the biological sample is transcribed into complementary DNA (cDNA), which is afterwards labeled with a fluorescent dye. This labeled cDNA is then hybridized with the probes on the microarray. The fluorescence intensity of the spots is measured using a laser scanner and can be directly used to identify relative changes in RNA expression across samples.

Generally, microarrays can be produced in two ways, spotted and *in-situ* synthesized. For spotted microarrays, the probes are prepared before the attachment to the array. This is either done by preparing cDNA from the mRNA or by creating respective oligonucleotides. These DNA sequences are then spotted on a glass slide. The advantage of spotted microarrays is the relatively cheap

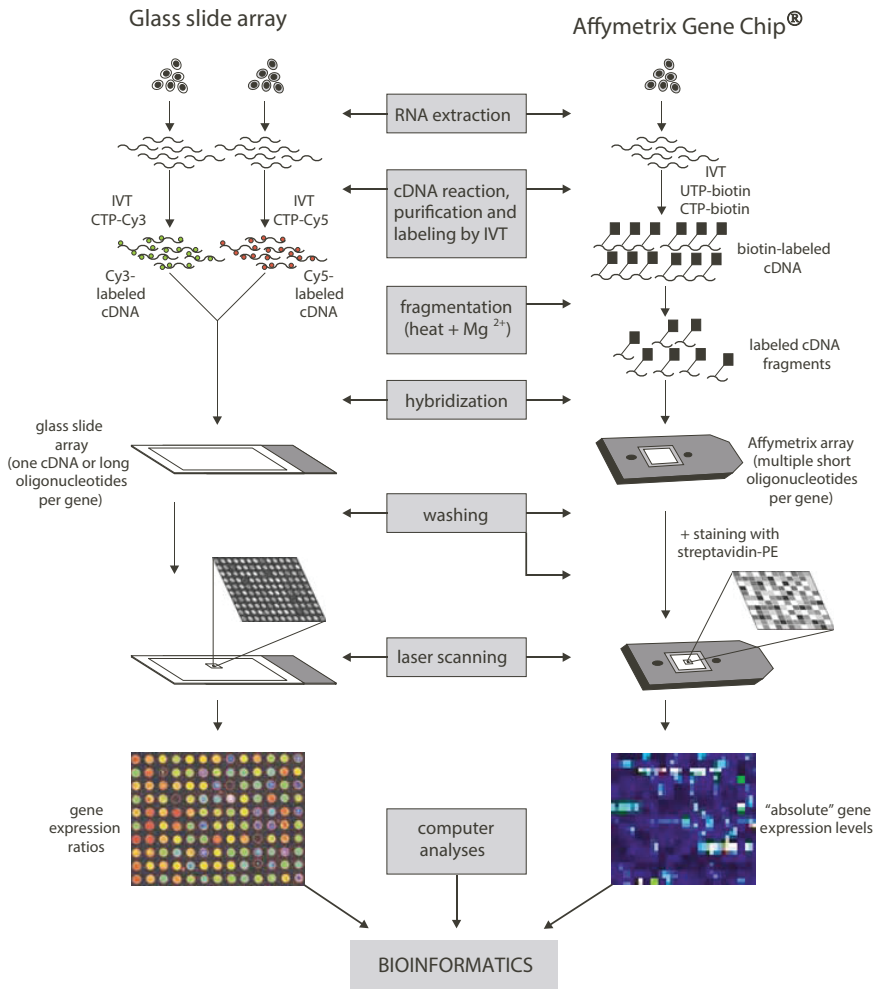


Figure 2.1: Comparison of spotted (left) and oligonucleotide microarrays (Affymetrix, right). For both workflows, the RNA is initially extracted from the cell population. The spotted array allows for two-channel measurements, which may correspond to a control and a diseased cell population. The RNA is transcribed into cDNA, which is then labeled with Cy3 (green) or Cy5 (red) by *in vitro* transcription (IVT). These cDNA fragments are then hybridized with the probes on the array, washed, and finally scanned with a laser to assess fluorescence intensity. In case of the oligonucleotide array, only one channel is available. The IVT step is used to incorporate biotinylated nucleotides, which are then stained with streptavidin after the hybridization. Figure adapted from [164].

and uncomplicated production. In addition, two different dyes may be used to analyze two samples simultaneously. However, these types of microarrays are prone to measurement errors since there exists a large amount of variation aris-

ing from the spotting procedure [163].

Nowadays, *in-situ* synthesized microarrays are usually preferred over spotted microarrays, as they yield more reproducible results [7]. The main difference between spotted and *in-situ* synthesized microarrays is that in the latter case oligonucleotides are synthesized directly on the array. These kind of microarrays are therefore also referred to as *oligonucleotide microarrays*, even though there exist spotted microarrays based on oligonucleotides, too. The most prominent procedure for the production of oligonucleotide microarrays is similar to the production of electronic chips where a photolithographic process is used to synthesize the oligonucleotides on the array [132] (Fig. 2.2). The surface of the array is covered with photolabile protecting groups, which prevent the attachment of nucleotides. By using a photolithographic mask, these groups can be specifically removed by light exposure. Special nucleotides, which are again protected by photolabile groups, are then presented to the surface. These nucleotides are then able to bind at the positions on the array, which were exposed to the light. The process is repeated until the desired oligonucleotides for each position are synthesized. This procedure allows for the exact synthesis of predefined probe sequences, which therefore can be specifically designed to prevent cross-hybridization between different transcripts [107].

Affymetrix was the first company to produce oligonucleotide microarrays, which are based on a photolitographic procedure [132]. The Affymetrix "Gene Chips" are still the most popular platform for measuring the transcriptome. The latest Affymetrix chips comprise of about 1 million distinct oligonucleotide probes, which are grouped into about 50,000 *probe sets*. These probe sets cover known transcripts of an organism.

Microarrays are available for different kinds of studies. The most common application of microarrays is for measuring mRNA expression within a cell population or tissue. But special microarray solutions are also available to measure, for example, miRNA expression. These miRNA microarrays cover the complete set of known miRNAs across different organisms. MRNA and miRNA measurement approaches primarily differ in the preparation of the samples. The amplification of the mRNA concentration prior to the microarray analysis typically requires the isolation of poly(A) RNA [107]. In case of mature miRNAs,

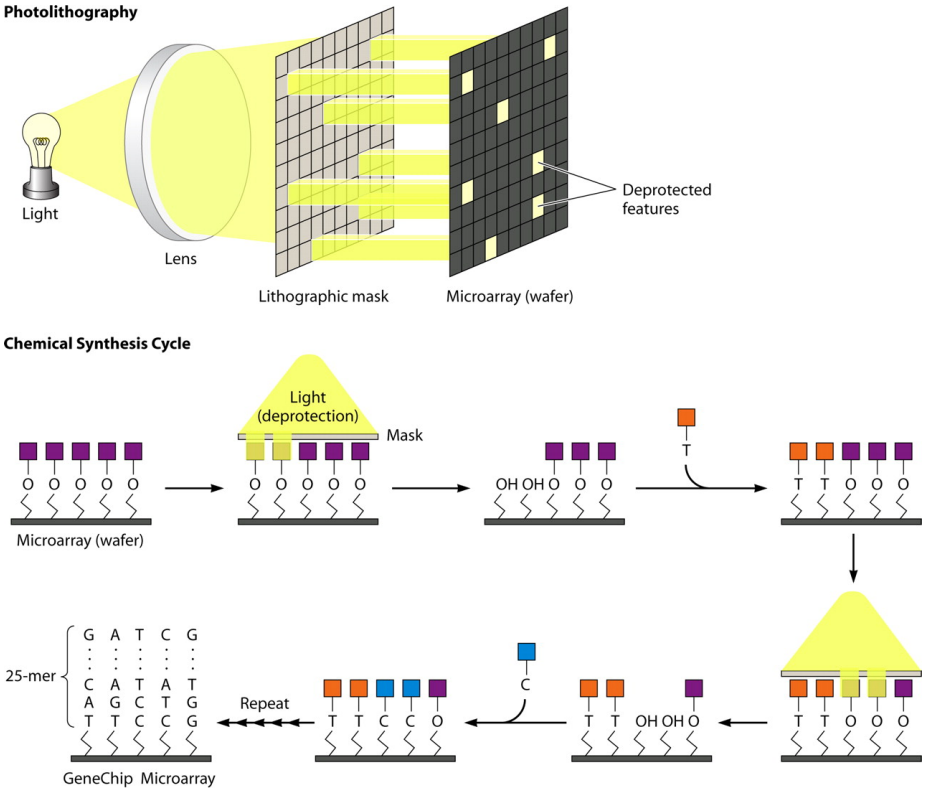


Figure 2.2: Principle of oligonucleotide synthesis using a photolithographic process. Initially, the surface of the microarray is covered with photolabile protecting groups, which can be specifically removed at predefined positions by light exposure using a photolithographic mask. This allows a binding of special nucleotides, which are themselves covered with a photolabile protecting group. By systematically repeating this procedure, the previously designed oligonucleotide sequences can be synthesized. Figure adapted from [118].

the isolation procedure usually aims to select the RNA molecules by their size [93].

2.1.2 Epigenetic gene regulation by DNA methylation

The study of stable and possibly heritable changes in phenotype or gene expression, which do not occur as a result of altered DNA sequence, is referred to as *epigentics* [51]. One of the most important epigenetic mechanisms is DNA methylation whose great importance, especially for cell differentiation, is well established [69].

DNA methylation is carried out through the attachment of a methyl group

on the 5C position of cytosine residues in DNA. This modification is maintained by a set of methyltransferases and predominantly happens in the context of a CpG site [52]. However, the methylation of cytosines may also occur in non-CpG regions [138]. CpG islands are regions on the DNA with a length greater than 500 bp and a GC content equal to or greater than 55% [167], which are usually unmethylated. Since methylated cytosines are prone to mutations, these regions underlie a lower mutagenic pressure as compared to other regions and are therefore well-conserved. However, the methylation of CpG islands in a promoter region can cause a stable inheritable transcriptional silencing by preventing the attachment of transcription factors [83].

It has been shown that stable methylation patterns are established during embryonic development when cells differentiate from embryonic stem cells to specific cell types [79]. However, changes in DNA methylation are also important during the differentiation process of adult stem cells [159, 189]. Furthermore, the methylation state of a cell highly depends on environmental factors [79].

Experimental techniques for large-scale methylation profiling

A variety of approaches have been proposed for the assessment of genome-wide DNA methylation. They differ either in the pretreatment or in the analysis technology [94]. A pretreatment step is necessary, since the methyl groups are removed from the sequences during the amplification of the DNA. Furthermore, methylation itself can not be identified by hybridization. Hence, array-based approaches would not be applicable. The pretreatment techniques include affinity enrichment or sodium bisulphite treatment. Affinity enrichment utilizes antibodies, which are specific for methylated cytosines, in order to separate methylated DNA fragments from the fragmented genomic DNA via immunoprecipitation [121]. This technique is referred to as *methylated DNA immunoprecipitation* (MeDIP) [176]. The treatment of denatured genomic DNA with sodium bisulphite makes use of the fact that deamination, which is caused by this treatment, happens much faster for unmethylated cytosines as compared to methylated ones [174] (Fig. 2.3). This leads to a rapid conversion of unmethylated cytosines into uracil, whereas the methylated ones remain unchanged. The bisulphite treatment allows for base-pair resolution, which is a great advantage

over the affinity enrichment.

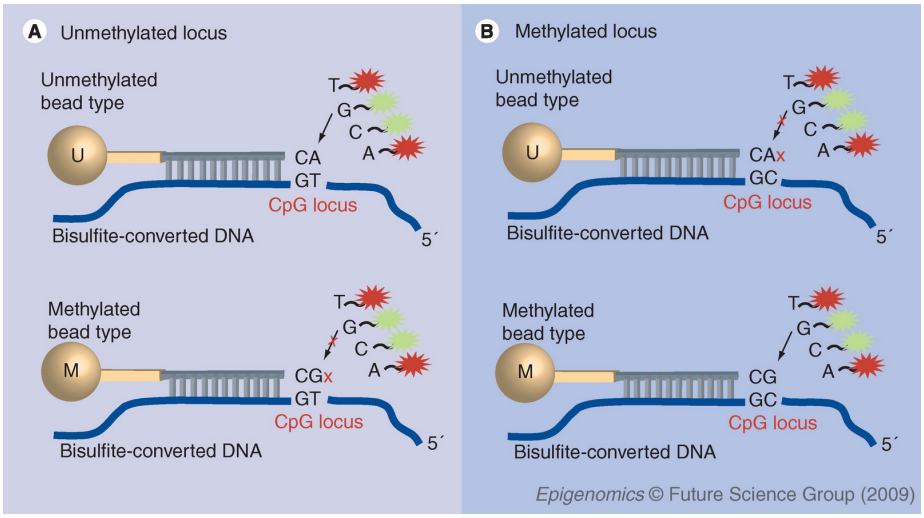


Figure 2.3: Principle of methylation profiling using the Illumina Infinium platform. Each CpG site is represented by two bead types. (**A**) Within a CpG site the bisulfite treatment causes a conversion of unmethylated cytosines (C) to uracil (T). It matches perfectly with a corresponding (U) probe, whereas a mismatch occurs for the (M) probe. The (U) signal will therefore become high and the (M) signal low. (**B**) Methylated CpG sites remain unchanged by the bisulfite treatment. In this case, the (M) probe matches perfectly and the matching of the (U) probe is imperfect. This causes a high signal for the (M) probe and a low signal for the (U) probe. Only perfect matches can be extended and become labeled with a fluorescent dye. Figure adapted from [14].

For both pretreatments, commonly used follow-up quantification techniques are microarray- or NGS-based. These two techniques have the same advantages and disadvantages in methylation analysis as in transcriptome analysis (see above). In case of MeDIP, the array-based variant is called *MeDIP-chip* [176] and the NGS-based variant *MeDIP-seq* [32]. A powerful array-based technology for bisulphite treated DNA is the Illumina Infinium platform [14] (Fig. 2.3). The amplified bisulphite treated DNA is hybridized with methylation-specific probes, which correspond to predefined CpG sites in the genome. The probe for an unmethylated site binds perfectly, if the respective cytosine was replaced by a uracil. Otherwise, the probe for the methylated site binds perfectly. Only perfect matches can be extended by a sequence with a fluorescent label, which enables the detection. A measure for the methylation of a CpG site is the β-value, which is the ratio of the fluorescent signals from the methylated probe to

the total locus intensity. This value can be transformed into *M-values*, which are an appropriate measure for the statistical analysis of methylation [33]. The M-value M_i for a CpG site i can be calculated from its β-value β_i as:

$$M_i = \log_2 \left(\frac{\beta_i}{1 - \beta_i} \right)$$

The *HumanMethylation450* chip is the most recent version of this technology and covers 485,577 methylation sites in the human genome [13]. Bisulphite-converted DNA can also be analyzed by using NGS approaches [105].

2.2 Hypothesis testing

In statistics, hypothesis testing is used to judge whether given phenomena can be rejected with a certain probability or not. Performing such a test involves the formulation of two hypotheses - the null hypothesis H_0 and the alternative hypothesis H_1. H_0 summarizes common events which are observed with high probability and H_1 captures rare events which can be observed only with a low probability. Deciding wether H_0 can be rejected and thus H_1 can be assumed is the goal of the hypothesis test. For this purpose, we have to identify an appropriate *test statistic*, whose distribution can then be used to determine a p-value. A p-value is the probability of observing a value of the test statistic which is equal or more extreme than the test statistic value obtained from the tested data. H_0 can then be rejected if the p-value is below a certain significance level α, which is typically chosen as 0.05 or 0.01. Statistical tests can be performed as *one-tailed* or *two-tailed* tests. If we perform a one-tailed test, we reject H_0 only if the test statistic is lower or higher than a given critical value, but not in both cases. In a two-tailed test, either lower or higher values of the test statistic lead to a rejection of H_0. Note that we have to modify the critical value to obtain the identical area under the tails for one- and two-tailed tests.

We distinguish between *parametric* and *non-parametric tests*. Parametric tests are used in cases when a specific probability distribution of the tested observations can be assumed, whereas non-parametric tests do not have any probability assumptions.

2.2.1 Student's t-test

Student's t-test is a parametric statistical test to determine whether the means μ_X and μ_Y of two normally distributed populations X and Y significantly differ from each other given two samples x of size n and y of size m from the two populations with sample means $\bar{x}$ and $\bar{y}$. Hence, the null hypothesis is formulated as $H_0 : \mu_X = \mu_Y$ and the alternative hypothesis as $H_1 : \mu_X \neq \mu_Y$. If we assume equal variance for X and Y and an unequal sample size, we can calculate the test statistic t as follows:

$$t = \frac{\bar{x} - \bar{y}}{s \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}},$$

with s denoting the square root of the weighted mean of the two sample variances s_x^2 and s_y^2:

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}.$$

The observed value of t is then compared to the quantiles of a Student's t distribution with $n+m-2$ degrees of freedom, which yields a p-value. Given a pre-defined significance level α we can thus decide whether the null hypothesis can be rejected or not.

2.2.2 Wilcoxon rank-sum test

If we can not assume that two populations X and Y are normally distributed, we can use the non-parametric Wilcoxon rank-sum test to test whether one of these populations is ranked as superior to the other. This can be also regarded as a shift between the two populations X and Y. The Wilcoxon rank-sum test, which is also known as MannWhitney U test [117], serves as a powerful alternative to the t-test if we can not make any assumptions on the underlying distribution. The null hypothesis H_0 states that an observation of X exceeds an observation Y with the same probability as vice versa. Under the alternative hypothesis H_1 this probability is not equal to 0.5. Given two samples x of length n and y of length m from the populations X and Y we can calculate the test statistic U to decide whether the null hypothesis can be rejected or not.

Initially, we rank the combined set of x and y from the lowest to the highest value where average ranks are assigned in case of ties. Let R_x be the sum of all

ranks for the observations of the sample x and R_y for sample y. We can then calculate

$$U_x = nm + \frac{n(n+1)}{2} - R_x$$

and

$$U_y = nm + \frac{m(m+1)}{2} - R_y,$$

from where we define the test statistic as $U = \min(U_x, U_y)$. To determine whether H_0 can be rejected, we have to compare the U statistics to a table of critical values for a significance level α. If the sample size is high, U is asymptotically normally distributed. We can thus calculate a z score as follows:

$$z = \frac{U - \frac{nm}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}$$

with $z \approx N(0,1)$. A p-value can be obtained using the cumulative distribution function of the standard normal distribution ϕ by $p = \phi(z)$.

2.2.3 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (KS) test can be used to test whether two random variables X and Y follow the same distribution given two samples x of length n and y of length m. The null hypothesis is formulated as $H_0 : F_X(z) = F_Y(z)$, where F_X and F_Y correspond to the empirical cumulative distribution functions obtained from x and y, respectively. the alternative hypothesis is thus formulated as $H_1 : F_X(z) \neq F_Y(z)$ meaning that X and Y do not follow the same distribution. The test statistic D is calculated as follows:

$$D = \sup_z |F_X(z) - F_Y(z)|.$$

The null hypothesis can then be rejected at significance level α if

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

where

$$K_\alpha = \sqrt{\frac{\ln\left(\frac{2}{\alpha}\right)}{2}}$$

for large sample sizes. If the sample sizes are small, we have to use a table of critical values for a significance level α.

2.2.4 Fisher's exact test

The Fisher's exact test can be used to test for a significant association between two properties A and B given n observations, where we can determine for each observation whether or not it possesses the property A and B. The numbers of observations with a certain property can be arranged in 2×2 contingency table as illustrated in Table 2.1):

Table 2.1: 2×2 contingency table

	A	$\bar{A}$	row sum
B	n_{11}	n_{12}	$n_{1\cdot}$
$\bar{B}$	n_{21}	n_{22}	$n_{2\cdot}$
column sum	$n_{\cdot 1}$	$n_{\cdot 2}$	n

The null hypothesis is formulated as $H_0 : n_{11}/n_{\cdot 1} = n_{12}/n_{\cdot 2}$ meaning that the observation of property A is independent from observing property B. Consider the case that number of observations with property A and B is x yielding $n_{11} = x$. The probability P of obtaining such a particular configuration is then given by the hypergeometric distribution [38]:

$$P(n_{11} = x) = \frac{\binom{n_{1\cdot}}{x}\binom{n_{2\cdot}}{n_{21}}}{\binom{n}{n_{\cdot 1}}}$$

If we assume a one-sided test, we can obtain a p-value for rejecting the null hypothesis by summing up the probabilities for the given configuration and all possible more extreme configurations, i.e. higher values of n_{11}: $p = \sum_{i=x}^{\min(n_{1\cdot}, n_{\cdot 1})} P(n_{11} = i)$

2.2.5 Multiple testing correction

If we perform a statistical test on a dataset several times, the probability increases that the alternative hypothesis for one of these tests was falsely considered as true. To correct for this error we can make use of methods for the adjustment of p-values. These methods usually increase the p-values, which

were calculated in a multiple testing procedure, depending on the amount of performed statistical tests and on the chosen significance level α.

Bonferroni correction

The Bonferroni correction is a simple and conservative method to correct for multiple testing errors arising from a repeated hypothesis testing procedure. Let $H_1, \ldots, H_m$ be the set of null hypothesis for m repeated tests, which yield a set of p-values $p_1, \ldots, p_m$. The Bonferroni method controls for the multiple testing error by rejecting a null hypothesis H_i only if $p_i \leq \frac{\alpha}{m}$ for a given significance level α [16]. In order to correct the given set of p-values to obtain a set of adjusted p-values $\hat{p}_1, \ldots, \hat{p}_m$, we can thus simply multiply each of them with the total number of tests: $\hat{p}_i = p_i m$ for $i = 1, \ldots, m$.

Benjamini-Hochberg False Discovery Rate

The p-value correction method by Benjamini and Hochberg is based on the concept of *false discovery rates* (FDR) [12]. Let $H_1, \ldots, H_m$ be the set of null hypothesis for m repeated tests, which yield a set of ordered p-values $p_1, \ldots, p_m$ with $p_1 \leq p_2 \leq \ldots \leq p_m$. To control the FDR, we can define the following multiple testing procedure for a given significance level α:

Let k be the largest i for which $p_i \leq \frac{i}{m}\alpha$; then reject all H_i with $i = 1, \ldots, k$. We can use this procedure to obtain a set of adjusted p-values $\hat{p}_1, \ldots, \hat{p}_m$ as follows:

Set $\hat{p}_m = p_m$ and $\hat{p}_i = \min(p_i \frac{m}{i}, \hat{p}_{i+1})$ for $i = m-1, \ldots, 1$.

2.3 Correlation analysis

One way to determine a dependency between two random variables is the application of correlation analysis. In this section, we will introduce Pearson's product-moment coefficient and Spearman's rank correlation coefficient, which are both measures for a correlation between two random variables. In addition, we will show how we can test for statistically significant correlation.

2.3.1 Pearson's product-moment coefficient

Pearson's product-moment coefficient [131] is a measure for the linear dependency between two random variables X and Y. It is defined as

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y},$$

where $Cov(X,Y) = E\left[(X - E\left[X\right])(Y - E\left[Y\right])\right]$ corresponds to the *covariance* between X and Y and $\sigma_X = \sqrt{E(X^2) - (E(X))^2}$ denotes the standard deviation of X. $\rho_{X,Y}$ takes on values between -1 and 1, where -1 and 1 corresponds to perfect anticorrelation and correlation, respectively. A value of 0 denotes that the two variables are uncorrelated.

The *sample Pearson correlation coefficient* r is defined as

$$r_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}},$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ represents the *sample mean* of x and N denotes the sample size of x and y.

2.3.2 Statistical significance

In many cases, we want to know whether the correlation between two variables is statistically significant. For this purpose, we can apply a statistical test that is based on the *Fisher transformation* [39] of the correlation coefficient r. It can be calculated as follows

$$F(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right).$$

If we assume that the two underlying random variables X and Y are normally distributed, also $F(r)$ is normally distributed with mean $\mu = F(\rho)$ and standard error $\sigma = \frac{1}{\sqrt{n-3}}$, where n denotes the sample size. Usually, we want to test the null hypothesis that $\rho = 0$. Hence, we can calculate a z-score as

$$z = \frac{F(r) - \mu}{\sigma} = F(r)\sqrt{n-3}.$$

By using the cumulative distribution function of the standard normal distribution ϕ we can then obtain a two-sided p-value indicating whether the correlation significantly differs from zero:

$$p = (1 - \phi(F(|r|)\sqrt{n-3})) * 2.$$

If we only want to test for significant anticorrelation, we can obtain the one-sided p-value by

$$p = \phi(F(r)\sqrt{n-3}).$$

2.3.3 Spearman's rank correlation coefficient

Spearman's rank correlation coefficient assesses the dependency between two random variables X and Y using a monotonic function. In contrast to Pearson's product-moment coefficient, it can also determine non-linear relationships and is robust to outliers.

Let x and y be two samples of size N, which were drawn from the random variables X and Y, respectively, and let $R(x)$ and $R(y)$ be the ranks of x and y. Spearman's rank correlation coefficient then can be calculated as

$$r_{x,y} = 1 - \frac{6\sum_{i=1}^{N}(R(x)_i - R(y)_i)^2}{N(N^2-1)}.$$

To assess the statistical significance of Spearman's rank correlation coefficient, we can apply the Fisher transformation to $r_{x,y}$ as for Pearson's product-moment coefficient. However, the standard error should be chosen as $\sigma = \sqrt{1.06/(n-3)}$ [21].

2.4 Linear regression analysis

Here we give a brief introduction into the principles of regression analysis and penalized regression, which is inspired by [63].

Regression is a method in statistics for modeling a quantitative output $\mathbf{y} = (y_1, \ldots, y_N)$ given a set of p input variables $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$. Based on a regression model, a relationship between the output variable and the input variables can be estimated. If we assume a linear relationship between $\mathbf{y}$ and $\mathbf{X}$,

we call the process *linear regression* and furthermore *multiple linear regression*, if $\mathbf{X}$ consists of more than one variable ($p > 1$). A linear regression model has the form

$$\mathbf{y} \sim \beta_0 + \mathbf{X}\boldsymbol{\beta}^T + \boldsymbol{\epsilon}$$

with normally distributed error $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma)$, parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and the intercept β_0. We can use the linear regression model to predict an outcome $\hat{\mathbf{y}}$ by estimating the corresponding coefficients $\hat{\boldsymbol{\beta}} = \{\hat{\beta_0}, \ldots, \hat{\beta_p}\}$. A common method for estimating $\hat{\boldsymbol{\beta}}$ is the least squares approach where an optimal coefficient set is chosen that minimizes the residual sum of squares

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2.$$

This can be expressed in matrix notation as follows

$$RSS(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Differentiating this equation with respect to $\boldsymbol{\beta}$ and setting it to zero results in

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0.$$

We can therefore obtain the parameter set $\hat{\boldsymbol{\beta}}$ that minimizes the residual sum of squares by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Note that $\hat{\boldsymbol{\beta}}$ is only computable by this expression if $\mathbf{X}^T\mathbf{X}$ is invertible. This only holds if $\mathbf{X}$ has full rank, which means that all predictors in $\mathbf{X}$ must be linearly independent.

The least square method is a suitable approach for the estimation of an optimal parameter set. However, the precision of the least squares estimates may become low if some input variables do not fit well into the model. In this case, we might shrink the coefficients of these variables or even set them to zero, thereby excluding them from the model. This furthermore provides a subset of input variables with the strongest effects on the output variable.

Several approaches have been proposed in order to identify a subset of input variables that describe the output variable best. The most intuitive one is *best subset regression*, where for each $k \in \{0, \ldots, p\}$ the subset of size k is determined by a *leaps and bounds* procedure [46] that gives smallest residual sum of squares.

However, this approach becomes infeasible for large p. For this reason, *forward- and backward-stepwise selection* may be more appropriate. Here, a greedy algorithm starts with the empty or full model and adds or removes variables in each step, respectively. In each step, the quality of the model is evaluated by using a quality measure like the Akaike information criterion (AIC) [1]. Finally, the model with the best quality is retained.

2.4.1 Shrinkage

Even though best subset regression provides a model that might have lower prediction error than the full model, several problems have to be considered [41]. One major drawback is the strongly biased R^2 value, which leads to an overestimation of the fit. Furthermore, the regression coefficients are usually estimated too large. In order to overcome these drawbacks, *shrinkage methods* have been proposed. These methods aim to shrink the regression coefficients by introducing a penalty on their size.

The coefficients obtained by *ridge regression* [67] minimize a penalized residual sum of squares as follows:

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

Note that this expression is equivalent to the expression of the unpenalized regression above if $\lambda = 0$. On the other hand, the magnitudes of $\hat{\boldsymbol{\beta}}$ are forced to become smaller dependent on the magnitude of λ. The parameter λ therefore controls the amount of shrinkage in the regression model.

The expression can be solved in closed form by solving the following equation in matrix form [63]:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

where $\mathbf{I}$ denotes the identity matrix. Since we add a positive constant to the diagonal of $\mathbf{X}^T\mathbf{X}$ for $\lambda > 0$, it is not necessary for ridge regression that $\mathbf{X}$ must have full rank, which is the prerequisite for unpenalized regression. However, we have to assure that the predictor variables are on the same scale in order to provide a fair penalization of the regression coefficients. We therefore have to standardize the data prior to the regression analysis.

Another shrinkage method is *the least absolute shrinkage and selection operator (lasso)* [170]. The lasso estimate is defined as

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

This expression differs only slightly from ridge regression. In case of the lasso, the penalty is defined as $\lambda \sum_{j=1}^{p} |\beta_j|$. This penalty can be also expressed as $||\boldsymbol{\beta}||_1$ and is referred to as the L_1 penalty. The ridge penalty in turn corresponds to a L_2 penalty and can be also expressed as $||\boldsymbol{\beta}||_2$. In contrast to the L_2 penalty of ridge regression, the nature of the L_1 penalty causes a shrinkage of the regression coefficients so that the may become exactly zero. It therefore actually performs a feature selection on the predictor variables by removing variables that do not fit the model well.

Another property of the L_1 penalty of the lasso is that it can not be solved in closed form as it is the case for ridge regression. Hence, we have to approximate the solution by a certain optimization algorithm. Originally, a quadratic programming approach was proposed for finding optimal lasso solutions [170]. However, more efficient algorithms have been introduced. The *Least Angle Regression (LARS)* [34] allows for the calculation of the lasso solution with the same computational costs as for ridge regression. Other methods for the efficient calculation of lasso solutions are based on coordinate descent approaches. These approaches were originally proposed in the late 90's [45] and meanwhile further improved to be very competitive with LARS, thereby providing a higher flexibility [42]. In general, all approaches yield an entire regularization path as λ is varied. The optimal choice of lambda can then be determined by cross-validation.

2.4.2 Elastic net

Even though the lasso provides a powerful tool for feature selection in a linear regression model, it has some major drawbacks. Primarily, it has been shown that the lasso has several issues with regard to predictor variables that are highly correlated with each other [190]. In this case, the lasso tends to select only one representative out of a set of highly correlated predictor variables. This leads to a certain loss of information, especially if one is interested in analyzing predictors that are correlated with each other. In this case, a method would be desirable that retains the correlated variables in the model, if a combined effect can be determined. For this purpose, the *elastic net* was proposed [190]. The elastic net solves the following problem:

$$\hat{\boldsymbol{\beta}}^{EN} = \arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda P_\alpha(\boldsymbol{\beta}) \right\},$$

where

$$\begin{aligned} P_\alpha(\boldsymbol{\beta}) &= (1-\alpha)\frac{1}{2}||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1 \\ &= \sum_{j=1}^{p} \frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|. \end{aligned}$$

P_α denotes the elastic net penalty and includes both, the lasso penalty (L_1) and the ridge penalty (L_2). The elastic net penalty is therefore a certain compromise between the two penalties. The additional parameter α controls how strong one of the two penalties is taken into account. In the extreme cases $\alpha = 0$ and $\alpha = 1$, an ordinary ridge regression and lasso is performed, respectively. If $0 < \alpha < 1$, both penalties are considered accordingly, which combines the advantages of both penalties: feature selection of the lasso and appropriate handling of correlated variables by ridge regression.

As for the lasso, the elastic net solution can not be determined in closed form. Therefore, the LARS algorithm has been adapted for the calculation of a whole elastic net regularization path with the computational costs of a single ordinary least squares fit (LARS-EN) [190]. It has been shown that the LARS-EN algorithm can even be outperformed by using a coordinate descent

approach [43]. This approach is implemented in the *R* package *glmnet* [43], which furthermore includes the ability of calculating the solutions of the pure lasso and ridge regression.

2.5 Analysis workflow for large-scale molecular profiling data

In the following we will describe a common workflow for the analysis of large-scale molecular profiling data, which can be used for the basic analysis of datasets from different experiments such as microarrays or RNA-seq. It is divided in two parts: statistical analysis and functional gene set analysis.

2.5.1 Statistical analysis

An important goal of large-scale molecular profiling is to gain comprehensive insight into molecular changes caused by a certain phenotype. These phenotypes include diseases, cell differentiation, treatments or tissue type. In order to identify these changes, statistical tests can be applied on a data set of replicated samples under the respective conditions. A typical approach is the application of a t-test combined with a predefined significance cutoff, which is typically $p < 0.05$. An generalization of the t-test is the analysis of variance (ANOVA), which can be used if we want to compare more than two groups. In this case, the ANOVA assesses whether at least one population mean is different from the others. ANOVA can also be further generalized if we express it in form of a linear regression model. This allows for a flexible modelling of relationships between the given phenotypic information and the observed molecular signature. It can be used for different scenarios for example time-resolved experimental setups and can also account for batch effects. A prominent example for a linear model-based approach is *Linear Models for Microarray Data (limma)* [162], which can be used for for the analysis of differential molecular abundances and which is part of *Bioconductor* [50] for the *R* environment [137]. Limma calculates a moderated t statistic if applied on two conditions, which, in contrast to an ordinary t-test, includes the information of all molecular signatures to calculate the individual variances. These methods yield a binary decision for

each molecule in the dataset whether its molecular signature is differentially abundant between the given conditions or not.

2.5.2 Functional analysis

After the identification of molecular changes related to a certain phenotype, we may want to interpret this result with regard to the underlying biological processes. A common approach to achieve this is the application of gene set enrichment methods.

The basic principle of gene set analysis is the identification of functional properties, which are enriched among a gene set of interest. These functional properties in turn can be regarded as sets of genes, which share a common functional property. We define these gene sets as *functional groups*, whose functional property is described by a *term*. A collection of terms is defined as *ontology*. An ontology can be interpreted in many ways but usually corresponds to biological processes or signaling pathways. However, an ontology may also have different meanings such as chromosomal regions. Several ontology databases have been set up to map functional information on genes. The most prominent resource is GO [5], which holds three different ontologies: biological processes, molecular functions and cellular components. Other resources, such as KEGG [84], WikiPathways [87] or Reactome [26], deal with the mapping of signaling pathways on genes. These resources are often used for gene set analysis.

The most commonly used method for gene set analysis is the application of Fisher's exact test [18, 58] (see Sec. 2.2.4) on the basis of functional categories from a certain ontology resource. It is implemented in several popular gene set analysis tools like GOstats [37] or DAVID [72]. Let R denote the set of significantly altered genes out of a total set of analyzed genes N, which were determined by the statistical analysis (see above). Given an ontology with m terms, we want to test for each term i whether its associated functional group D_i is enriched among the set of interest R with $i = 1, \ldots, m$. For this purpose, we can apply Fisher's exact test on a 2×2 contingency matrix (see Sec. 2.2.4), by defining property A as "assigned to R" and property B as "assigned to D_i". The corresponding table is illustrated in Tab. 2.2.

Applying this procedure to each of the functional categories D_i available in

Table 2.2: 2×2 contingency table for gene set analysis given a functional category D_i and a gene set of interst R.

	A	$\bar{A}$
B	$\lvert R \cap D_i \rvert$	$\lvert D_i \setminus R \rvert$
$\bar{B}$	$\lvert R \setminus D_i \rvert$	$\lvert M \setminus (R \cup D_i) \rvert$

the database yields m p-values. This repeated testing procedure results in a multiple testing problem. We therefore have to correct the p-values accordingly. However, the functional groups retrieved from the established databases are usually not independent of each other. Especially the functional groups from GO [5] are highly dependent on each other, since the categories are built up as a directed acyclic graph (DAG), which gives rise to a tree-like representation. Due to the *true-path rule* of GO, all parent categories also contain the genes associated with their children categories. p-value correction is therefore strongly biased. Another problem caused by these overlaps is the high amount of redundancy among the resulting categories from the enrichment analysis.

A number of methods have been proposed to overcome these issues. The *elim* algorithm [2] decorrelates the GO terms by processing the GO DAG in a bottom-up fashion to systematically remove genes from the assigned gene set. Starting at the bottom of the GO DAG, the *elim* algorithm performs Fisher's exact test on each GO term to calculate its p-value. It then marks the corresponding node as significant if the p-value (Bonferroni corrected) is below a significance level of 0.01. If a node is marked as significant, the algorithm removes all assigned genes from its ancestor nodes. A related approach is the *weight* method [2], which down-weights the genes in less significant neighbor categories instead of removing them completely from the analysis.

Furthermore, model-based approaches were introduced in order to deal with redundancies, which were initially based on the combination of the model likelihood and a penalization [110] and were further optimized by using a Bayesian modelling approach [9] (see Chapter 7.1).

Since nowadays more and more studies make use of multiple omics techniques at once, methods have been proposed that perform gene set analysis on several levels simultaneously. Thomas et al. [169] have addressed this issue by introducing an ontology jointly representing disease risk factors and causal

mechanisms based on genome typing and epidemiology studies. The proposed ontology is disease-specific (nicotine addiction and treatment) and only applicable to very specific research questions. Recently, an algorithm was introduced for the combined analysis of the protein and mRNA level [25].

2.6 Bayesian networks

In this section, we give a brief introduction into Bayesian networks and the inference of posterior probability distributions, which is inspired by [15].

2.6.1 The concept of Bayesian networks

A Bayesian network is a graphical model that represents the joint probability distribution of a set of random variables. It is a directed acyclic graph (DAG) where the nodes correspond to the random variables and the edges to the conditional dependencies between them. Each of the random variables is associated with a conditional distribution, which is conditioned on the parent nodes in the graph. As these graphical models represent the causal processes, by which the data was generated, they are often called *generative models*.

The joint distribution corresponding to a Bayesian network is given by the product over all conditional distributions of the nodes with pa_k denoting the set of parents of x_k and $\boldsymbol{x} = \{x_1, \ldots, x_K\}$:

$$p(\boldsymbol{x}) = \prod_{k=1}^{K} p(x_k|pa_k)$$

For example, in the case of three random variables a, b, c, the joint distribution is given by

$$p(a, b, c) = p(b|a, c)p(c|a)p(a).$$

Then, this joint distribution then can be represented as a DAG as illustrated in Fig. 2.4.

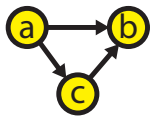


Figure 2.4: A directed acyclic graph corresponding to the joint distribution of the three random variables a,b, and c. In this case, the joint distribution is given as $p(a, b, c) = p(b|a, c)p(c|a)p(a)$.

2.6.2 Bayesian inference

Consider two nodes x and y in our network with joint probability distribution $p(x, y) = p(x)p(y|x)$. Let us suppose that we have obtained observations of a node y in our network. This enables us to compute the conditional probability $p(x|y)$ using Bayes' theorem

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}.$$

The joint distribution is then expressed in terms of $p(y)$ and $p(x|y)$. In general, we assume that a subset of variables is observed whereas the others are hidden. We then aim to calculate the posterior probability distributions conditioned on the observed variables. In the following, we refer to the observed variables as the given data $\boldsymbol{D}$ with probability $p(\boldsymbol{D})$. The hidden variables correspond to the model parameters $\boldsymbol{\theta}$ with prior $p(\boldsymbol{\theta})$ and posterior probability distribution $p(\boldsymbol{\theta}|\boldsymbol{D})$ conditioned on the data.

In practice, the posterior probability distribution of a Bayesian network usually can not be analyzed analytically, which is due the fact that $p(\boldsymbol{\theta}|\boldsymbol{D})$ can only be evaluated up to the normalization constant $p(\boldsymbol{D})$. The evaluation of $p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ is feasible, whereas the marginalization over $\boldsymbol{\theta}$ for $p(\boldsymbol{D}) = \int p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ is usually intractable. However, several approximate inference algorithms are available that allow for an efficient computation of the posterior probability distribution given a Bayesian network. These methods are either based on stochastic or deterministic approximations. The most prominent stochastic inference technique is *Markov chain Monte Carlo* (MCMC). However, sampling methods can become computationally demanding especially for large-scale setups. Deterministic approximation techniques like *expectation propagation* or *variational Bayes* may be the better choice in these cases.

Markov chain Monte Carlo

The arrangement of the joint probability distribution implies a certain hierarchy among the random variables in the network. In order to draw samples from such a distribution, we have to consider this hierarchy. Therefore, we have to start with drawing samples from all nodes with no incoming edges. Following the hierarchy given in the model, we are then able to successively draw a sample from the conditional distribution $p(x_n|pa_n)$ of the child node n assuming that a sample of the parent nodes pa_n has already been drawn. Drawing a sample from the conditional distribution of a lowest node in the hierarchy then corresponds to a sampling from the joint probability distribution $p(\boldsymbol{x})$. It is furthermore possible to sample from the marginal probability distribution by drawing from a subset of nodes corresponding to a certain node n in the network. To do so, we just have to take the sampled value of n into account and discard the others.

The sequential drawing of samples is the basic principle in the inference of the posterior using MCMC. However, the generation of these samples is not straight-forward and requires sophisticated approaches. MCMC methods allow for the sampling from a variety of distributions, thereby scaling well with the dimensionality of the sampling space. It has the ability to generate exact results if the algorithm is run with infinite computational resources.

The goal of MCMC is the use of Markov chains to sample from a given distribution. A first-order Markov chain is defined as a sequence of random variables $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(M)}$ such that

$$p(\boldsymbol{z}^{(m+1)}|\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}) = p(\boldsymbol{z}^{(m+1)}|\boldsymbol{z}^{(m)})$$

holds for $m \in \{1, \ldots, M-1\}$. Hence, the random variable $\boldsymbol{z}^{(m+1)}$ is exclusively conditioned on its predecessor $\boldsymbol{z}^{(m)}$. Furthermore, $Tm(\boldsymbol{z}^{(m)}, \boldsymbol{z}^{(m+1)})$ is the transition probability between each random variable $\boldsymbol{z}^{(m)}$ and its successor $\boldsymbol{z}^{(m+1)}$. Such a Markov chain is illustrated in Fig. 2.5.

A Markov chain is called *homogenous* if all transition probabilities are the same for all m. A distribution is called *invariant* with respect to a Markov chain if that distribution remains invariant after each step in the Markov chain. If the transition probabilities are chosen in a way such that $p^*(z)T(z, z') =$

Figure 2.5: Concept of a Markov chain with regard to the random variables $\boldsymbol{z}$. Each random variable $\boldsymbol{z}^{(m+1)}$ exclusively depends on its predecessor $\boldsymbol{z}^{(m)}$ with the transition probability $Tm(\boldsymbol{z}^{(m)}, \boldsymbol{z}^{(m+1)})$.

$p^*(z')T(z', z)$ (*detailed balance*), the distribution will be always invariant. In order to be able to sample from a given distribution using Markov chains, we have to ensure invariance of the distribution as well as homogenity of the Markov chain. In addition, for $m \to \infty$, the distribution $p(z(m))$ has to converge to the required invariant distribution $p^*(z)$, irrespective of the choice of initial distribution $p(z(0))$ (*ergodicity*). A homogenous Markov chain is ergodic with very few restrictions on the invariant distribution and the transition probabilities [124].

The *Metropolis-Hastings algorithm* [64] is a popular MCMC method and is based on a random walk through the sample space. Instead of directly sampling from a given distribution $p(\boldsymbol{z})$, which might be infeasible, the samples are drawn from a simpler *proposal distribution* $q(\boldsymbol{z})$. These samples $\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots$ then form a Markov chain, where each sample corresponds to the current state $\boldsymbol{z}^{(\tau)}$ and the proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$ directly depends on this current state. We therefore draw at each state $\boldsymbol{z}^{(\tau)}$ a sample $\boldsymbol{z}^*$ from the proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$. This sample is then accepted with the probability

$$A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\boldsymbol{z}^*)q(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)})q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})}\right).$$

Here we suppose that the evaluation of $\tilde{p}(\boldsymbol{z})$ for the given sample $\boldsymbol{z}$ is feasible, where $\tilde{p}(\boldsymbol{z})$ is proportional to $p(\boldsymbol{z})$ with $p(\boldsymbol{z}) = \tilde{p}(\boldsymbol{z})/Z_p$ for an unknown constant Z_p, which corresponds to $p(\boldsymbol{D})$ in our context.

In fact, we accept a sample if $A_k(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) > u$, where u is some random number drawn from an uniform distribution over the interval $(0, 1)$. The new state $\boldsymbol{z}^{(\tau+1)}$ is then set to $\boldsymbol{z}^*$, if the sample was accepted. Otherwise, it becomes the same as before, which means that $\boldsymbol{z}^{(\tau+1)}$ is set to $\boldsymbol{z}^{(\tau)}$.

It can be easily shown that $p(\boldsymbol{z})$ is in detailed balance [15]. The required invariance condition is therefore satisfied.

We finally obtain the desired posterior probability distribution $p(\boldsymbol{z})$, since the distribution of $\boldsymbol{z}^{(\tau)}$ tends to $p(\boldsymbol{z})$ as $\tau \to \infty$ [15]. Alternative sampling methods include slice sampling [125] and Gibbs sampling [49]. The Metropolis-Hastings algorithm is summarized in Algorithm 2.1.

Input: Proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$, initial state $\boldsymbol{z}^{(1)}$.
Result: $\boldsymbol{z}^{(\tau)}$, which tends to $p(\boldsymbol{z})$ as $\tau \to \infty$.
$\tau = 1$;
repeat
  Draw sample $\boldsymbol{z}^*$ from $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$;
  $A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\boldsymbol{z}^*)q(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)})q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})}\right)$;
  if $A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) > rand.uniform(0,1)$ **then**
    $\boldsymbol{z}^{\tau+1} = \boldsymbol{z}^*$;
  end
  $\tau = \tau + 1$;
until *converge*;
return $\boldsymbol{z}^{(\tau)}$

Algorithm 2.1: Metropolis-Hastings algorithm [64]

Approximate inference

The inference of posteriors using sampling methods may become very challenging for large-scale problems. Therefore, we often wish to use deterministic approximations for this purpose. These approaches perform well with regard to computational costs also for large applications. In contrast to MCMC, they will always only provide an approximation of the posterior. However, this is usually sufficiently accurate for many problems.

Expectation propagation (EP) is an approximate inference method proposed by Minka et al. [120].

The basic assumption of EP is that the posterior probability distribution factorizes in the following way:

$$p(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{1}{p(\boldsymbol{D})} \prod_i f_i(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ are all parameters of the model and f_i functions as defined in the model specifications, which depend on the definition of the specific generative model.

The model evidence $p(\boldsymbol{D})$ is given by

$$p(\boldsymbol{D}) = \int \prod_i f_i(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

This marginalization over $\boldsymbol{\theta}$ however is usually intractable and demands for some sort of approximation. In EP this is realized by an approximation of the form

$$q(\boldsymbol{\theta}) = \frac{1}{Z} \prod_i \tilde{f}_i(\boldsymbol{\theta}),$$

where the factors $f_i(\boldsymbol{\theta})$ in the true posterior are approximated by $\tilde{f}_i(\boldsymbol{\theta})$ coming from the exponential family. The factor $1/Z$ serves as a normalizing constant. In order to obtain an optimal approximation, we aim to minimize the Kulback-Leibler (KL) divergence between the true posterior and the approximation, which is given by

$$\mathrm{KL}(p||q) = \mathrm{KL}\left(\frac{1}{p(\boldsymbol{D})} \prod_i f_i(\boldsymbol{\theta}) \middle\| \frac{1}{Z} \int \prod_i \tilde{f}_i(\boldsymbol{\theta}) \right),$$

where

$$\mathrm{KL}(p\|q) = \int p(\boldsymbol{\theta}|\boldsymbol{D}) \ln \left(\frac{p(\boldsymbol{\theta}|\boldsymbol{D})}{q(\boldsymbol{\theta})} \right) \mathrm{d}\boldsymbol{\theta}.$$

This approximation is again intractable but could be roughly approximated by a pairwise calculation of the KL divergence between the factors $f_i(\boldsymbol{\theta})$ and $\tilde{f}_i(\boldsymbol{\theta})$.

However, this would lead to a very poor approximation. Hence, EP aims to optimize each factor in the context of the remaining factors. We iteratively refine each of the factors $\tilde{f}_j(\boldsymbol{\theta})$ by initially removing this factor from the product yielding $\prod_{i \neq j} \tilde{f}_i(\boldsymbol{\theta})$. We then generate a revised form of the factor $\tilde{f}_j$ by ensuring that

$$q^{new}(\boldsymbol{\theta}) = \tilde{f}_j(\boldsymbol{\theta}) \prod_{i \neq j} \tilde{f}_i(\boldsymbol{\theta})$$

is as close as possible to

$$f_j(\boldsymbol{\theta}) \prod_{i \neq j} \tilde{f}_i(\boldsymbol{\theta}),$$

which can be easily evaluated by the KL divergence. This refinement is then repeated for each factor in several steps. We can then obtain the approximated normalization constant Z by integrating over the product of all optimized fac-

tors. The algorithm for approximating the posterior probability distribution using EP is summarized in Algorithm 2.2.

Input: Factorized posterior

$$p(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{1}{p(\boldsymbol{D})} \prod_i f_i(\boldsymbol{\theta})$$

Result: Gaussian approximation $q(\boldsymbol{\theta}|\boldsymbol{D})$ of posterior.
Initialize Gaussian term approximations $\tilde{f}_j(\boldsymbol{\theta})$;
repeat
  for *j=1:Number of factors* **do**
    Update $\tilde{f}_j$ such that $\tilde{f}_j(\boldsymbol{\theta}) \prod_{i \neq j} \tilde{f}_i(\boldsymbol{\theta})$ minimizes KL-divergence from $f_j(\boldsymbol{\theta}) \prod_{i \neq j} \tilde{f}_i(\boldsymbol{\theta})$
  end
until *all* $\tilde{f}_j$ *converge*;
Approximate $p(\boldsymbol{D})$ as $Z = \int \prod_i \tilde{f}_i(\boldsymbol{\theta}) d\boldsymbol{\theta}$;
return $q(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{1}{Z} \prod_i \tilde{f}_i(\boldsymbol{\theta})$

Algorithm 2.2: Expectation Propagation for approximating the posterior [120]

2.7 Molecular mechanisms of adipogenesis

Adipocytes derive from multipotent mesenchymal stem cells [145] from where they undergo certain intermediate states until they become fully differentiated mature adipocytes. These intermediate states can be divided in two major steps. It starts with the commitment to the adipocyte lineage, which leads to the conversion of the multipotent stem cells to preadipocytes. These preadipocytes then lack the ability to differentiate into other cell types. In a second step, they differentiate into mature adipocytes, which are equipped with the cellular machinery that is needed to act as regulators of the organism's energy household.

The process of adipocyte differentiation involves a variety of molecular mechanisms. The most intensively studied transcription factor, which is also regarded as the "master regulator" of adipogenesis, is the peroxisome proliferator-activated receptor γ (PPARγ) [145]. It has been shown that the activation of this receptor is necessary for adipogenesis and at the same time sufficient for its induction [146]. Hence, pro-adipogenic factors like CCAAT-enhancer-binding proteins (C/EBPs) and Krüppel-like factors (KLFs) generally act as activators of PPARγ [145] (Fig. 2.6a). Essential pathways in adipogenesis furthermore activate or repress the PPARγ gene transcriptionally in order to influence the process of differentiation. For instance, the activation of bone morphogenetic proteins (BMPs), which are part of the transforming growth factor-β (TGFβ) superfamily, is known to inhibit adipogenesis through activation of SMAD3, which in turn binds to C/EBPs and inhibits their transcriptional activity [24]. Further important signaling pathways in adipogenesis are amongst others the WNT pathway [148], insulin growth factor-1 (IGF1) receptor signaling [161] and the mitogen-activated protein kinase (MAPK) pathway [17].

Besides the formation of adipocytes, mesenchymal stem cells may also differentiate in other cell types including osteoblasts [80], which form bone tissue. The key factor for the commitment of this lineage is the runt-related transcription factor 2 (RUNX2). In fact, the decision whether a mesenchymal stem cell differentiates into the osteogenic or adipogenic lineage primarily depends on the presence of RUNX2 and PPARγ, respectively, which are in turn found to be down-regulated in the other cell type [102]. Several cofactors are involved in the lineage decision of which some are able to act as both co-activators and

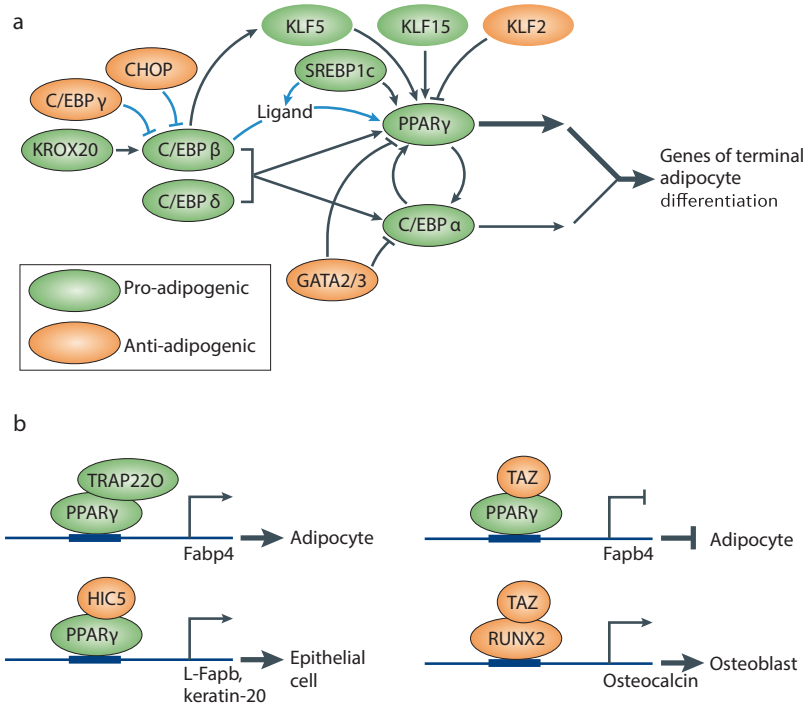


Figure 2.6: **(a)** Peroxisome proliferator-activated receptor γ (PPARγ) is the key regulator of adipocyte differentiation, which is itself regulated by a set of other factors. Other pro-adipogenic factors such as CCAAT-enhancer-binding proteins (C/EBPs) and Krüppel-like factors (KLFs) act as activators of PPARγ. Effects on gene expression are indicated by black lines, whereas effects on protein activity are indicated by blue lines. **(b)** Depending on the activity of nuclear cofactors of such as HIC5 and TRAP220, the differentiation process may end up in different cell fates. Some cofactors like the transcriptional co-activator with PDZ-binding motif (TAZ) can act as both co-activators and co-repressors of gene expression. TAZ can repress the pro-adipogenic activity of PPARγ and at the same time promote osteoblast differentiation through the activation of RUNX2. Figure adapted from [145].

co-repressors of gene expression. For example TAZ can promote osteogenesis and at the same time inhibit adipogenesis by activating RUNX2 and repressing PPARγ, respectively (Fig. 2.6b) [70].

Chapter 3

Materials

Throughout this thesis, we will investigate the molecular mechanisms of adipocyte differentiation using a dataset, which comprises mRNA and miRNA expression as well as DNA methylation. The dataset was generated by collecting preadipocytes from 20 highly obese probands (body-mass index >40) at the group of Harald Staiger (Universitätsklinikum Tübingen). These preadipocytes were differentiated *in vitro* for 20 days. Measurements were taken of the preadipocytes (day 0) and the fully differentiated mature adipocytes (day 20). For mRNA profiling the Affymetrix GeneChip®Human Gene 1.0 ST Array, for miRNA profiling the Affymetrix GeneChip®miRNA 2.0 Array and for DNA methylation the Illumina Infinium HumanMethylation450 Array was used. All these experiments were conducted in the group of Johannes Beckers (Institute of Experimental Genetics at the Helmholtz Center Munich).

Microarray quality analysis identified one miRNA array, which exhibited an irregular expression pattern. Therefore, this array was excluded from further analysis. We processed the miRNA and mRNA samples within the *R* framework for statistical computing [137] using the *affy* package [48], which is included in *Bioconductor* [50]. We used robust multi-array average (RMA) [77] for microarray normalization.

We analyzed the mRNA and the methylation data separately using statistical methods to determine individual changes on these two levels. Prior to the statistical analysis, the mRNA data was filtered independently using the *gene-*

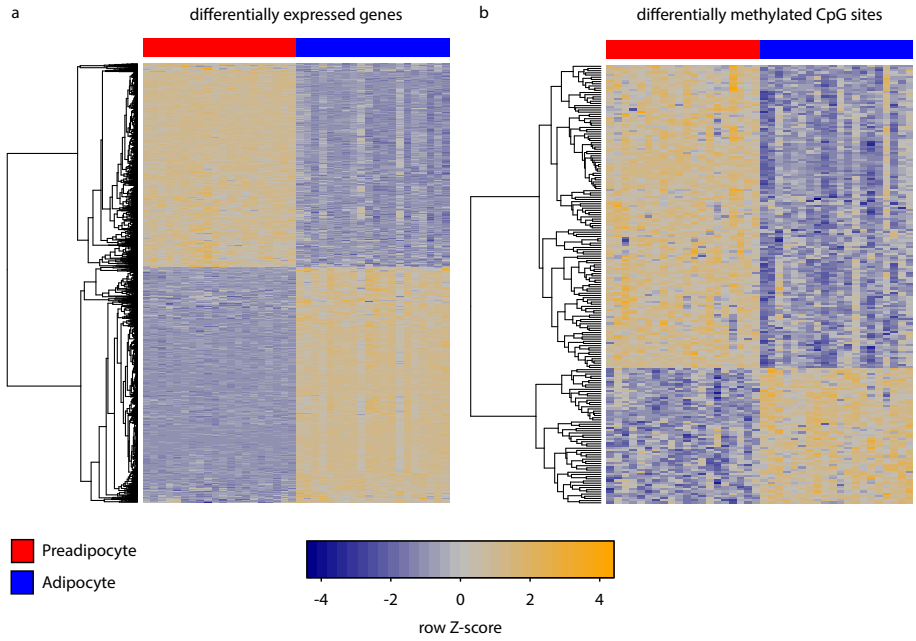


Figure 3.1: Differentially expressed genes **(a)** and differentially methylated CpG sites **(b)** between preadipocytes (red samples) and adipocytes (blue samples). The gene expression and methylation values were standardized row-wise, respectively. Low values are indicated in blue whereas high values are colored orange.

filter package for eliminating probesets, which are unexpressed or not annotated. For all datasets, we used *limma* [162] to identify statistically significant changes between preadipocytes and adipocytes. For the mRNA dataset, we calculated the moderated t statistics (see Chapter 2.5.1) on the mRNA expression dataset for inferring the p-values that indicate statistically significant gene expression changes between the two cell types. We applied the p-value correction by Benjamini and Hochberg [12] to account for multiple testing. We considered a gene to be differentially expressed if its adjusted p-value was below 0.05 and if it was at least two fold up- or down-regulated.

For the methylation data, we used M-values, which are an appropriate measure for the statistical analysis of methylation across the samples [33] (see Chapter 2.1.2). The moderated t statistics were computed for each CpG site on the array and the resulting p-values were corrected for multiple testing accordingly. A CpG site was considered to be differentially methylated between the two cell types if its corrected p-value was below 0.05. The testing procedure revealed

1,396 differentially expressed mRNAs and 201 differentially methylated CpG sites (Fig. 3.1).

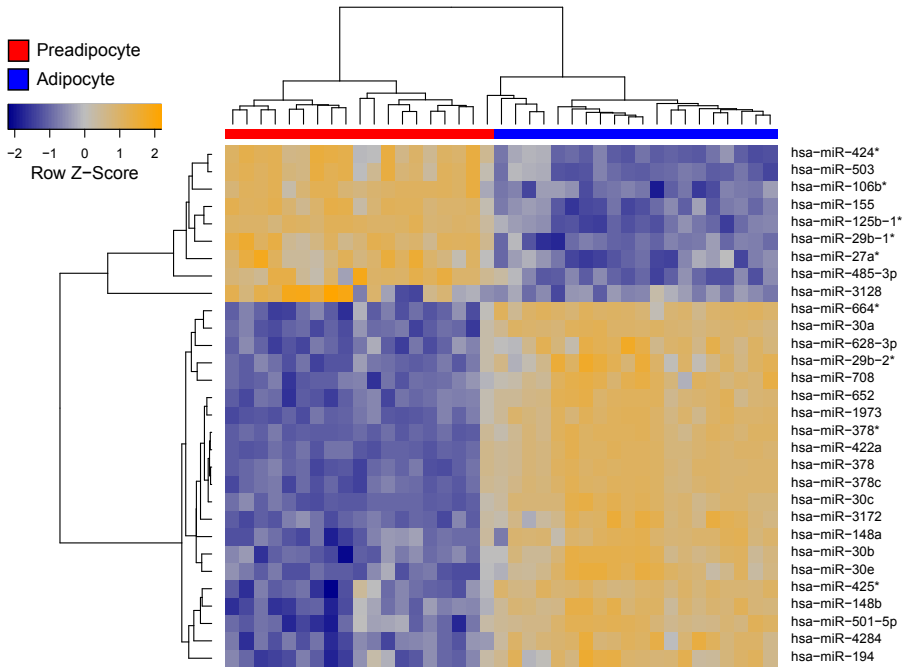


Figure 3.2: Heatmap of differentially expressed miRNAs between the samples derived from adipocytes and the ones from preadipocytes. The coloring indicates the row scaled expression values from low expression (blue) over medium expression (grey) to high expression (orange). The columns are marked blue if the respective sample is derived from an adipocyte and red if it is derived from a preadipocyte. The clustering of miRNAs and samples is based on the correlation of the miRNA and sample expression profiles, respectively.

In case of the miRNA dataset, we applied Bonferroni p-value correction. Hence, the raw p-values were multiplied by the total number of miRNAs in the dataset. A miRNA was supposed to be differentially expressed if their adjusted p-value was below 0.05 and its mean expression was at least 4-fold up- or down-regulated between the adipocyte and preadipocyte samples. The statistical analysis revealed 30 differentially expressed miRNAs, with 9 of them down- and 21 up-regulated after the differentiation process (Fig. 3.2).

Chapter 4

MicroRNAs and their role in gene regulation

In this chapter, we will introduce the basic principles of post-transcriptional gene regulation through miRNAs. We subdivide this chapter in two parts. The first part introduces the process of miRNA biogenesis and gene regulation as well as experimental methods and bioinformatic resources, which are commonly used in miRNA research. We furthermore discuss the properties of miRNAs with regard to their chromosomal arrangement and transcription. As the underlying mechanisms of miRNA regulation are still only poorly understood, we aim to further broaden this knowledge by investigating coordinated miRNA regulation of protein complexes. This work is introduced in the second part of this chapter. By showing that co-expression of miRNAs is an important aspect in the coordinated regulation of protein complexes we can provide insights into regulatory mechanisms that are essential for the development of novel methods dealing with the identification of miRNA-target relationships. This part of this chapter was published in:

- **Steffen Sass***, Sabine Dietmann*, Ulrike C. Burk, Simone Brabletz, Dominik Lutter, Andreas Kowarsch, Klaus F. Mayer, Thomas Brabletz, Andreas Ruepp, Fabian J. Theis, and Yu Wang. MicroRNAs coordinately regulate protein complexes. *BMC Systems Biology*, 5(1):136, August 2011.

$*$ = equal contributions

4.1 MicroRNA biogenesis

Similar to protein-coding genes, miRNA genes are transcribed by Polymerase II. They may be located in the intronic regions of coding or non-coding genes leading to the transcriptional regulation through the promoters of these host genes [92]. A set of miRNAs may be furthermore located in close proximity on a chromosome, which is referred to as *miRNA cluster*. These clustered miRNAs are usually under the control of a common promoter and therefore transcribed as a single transcript [92]. Due to this co-regulation, miRNAs are frequently co-expressed either with their host genes or clustered miRNAs [8].

The original transcript (pri-miRNA) is processed in two major steps yielding the functional mature miRNAs (Fig. 4.1). Initially, the pri-miRNA is cleaved by the Drosha protein inside of the nucleus resulting in ∼70-nucleotide precursor miRNAs (pre-miRNAs). These pre-miRNAs consist of stem-loop structures, which are further processed outside of the nucleus. The export from the nucleus to the cytoplasm is mediated by exportin 5, which specifically binds to correctly processed pre-miRNAs [112]. In the cytoplasm, the hairpin of the pre-miRNAs is cleaved by the Dicer protein into ∼21-25-nucleotide double-stranded RNA which consists of the 3' miRNA and the 5' miRNA.

miRNAs are able to post-transcriptionally regulate the expression of genes in a specific manner. The mature miRNA is loaded into the *RNA-induced silencing complex (RISC)*. The double-stranded mature miRNA is initially inserted into Argonaut (Ago) proteins, which form the core component of the RISC [86]. Within the Ago proteins, the RNA duplex is separated. Only one strand remains bound to the complex (functional strand) while the other one is discarded (passenger strand). Usually, one of the two pre-miRNA strands is preferably selected as functional strand. Hence, it was common to indicate this by the nomenclature. The strand of the miRNA, which was supposed to serve as passenger strand more frequently, was annotated as miRNA while the other one as "star" miRNA (miRNA*). However, several recent publications report biological functions of miRNA* [183, 127]. Therefore, the miRNA/miRNA*

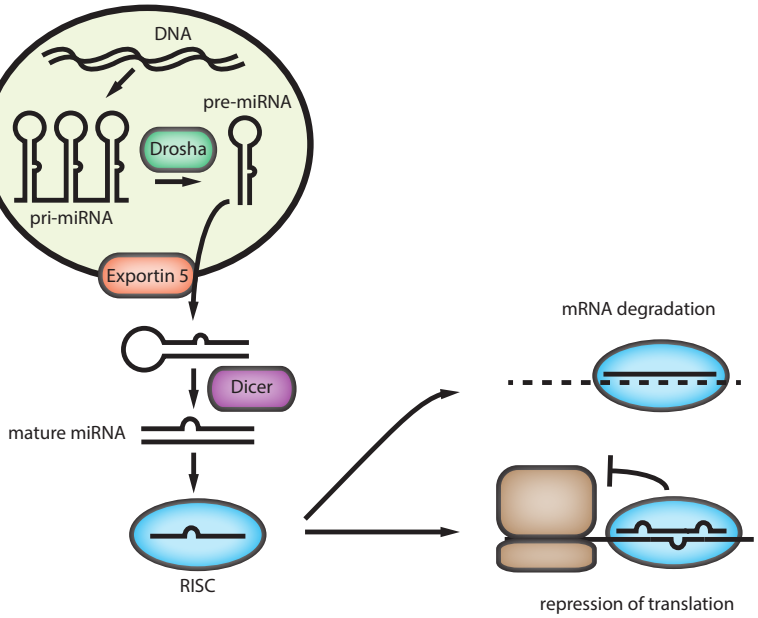


Figure 4.1: MiRNA biogenesis and post-transcriptional gene regulation. The miRNA-containing gene on the genome is transcribed by Polymerase II. The resulting pri-miRNA, which may consist of several miRNAs, is then cut by Drosha within the nucleus, which yields the pre-miRNAs. These pre-miRNAs are then exported to the cytosol by exportin 5 where the hairpin is removed by Dicer. This results in the double-stranded mature miRNA, which is then loaded in the RISC, where one of the two strands becomes degraded. The RISC then can bind to the target mRNA where it either leads to the degradation of the mRNA or the repression of its translation. The translational repression may also happen if the binding is not perfect. Figure outline partly adapted from [66].

nomenclature becomes more and more outdated [90].

4.2 Post-transcriptional gene regulation

The miRNA strand incorporated in the RISC serves as guide strand, which can bind specifically to the 3' UTR of target mRNAs. The bases 2-8 can form a Watson-Crick-paired, A-form double helix with complementary regions in the target mRNA [135]. This region is also referred to as the "seed region", which is the most important region for target recognition [99]. Bases 2-6 of the guide strand are fully exposed to the outside of the RISC and therefore play an even more important role in the binding of the RISC to the target mRNA [135].

miRNAs that have evolved from a common ancestor are commonly grouped in so-called *miRNA families* [56]. They are supposed to be similar in function, thereby usually sharing a common target set. The affiliation of miRNAs to the same family is usually indicated by the name.

The regulation of target genes by miRNAs can happen in different ways. In general, miRNAs are supposed to down-regulate the expression of the target genes, even though there is evidence for positive effects on gene expression [171]. The negative regulation happens either by the degradation of the bound mRNA or by the repression of its translation. The mRNA degradation demands for a near-perfect match of the guide strand and the target mRNA [106]. If such a perfect match is given, the RISC component Ago2 can act as ribonuclease. The ribonuclease activity of Ago2 then leads to the cleavage and the degradation of the target mRNA. For the translational repression, a perfect complementary of the guide strand and the target mRNA is not required [135]. Although the translational repression was initially supposed to be the most prevalent mode of target gene down-regulation, more recent studies report a predominant role of mammalian miRNAs in decreasing target mRNA levels [60]. Independently of the mechanism, the effect of miRNA down-regulation on the protein outcome often tends to be moderate [6]. Therefore, miRNAs usually act as fine-tuners of gene regulation [76].

4.3 Experimental identification of target relationships

Several experimental efforts were made to identify potential targets of miRNAs for example by the transfection or knock-down of miRNAs [104, 81]. The altered abundance of miRNAs is then supposed to directly influence the expression of the target gene expression, which can be measured for example by using PCR or microarrays. However, it is not clear if the miRNA has direct effects on the determined genes, since the set of altered genes in the expression analysis may also contain genes whose expression changes were caused by targets of the miRNA and not by the miRNA itself.

More recent approaches are focused on identifying mRNAs with bound

RISC, namely high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) [23] and photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) [61]. These techniques actually allow for the identification of direct targets. The basic principle of both techniques is an *in vivo* crosslinking of the mRNA with the Ago proteins by UV irraditaion. In case of PAR-CLIP, the cells are fed with 4-thiouridin to facilitate crosslinking. The mRNA-protein complex is immunoprecipitated by utilizing an anti-Ago monoclonal antibody. The sequences of the extracted mRNAs, which were bound to the Ago proteins, are then determined by RNA-seq.

4.4 Bioinformatic resources

After the identification of miRNA genes initially in *C. elegans*, efforts were made to characterize miRNAs bioinformatically [96, 91]. The goal was to predict hairpin structures among cloned small RNA sequences in order to classify them as miRNAs. Due to the rising number of identified miRNAs, comprehensive databases were built up, which make miRNA annotations easily accessible [57]. The advances in experimental and computational techniques allow for a constant increase in identified miRNAs. Especially the application of small RNA deep sequencing experiments identified a multitude of formerly unknown miRNA genes [90]. The most prominent database for miRNA annotations (miRBase) currently contains 24,521 microRNA loci from 206 species, processed to produce 30,424 mature microRNA products (v20, June 2013) [90].

Besides the identification and annotation of miRNAs, one of the most important fields in miRNA research is the prediction of target genes. A functional characterization of miRNAs is actually only possible with the knowledge of potentially regulated genes. Due to the short length of miRNAs and the fact that their regulatory mechanisms are still not fully understood, miRNA target prediction is a non-trivial task.

In order to make the information on experimentally validated miRNA-mRNA interactions available to the scientific community, several databases were set up containing interactions determined by various experimental approaches. In par-

ticular, these are TarBase [172], miRecords [180] and miRTarBase [71]. Another database, starBase, is specially focused on results from HITS-CLIP and PAR-CLIP experiments [182].

As mentioned above, experimental methods may often not lead to direct miRNA-mRNA interactions. Even the more advanced CLIP approaches have several limitations, especially with regard to the specificity [28]. All of the experimental techniques are furthermore conducted under specific conditions. This makes them prone to missing miRNA-mRNA interactions, which heavily depend on the experimental conditions like the tissue [3]. Computational target predictions are thus essential, especially for narrowing down potential miRNA-mRNA interactions for the experimental validation. Many approaches were proposed in order to overcome primarily the high amount of false positives arising from the short length of the binding sites and basically include some of these four common features, which are reviewed in [134]:

Seed match Only the nucleotides in the seed region (nucleotides 2-8) of the guide strand are able to form a stable association with the target mRNA [99]. Nearly all approaches therefore test for a sequence match between the seed region of the miRNA and the target mRNA.

Conservation miRNAs are often observed to be conserved across different species, especially within mammals, which indicates their important role in developmental processes [100]. Information on the conservation of binding sites is therefore usable for the prediction of "true" functional binding sites of miRNAs. Conservation analysis is usually restricted to the 3' UTR of the target mRNAs [44]. Even though conservation plays an important role, regulation of target genes by miRNAs through non-conserved binding sites is also observed frequently [6].

Free energy An important aspect in the functionality of binding sites is the free energy (or Gibbs free energy), which is an indicator for the stability of the miRNA-mRNA association. The prediction of the free energy can therefore be used to infer functional binding sites [185].

Site accessibility The stability of the mRNA secondary structure can influ-

ence the ability of forming a miRNA-mRNA association. Unwinding of the target mRNA secondary structure at the respective position is necessary in order to allow for the binding of a miRNA. Regions with less stable secondary structures are therefore preferred as functional binding sites [108]. The stability of the mRNA secondary structure can be predicted and utilized to determine functional target sites.

Among the whole set of target prediction approaches, the most prominent ones are TargetScan [99], miRanda [36] and DIANA-microT-CDS [128]. All of these three approaches include seed matching, conservation and free energy. DIANA-microT-CDS is based on a machine-learning approach that also takes further information like experimental data, binding site position and site accessibility into account.

Besides the sequence-based target prediction algorithms, mRNA targets of miRNAs are also predicted by joint gene expression analysis [103]. Both, *ab initio* and sequence-based predictions may infer false positive mRNA targets for individual miRNA regulators. Hence, several methods have been proposed, which combine both, sequence information and expression data. The first study to match transcriptome-wide miRNA and mRNA profiles aimed to validate putative miRNA-mRNA target relationships by proposing GenMiR++, which utilizes Bayesian networks for finding functional miRNA targets [73]. For selecting potential miRNA-mRNA relationships on a "1:1" basis, regression and correlation based approaches have been proposed [153, 101, 175]. Other methods have been proposed to select miRNA-mRNA relationships in a "n:1" fashion by using penalized regression analysis. This has been done by using the lasso penalty [111, 122] or by applying penalized Cox regression [20]. It has been shown that the introduction of an negativity constraint in the regression analysis allows for the identification of more experimentally validated interactions and better biological interpretation [111].

With the knowledge on target interactions of miRNAs, further functional analyses become possible. In order to provide insight into the functional role of miRNAs, a variety of bioinformatic resources were built up. These resources primarily deal with the association of miRNAs to functional categories like path-

ways [89], diseases [109] and other biological processes [152]. Other approaches use the network information of the miRNA-target network for the identification of miRNA modules, which include related miRNAs together with their target sets [19, 133]. These modules can then be functionally characterized independently.

4.5 Coordinated protein complex regulation

In this section, we will show our own bioinformatic analyses to reveal novel insights into the regulatory mechanisms of miRNA regulation. This section is part of a publication [155], which also comprises further analyses such as the identification of miRNA-regulated complexes. In this section we will focus on the global properties of clustered and co-expressed miRNAs with regard to their coordinated regulation of protein complexes, as it is the basis for the development of our miRNA-target identification approach (see Chapter 5).

Several components of protein complexes may be regulated simultaneously by a single miRNA or by several co-expressed miRNAs. Thus, we hypothesize that regulation of protein synthesis may be marginal for some of the miRNA targets. A cumulative effect for substantial phenotypic consequence may be achieved for those targets, which are members of the same protein complexes.

To test our hypothesis, we developed a robust computational framework to infer protein complexes, of which several distinct components are simultaneously regulated by either single miRNAs or co-expressed miRNAs. We applied the framework to characterize the protein complex networks, which consist of 722 experimentally verified protein complexes and protein-protein interactions from the CORUM database, which provides a resource of manually annotated protein complexes from mammalian organisms [151]. We furthermore assembled a miRNA-protein target network for 677 human miRNAs and 18,880 targets which are listed in the TargetScan database (v5.2, June 2011) [99]. The protein complex networks from the CORUM database are regulated by 677 miRNAs and 154 known miRNA clusters in humans.

Protein complexes and miRNA expression

Initially, we tested whether miRNAs which target different components of the same protein complex, are more likely to be co-expressed. For this purpose, we used a dataset from a previously published study where miRNA expression profiles were assessed across 26 different organ systems and cell types [95]. Hence, this dataset provides comprehensive insights into the properties of miRNA expression, which may be very specific for different cell types [178]. To assess the co-expression among the measured miRNAs, we retrieved the pairwise Pearson correlation coefficients of the miRNA expression profiles from the publication website. To test for statistical significance, we combined all pairwise correlation values obtained from the sets of miRNAs which significantly target the same complex. These correlation values were then compared to all other pairwise correlation values which are present in the dataset from [95] (Fig. 4.2).

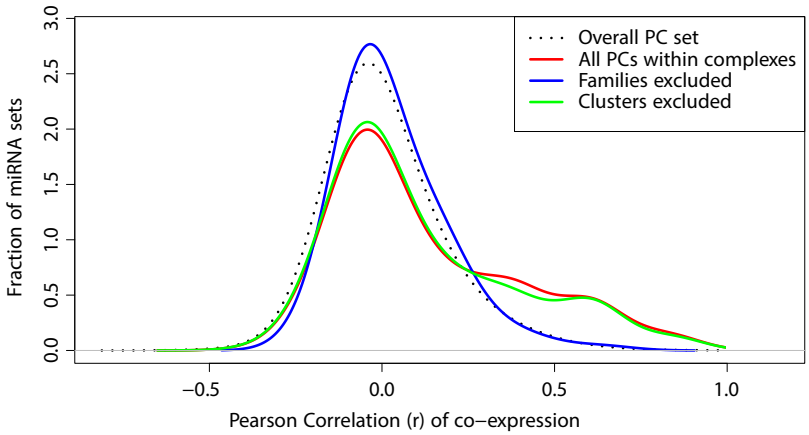


Figure 4.2: Pearson correlation distributions of miRNAs that target the same complex (red line) is plotted against the distribution of all observed Pearson correlation values (black dotted line). Also the distributions of excluded Pearson correlations of miRNAs from the same family (blue) and the same cluster (green) are plotted.

We performed a one-sided KS test for the two correlation value distributions and obtained a significantly ($p = 6.106 \times 10^{-24}$) higher co-expression within the sets of miRNAs that target the same complex. Since we are interested in coexpression of miRNAs that are not in one transcriptional unit, we also tested for increased correlation only for miRNAs of different miRNA clusters. Only a few (3.3%) of the correlated miRNAs were actually contained in one miRNA

cluster. The result remains highly significant ($p = 2.11 \times 10^{-18}$). Another bias of our results might occur due to fact that all miRNAs from one family must target the same complex since they target the same set of mRNA. We compared only miRNAs within one complex which belong to different families. The KS test resulted in a p-value of 5.8×10^{-3}. Taken together, our statistical test indicates that miRNAs targeting different components of a protein complex are significantly co-expressed.

Protein complex networks coordinately regulated by miRNA clusters

We systematically characterized the protein complex networks, which are simultaneously regulated by clustered miRNAs in 154 transcription units gained from miRBase (v17, April 2011) [90]. The interconnectivity of the target sets of the miRNA gene clusters was first assessed as follows: the number of PPIs between the target sets of each pair of miRNAs in the cluster was counted, and these values were compared to 1,000 randomly sampled sets of miRNAs. To avoid miRNA target prediction bias arising from redundant prediction of clustered miRNA family members, only targets of one family member were counted within each cluster. Comparing the observed number of interactions (Fig. 4.3) with the corresponding distributions of randomly sampled sets of miRNAs provides a strong indication that a significant fraction of miRNAs in clusters might coordinately regulate targets (p-value <0.02 Wilcoxon signed rank test).

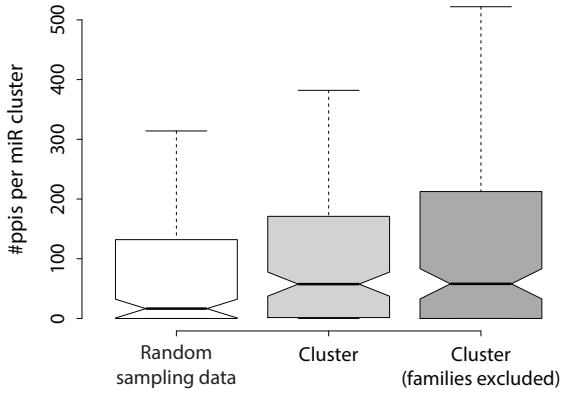


Figure 4.3: Boxplot for direct interactions of proteins targeted by N miRNAs within a cluster as compared to a null model of N randomly sampled miRNAs, respectively.

4.6 Conclusion

In this chapter, we introduced the concept of post-transcriptional gene regulation by miRNAs. We pointed out that there commonly exist relationships between miRNAs in form of clusters families. Clustered miRNAs are located in near proximity on the genome and often under the control of a common promoter. They are therefore frequently found to be co-expressed. We statistically analyzed the regulatory effect of these clustered miRNAs on protein complexes and could show that clustered miRNAs also have functional similarities. This functional similarity is illustrated by the fact that the genes among the predicted target sets are found to interact on protein level more often as compared to random target sets. On the other hand, we could also show that miRNAs targeting components of the same protein complex tend to be co-expressed. Hence, co-expression between miRNAs is directly linked to a coordinated gene regulation. This knowledge is important for the understanding of miRNA regulation and should also be taken into consideration when investigating regulatory relationships between miRNAs and genes.

Chapter 5

MicroRNA-target network construction

The main focus of miRNA research is the identification of regulatory influences on their target genes. Especially in specific experimental setups, we are interested in the functional role of miRNAs, which may be affected by a certain biological condition. One natural way to determine putative targets for the miRNAs of interest is the use of *in silico* target prediction approaches, which are typically based on sequence features. These approaches are prone to a large number of false positives and very unspecific for the given setup. In order to gain specific target relationships between miRNAs and genes, the integration of additional data is therefore essential. One resource, which is suitable for improving the quality of miRNA-target relationships, is gene expression data. These data can be used to select only observable miRNAs-target relationships out of the predicted target set. While simultaneous measurement of entire transcriptomes including mRNAs and miRNAs is relatively easy with high-throughput techniques, their integration is not straightforward. Several methods have been proposed to integrate miRNA and mRNA data for the identification of miRNA-target networks either based on correlation [153, 101, 175] or multiple linear regression analysis with lasso penalty [111, 122]. However, correlation analysis does not take into account a combined effect of miRNAs whereas lasso tends to exclude co-expressed miRNAs.

In this chapter, we introduce *miRlastic*, a method for the identification of data-driven miRNA-mRNA interactions by integrating *in silico* target predictions with paired miRNA and mRNA expression measurements. To meet drawbacks of existing methods, we use a multiple linear regression model with elastic net penalty and thus accounts for both, joint effects of several miRNAs on a common target and co-expression between miRNAs. We will point out why this co-expression is an important aspect in the joint-analysis and will show that miRlastic outperforms other common methods for the joint analysis of miRNA-mRNA data both on simulated and real data. With miRlastic we neglected the obvious and dominant modulator of mRNA expression, namely transcription factors. Thus, we performed additional analysis with a combination of miRNA and transcription factor data and provide evidence that miRlastic is preferable when compared to experimentally verified miRNA-target interactions. Finally, we use miRlastic to identify potential target genes of miRNAs, which are altered during adipogenesis. The resulting miRNA-mRNA regulatory network serves as a basis for the investigation of miRNA influences on the differentiation process.

The basic principle of the approach described in this chapter is discussed in:

- Swanhild U. Meyer, Katharina Stoecker, **Steffen Sass**, Fabian J. Theis, and Michael W. Pfaffl. Posttranscriptional regulatory networks: from expression profiling to integrative analysis of mRNA and microRNA data. *Methods Mol Biol*, 1160:165–188, 2014.

5.1 Dependencies of microRNA expression

To design a biologically-driven method to filter miRNA-mRNA profiles, we first had to interrogate typical expression characteristics. We therefore assessed the typical correlations of miRNA profiles alone.

Among miRNA expression profiles from the adipocyte differentiation dataset, we calculated pairwise Pearson correlations of all miRNAs predicted to have a common target (Fig. 5.1a). Besides a partitioning of the miRNAs in two groups of moderate correlation, we observe subgroups of very high correlation, for instance the miRNAs miR-30a/b/c/d/e and miR-320a/b/c. While the partitioning in the two groups can be explained by the biological variation arising from

the differentiation process (up- or down-regulation in adipocytes), the grouping into the smaller highly correlated subgroups exhibits strong functional similarities of the respective miRNAs. The two groups of miRNAs miR-30a/b/c/d/e and miR-320a/b/c both form a set of miRNAs coming from the same family. Furthermore, the sets of miRNAs miR-30b/d, miR-30c/e, miR-132/212 are located in the same miRNA cluster on the chromosome. All of these miRNAs are also highly correlated with each other, which corresponds to the expectation that functionally related miRNAs or miRNAs in high proximity on the chromosome tend to be co-expressed [155].

5.1.1 Correlation strength among a set of variables

To systematically analyze whether miRNA expression profiles are typically correlated, if they share a putative target, we next define a measure of correlation strength of a pairwise correlation matrix.

Let $\mathbf{X}$ be a matrix of miRNA expression measurements (x_{ik}) for n measured miRNAs, which is assessed across s observations with $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, s\}$. To summarize the pairwise correlation values, we introduce a measure of correlation strength $c(\boldsymbol{X})$ as

$$c(\boldsymbol{X}) = \frac{||\mathcal{R}(\boldsymbol{X})||_F}{\sqrt{(n^2 - n)/2}},$$

with the Pearson correlation matrix $\mathcal{R}(\boldsymbol{X}) = (\rho_{i_1 i_2}) = (r_{\boldsymbol{X}_{i_1 \cdot} \boldsymbol{X}_{i_2 \cdot}})$ for $i_1, i_2 \in i$. The Frobenius norm of the correlation matrix $\mathcal{R}(\boldsymbol{X})$ is calculated as

$$||\mathcal{R}(\boldsymbol{X})||_F = \sqrt{\sum_{i_1 < i_2} \rho_{i_1 i_2}^2}.$$

Note that only the upper triangular matrix with $(n^2 - n)/2$ elements is considered for the calculation of the Frobenius norm. As all elements of $\mathcal{R}(\boldsymbol{X})$ range between $[-1, 1]$, all values of $c(\boldsymbol{X})$ range between $[0, 1]$. The extreme values $c(\boldsymbol{X}) = 0$ and $c(\boldsymbol{X}) = 1$ indicate an entirely uncorrelated and perfectly (anti-)correlated set of miRNAs, respectively. When miRNAs are strongly correlated in groups while correlating little between groups (Figure 5.1a), the value of $c(\boldsymbol{X})$ decreases. In principle, $c(\boldsymbol{X})$ can be interpreted as a measure of correlation

strength: the more entries of miRNA expression profiles $\boldsymbol{X}$ are (anti-)correlated amongst each other, the higher $c(\boldsymbol{X})$.

5.1.2 Principles of collective miRNA regulation

We further evaluated the properties of collective miRNA regulation by assessing the (anti-)correlation strength across all miRNAs, which are predicted to target a common mRNA. We found that these sets of miRNAs are more correlated among each other than randomly sampled sets of miRNAs (Fig. 5.1b). As a result of coordinated targeting, individual miRNA subsets are highly correlated.

Our general assumption is that for a single mRNA, we end up with several miRNAs (1:n relationship) where the miRNA profiles themselves are correlated, which is a typical observation e.g. for clustered miRNAs [8]. Generally, correlation of miRNA profiles imply that they are again commonly regulated by some unknown process (Figure 5.1c). Even though a binding might be predicted for several of these co-expressed miRNAs, we may only observe an effect for a subset of these predicted target interactions. We thus aim to select functional predicted target interactions by simultaneously accounting for clusters of correlated miRNAs as effect from a number of unobserved regulatory layers. We furthermore assume that the target gene expression is modulated due to a combinatorial effect of several miRNAs, rather than a single miRNA, which has been suggested previously [116, 143].

5.2 Related methods

Previously proposed methods, such as the web application MAGIA [153], did not simultaneously capture all facets of miRNA dynamics: On the one hand, correlation-based approaches assume that expression of all miRNA regulators is well correlated together explaining mRNA target expression while other groups of correlated true miRNA regulators are neglected. On the other hand, L_1 regression-based approaches select one representative miRNA from each correlated group, which in turn allows to explain mRNA target expression. The web application TaLasso [122] is one example for a method, which is based on L_1 regression. Another approach for scoring putative miRNA-mRNA targets is

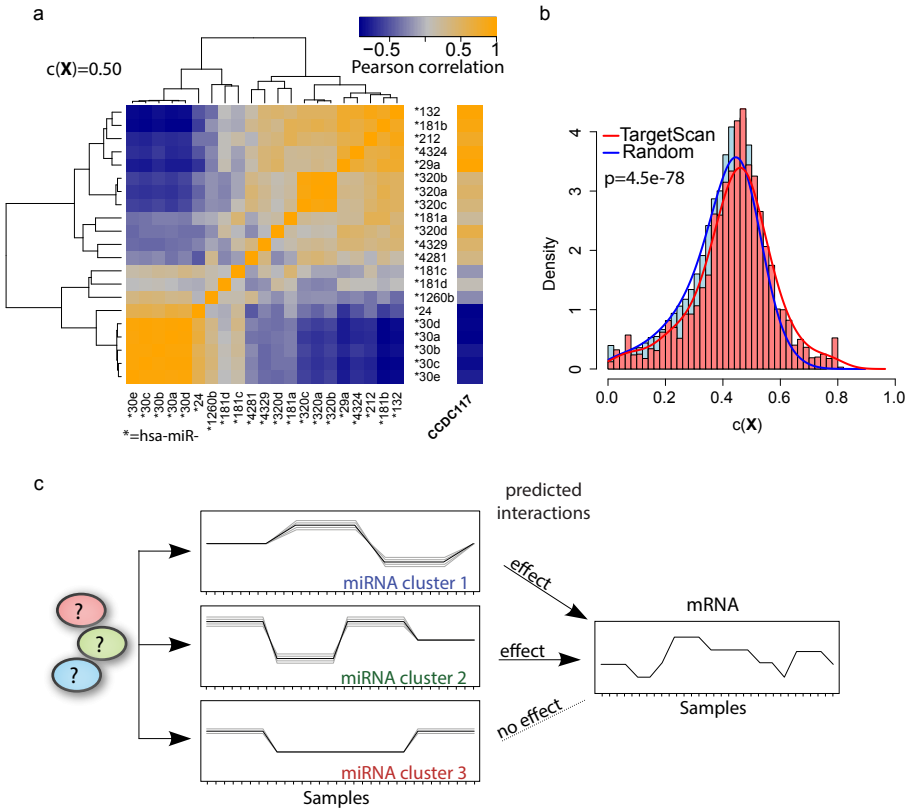


Figure 5.1: Collective effects of co-expressed miRNAs. **(a)** Pairwise correlation of putative, expressed miRNA regulators together with their target gene CCDC117 (obtained from the adipocyte differentiation dataset). The miRNAs are themselves clustered into several co-expressed groups. We observe a correlation strength of $c(\boldsymbol{X}) = 0.5$ for the given example. **(b)** The distribution of correlation strengths $c(\boldsymbol{X})$ of miRNA sets, which are predicted to target a common gene, (red curve and histogram) is higher than for randomly resampled miRNA-mRNA associations (blue, Wilcoxon rank sum test has $p = 4.5 \times 10^{-78}$) in the adipocyte differentiation miRNA expression dataset. **(c)** Schematic drawing of miRNAs co-expressed in clusters induced by an yet unknown regulatory layer. Only several of the putative miRNA regulators are finally collectively regulating mRNA expression.

based on Bayesian inference and is implemented in GenMiR++. This method ranks the miRNA-target interactions according to the calculated score and then selects a set of validated interactions via an arbitrary threshold of e.g. 50% of the input interactions.

A detailed overview of methods for the joint analysis of miRNA and mRNA expression data is given in [123].

In addition, miRNA expression has been included in the construction of

miRNA-transcription factor regulatory networks. Besides correlation-based approaches for setting up these kind of networks [157], regression analysis using the elastic net penalty has been used to combine miRNA and transcription factor information [10].

5.3 miRlastic for miRNA-target networks

We developed a novel approach, called miRlastic, to detect data-specific miRNA-mRNA targeting relationships inspired by biological principles. Beyond sole correlation- or standard regression-based models, we use the elastic net technique to identify all associations which are explained by the measured expression values.

By first clarifying our biological aims, the methodological implementation will be naturally explained. We aim to identify all and only those miRNA-mRNA targeting relationships, which are biologically sound with respect to the underlying experimental setup. For example, different cell types often employ a condition-specific miRNA-mRNA regulatory network required to fulfill its function. By integrating experimental data to filter putative miRNA-mRNA target predictions, we can account for the situative regulatory mechanisms.

Our novel method miRlastic filters those miRNAs out which are commonly and selectively regulating mRNA expression values. By trying to best explain mRNA expression by putative miRNA regulators, we have to account for both, additive effects of several miRNAs on the target mRNA as well high correlation between miRNA profiles. We use a multiple regression approach with an elastic net penalty to best balance feature selection without arbitrary thresholding. Our proposed method allows to derive biologically sound miRNA-mRNA (in a "n:1"-manner) relationships.

5.3.1 Preliminaries

The datasets $\mathbf{Y}$, $\mathbf{X}$ contain mRNA and miRNA measurements (y_{jk}) and (x_{ik}), respectively. Both of them are simultaneously assessed across s observations, typically the biological samples, with $k \in \{1, \ldots, s\}$, $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$. n and m are the number of measured miRNAs and mRNAs, thus,

$\mathbf{Y} \in \mathbb{R}^{m \times s}$ and $\mathbf{X} \in \mathbb{R}^{n \times s}$, respectively.

We denote the regulatory relationships between miRNA regulators and their putative mRNA targets as a bipartite graph. A bipartite graph has two disjoint node sets. Here the miRNA and the mRNA both constitute the node in the graph G. The bipartite graph captures all putative miRNA-mRNA relationships in $G = (V^{miR}, V^{mR}, E)$ with disjoint clusters of two node sets as V^{miR} and V^{mR}. The set of all mRNAs are the nodes listed in $V^{mR} = \{v_1^{mR}, \ldots, v_m^{mR}\}$, and likewise the miRNA are the nodes in $V^{miR} = \{v_1^{miR}, \ldots, v_n^{miR}\}$. The set of edges $E = \{e_1, ..., e_z\}$ connect nodes from V^{miR} to V^{mR} as $e_l = (v_u^{miR}, v_w^{mR})$ with $l \in \{1, \ldots, z\}$.

Edges e_l are exactly those yielded by target prediction algorithms, which are subjected to data-based filtering. G is validated with miRlastic and yields $G' = (V^{miR}, V^{mR}, E')$ with $E' \subseteq E$.

5.3.2 miRNA-mRNA models

Observing the miRNA-mRNA regulator-target graph G from the perspective of a single mRNA j, it is targeted by a set of miRNAs with indices $i^* = \{i | \exists (v_i^{miR}, v_j^{mR}) \in E\}$ connecting n^* miRNA nodes $v_{i^*}^{miR}$ with mRNA node v_j^{mR}. We refer to the observations of one mRNA j as $\mathbf{y}_j$ and its associated miRNA observations as $\mathbf{X}(j) = \mathbf{X}_{i^*}^T$. The 1-dimensional vector $\mathbf{y}_j$ and the $s \times n^*$-dimensional matrix $\mathbf{X}(j)$ are response and predictors of a corresponding regression model, respectively, which is given by

$$\mathbf{y}_j \sim \beta_{j0} + \mathbf{X}(j)\boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j.$$

with normally distributed error $\boldsymbol{\epsilon}_j \sim \mathcal{N}(0, \sigma)$, parameters $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jn^*})$ and the intercept β_{j0}.

5.3.3 miRNA-mRNA feature selection

Several penalization techniques have been proposed to shrink the regression coefficients $\boldsymbol{\beta}$ by imposing a penalty on their size (see Chapter 2.4.1). With ridge regression, a L_2 penalty is applied, which has only nonzero $\boldsymbol{\beta}$ coefficients. Thus, ridge regression does not provide any feature selection and maintains all

predictors in the model. With lasso, a L_1 penalty is implemented yielding many coefficients equal to zero and only a subset not equal to zero. Subsequently, lasso has a sparse solution performing feature selection on $\mathbf{X}$. For highly correlated predictors, only the strongest predictor will have non-zero coefficients with lasso [43].

For modeling miRNA-mRNA relationships, we want the regression model not only to maintain correlated miRNA predictors in the model but also to have a feature selection option efficiently excluding miRNAs with no effect on the mRNA response. The elastic net penalty was introduced to balance between L_1 and L_2 penalties [190].

We propose a penalized regression model to systematically evaluate all putative miRNA-mRNA interactions. microRNA target predictions G serve as putative interaction graph to be validated by given transcriptome expression measurements, i.a. given graph G, for each mRNA j a penalized regression model is calculated. In order to allow only for the down-regulation of mRNA abundances, we introduce a negativity constraint on the coefficients $\boldsymbol{\beta}_j$. The values of $\boldsymbol{\beta}_j$ then serve as an indicator for the strength of regulation for each individual miRNA.

Assume that the expression values $\mathbf{y}_j$ are standardized with mean 0 and standard deviation 1 as well as the columns of $\mathbf{X}(j)$. The miRNA coefficients $\boldsymbol{\beta}_j$ for a mRNA j are estimated by the following optimization:

$$\hat{\boldsymbol{\beta}}_j = \arg\min_{\boldsymbol{\beta}_j} |\mathbf{y}_j - \mathbf{X}(j)\boldsymbol{\beta}_j|,$$
$$\text{subject to } (1-\alpha_j)\frac{1}{2}||\boldsymbol{\beta}_j||_2^2 + \alpha_j||\boldsymbol{\beta}_j||_1 \leq t_j \text{ and } \boldsymbol{\beta}_j \leq 0,$$

with α_j denoting the elastic net mixing parameter for mRNA j with $0 \leq \alpha_j \leq 1$. Note that we do not have to estimate an intercept as it is supposed to be zero for scaled variables and response. The regularization parameter t_j can be chosen using a 10-fold cross-validation procedure. The non-zero entries of $\hat{\boldsymbol{\beta}}_j$ are then considered as the evaluated edges of the input network G and as a result, miRlastic returns the validated miRNA-mRNA relationships as G' gathered from all models for $\mathbf{y}_j$ and non-zero coefficients in $\boldsymbol{\beta}_j$ of corresponding miRNA predictors $\mathbf{X}(j)$ for $j = \{1, \ldots, m\}$.

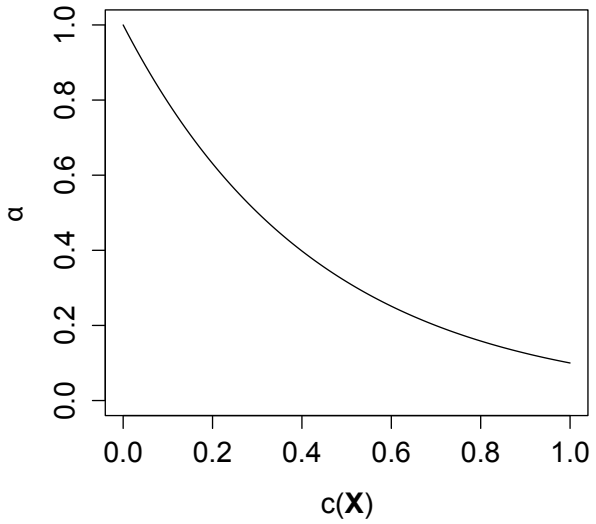


Figure 5.2: Choice of the elastic net parameter α_j. Given the measure of correlation strength $c(\boldsymbol{X}(j))$ among a set of miRNAs collectively targeting a mRNA j, we determine the value of α_j by $\alpha = 10^{-c(\boldsymbol{X}(j))}$. Lower values of α_j are preferred over high values. In the case of perfectly uncorrelated miRNAs ($c(\boldsymbol{X}(j)) = 0$) we perform native lasso ($\alpha_j = 1$). If the miRNAs are perfectly correlated ($c(\boldsymbol{X}(j)) = 1$), we aim to keep the correlated variables within the model ($\alpha_j = 0.1$). However, we always want to perform feature selection, which would be prevented by ridge regression. Therefore, we never set α_j to zero.

To finally tune the elastic net penalty chose to adjust α with respect to the potentially expected fraction of correlated predictor groups. Another possibility to tune α is cross-validation, but results were not satisfactory and the computation is very time-consuming. Thus, as an educated guess of α to balance the L_1 and L_2 penalties, we make use of the previously introduced measure of miRNA correlation strength $c(\boldsymbol{X})$.

Let $\boldsymbol{X}(j)$ be the expression matrix of miRNAs, which are predicted to simultaneously target a common mRNA j with expression profile $\mathbf{y}_j$. The parameter α_j of the elastic net regression model of $\mathbf{y}_j$ given $\boldsymbol{X}(j)$ is then defined as $\alpha_j = 10^{-c(\boldsymbol{X}(j))}$. This allows for an unbiased parameter tuning whereas lower values of α_j are slightly preferred. However, we do not want to set α_j too low, since we never want to prevent the feature selection procedure by performing ridge regression. Therefore, the choice of α_j in the given way is a good trade-off (see Fig. 5.2). The algorithm for the construction of the evaluated miRNA-mRNA network G' is summarized in Algorithm 5.1.

Input: Predicted miRNA-target network $G = (V^{miR}, V^{mR}, E)$, miRNA expression matrix $\mathbf{X}$, mRNA expression matrix $\mathbf{Y}$.
Result: Evaluated miRNA-target network G'.
n:=Number of mRNAs in G;
m:=Number of mRNAs in G;
Initialize $G' = (V^{miR}, V^{mR}, E')$ with $E' = \emptyset$;
for *j=1:m* **do**
  $i^* = \{i | \exists (v_i^{miR}, v_j^{mR}) \in E\}$;
  $\mathbf{X}(j) = \mathbf{X}^T[, i^*]$;
  $\alpha_j = 10^{-c(\mathbf{X}(j))}$;
  $\mathbf{y}_j = \mathbf{Y}[j,]$;
  $\hat{\boldsymbol{\beta}}_j = \arg\min_{\boldsymbol{\beta}_j} |\mathbf{y}_j - \mathbf{X}(j)\boldsymbol{\beta}_j|$,
   subject to $(1-\alpha_j)\frac{1}{2}||\boldsymbol{\beta}_j||_2^2 + \alpha_j||\boldsymbol{\beta}_j||_1 \leq t_j$ and $\boldsymbol{\beta}_j \leq 0$;
  Determine best t_j via 10-fold cross-validation;
  for $i' = \{i^* | \hat{\boldsymbol{\beta}}_{ji^*} \neq 0\}$ **do**
    $E' = E' \cup (v_{i'}^{miR}, v_j^{mR})$ with weight $\hat{\boldsymbol{\beta}}_{ji'}$;
  end
end
return *G'*

Algorithm 5.1: Construction of the evaluated miRNA-mRNA network G' using miRlastic.

We here use TargetScan 6.2 [99], but any prediction method may be used or even a combination of several. MiRlastic is implemented as an R package using the elastic net implementation from the *glmnet* R package [43].

5.3.4 Evaluation on synthetic data

Quality Measure

By counting the number of incorrectly selected miRNAs (false positives, FP) and the number of missed correct miRNAs (false negatives, FN) for several runs, we can compute the F_1 measure based on precision and recall as

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

With precision $= TP/(TP + FP)$ and recall $= TP/(TP + FN)$ of any related confusion matrix comparing actually classes (true and false) to any classification results (positive and negative).

Synthetic data

In order to assess the performance of our approach, we built up a test environment. In each test run, a set of synthetic miRNA and mRNA expression values was generated adapting to biological features. We modeled a set of 25 miRNAs $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{25})$ with expression values $\boldsymbol{x}_i \sim \mathcal{N}(0, 1)$ corresponding to the set of miRNAs predicted to target a common gene. Since we assume a coordinated regulation among this set of miRNAs, we furthermore model 4 expression profiles of unknown regulatory factors $\{\mathbf{h}_1, \ldots, \mathbf{h}_4\}$ with $\boldsymbol{h}_i \sim \mathcal{N}(0, 1)$, which are assumed to target a certain subset of the predicted miRNAs. We therefore randomly pick one of these factors j for each miRNA i and introduce a correlation of $corr(i, j) = 0.99^{|i-j|}$ between the profiles $\boldsymbol{x}_i$ and $\boldsymbol{h}_j$ (Fig 5.3a). Hence, we

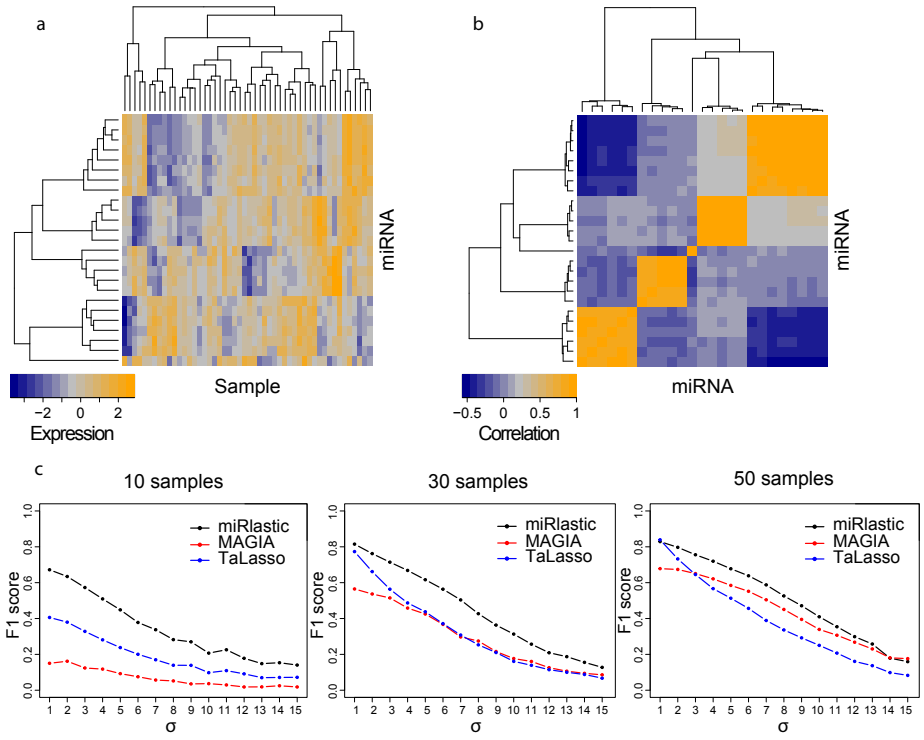


Figure 5.3: Application to synthetic data. **a.** Heatmap illustrating a set of randomly generated synthetic miRNAs with 30 samples. **b.** Heatmap of pairwise correlations between the generated miRNAs. **c.** Success-rates (measured F_1) of all algorithms across varying sample numbers and noise levels to recover the true synthetic miRNA-mRNA associations.

obtain groups of miRNAs that are highly correlated among each other while

the correlation to miRNAs in different groups is rather low (Fig 5.3b). As illustrated in Figure 5.1, this characteristics is in accordance with the biological data.

The whole set of 25 generated miRNA profiles was then randomly partitioned into 15 profiles $\hat{\mathbf{X}} = (\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_{15})$ corresponding to the actual regulators of the mRNA and another set of 10 miRNAs that did not have an influence on the mRNA expression. Assuming a repressive effect of the targeting miRNAs, the mRNA expression profile y was then generated as

$$y = \sigma\epsilon + \sum_{i=1}^{15} -\hat{\boldsymbol{x}}_{\boldsymbol{i}}$$

with $\epsilon \sim \mathcal{N}(0, 1)$ corresponding to the noise arising out of biological reasons or from measurement errors. In order to evaluate our method on different noise levels, we performed the test runs with different magnitudes of σ. To check whether miRlastic is competitive with other common methods, we also applied correlation analysis and lasso on the generated data. For correlation analysis, a synthetic miRNA was considered as true regulator of the corresponding mRNA, if the adjusted p-value (Bonferroni corrected) of negative Pearson correlation was below 0.05, which corresponds to analysis workflow of MAGIA [153]. For lasso, we used an approach, which is basically identical to miRlastic. However, the value of α was set strictly to 1. This kind of analysis is similar to the approach, which is implemented in TaLasso [122].

The whole test procedure was repeated three times where each time a different number of samples was determined for the synthetic miRNAs and mRNA, namely 10, 30 or 50 samples. Every test procedure consisted of 500 runs for 15 different values of the noise level σ. For each noise level, the amount of arising false positives and false negatives was recorded for every method. Finally, the F_1 measure was calculated for the respective noise level.

Performance on synthetic data

We observe a good performance of miRlastic with regard to the arising false positives and false negatives in comparison to the other methods for each of the three test procedures (Fig. 5.3c). It outperforms correlation analysis and

lasso especially for low sample numbers. This indicates that miRlastic is able to provide reasonable results even when applied on datasets with low sample numbers. Low sample numbers are an important issue in biological research since the preparation of samples using large-scale techniques can still become costly. Especially for combined expression data, the number of matched samples can be low since measurements have to be performed twice. Correlation analysis performs weakly for low sample numbers whereas the results improve for high sample numbers. Lasso performs well for medium and high sample numbers only for low noise level indicating a low robustness against noisy observations.

In summary, we can show that miRlastic is able to reliably identify true regulators with high specificity and sensitivity in a biologically reasonable synthetic test environment. It outperforms other methods, since it has a high tolerance against noisy observations and low sample numbers.

5.4 miRlastic on adipogenesis data

Given the set of differentially expressed miRNAs in the dataset, we applied miRlastic on the expression measurements of these miRNAs combined with the mRNA expression measurements for potential targets. The underlying target predictions were downloaded from TargetScanHuman (version 6.2) [99] by only considering conserved target sites for conserved miRNAs families. Target predictions were available for 20 of the 30 differentially expressed miRNAs. Especially for miRNA* target predictions were unavailable in TargetScan. MiRNA interactions were predicted for 14,242 genes in our dataset whereas 3,498 of them were predicted to be targeted by at least one of the differentially expressed miRNAs. Overall, we used 9,995 target interactions in combination with the respective miRNA and mRNA expression values as input for miRlastic.

The miRlastic approach then selected 4,020 miRNA-mRNA interactions out of the given target predictions (Fig. 5.4).

We observe a high amount of mRNAs which are jointly targeted by the miR-30 family. Interestingly, this set contains the runt-related transcription factor 2 (RUNX2). RUNX2 is the key factor in the formation of osteoblasts, which are similarly to adipocytes derived from mesenchymal stem cells [80] (see

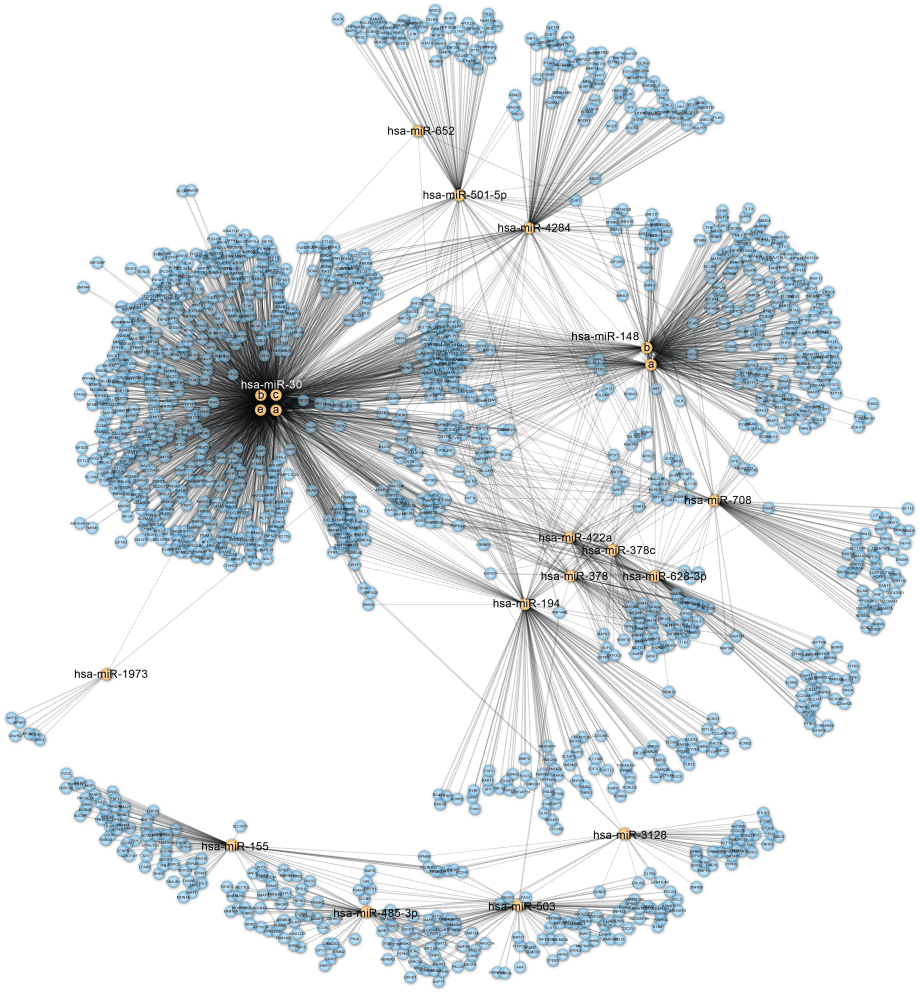


Figure 5.4: MiRNA-mRNA network for the adipogenesis data obtained by miRlastic. Based on the combined expression data, 4,020 miRNA-mRNA interactions were selected from the 9,995 given target predictions. The network contains nodes for the differentially expressed miRNAs (orange) and for the targeted genes (blue). The edge width corresponds to the respective coefficient in the regression model and indicates the strength of the interaction. We provide an interactive representation of this network at http://icb.helmholtz-muenchen.de/sass/adipo/ .

Chapter 2.6). The fact that the miR-30 family is up-regulated during adipogenesis, thereby targeting RUNX2, indicates a potential role of this family in this lineage-decision. In fact, the negative regulation of osteoblast differentiation through targeting of RUNX2 by the miR-30 family has already been shown [179].

5.5 Comparison with transcription factor model

Even though miRNAs are known to play essential roles in almost all cellular processes [30], their regulatory effect on gene expression is generally moderate in comparison to gene expression changes driven by transcription factor (TF) activity [6]. We could therefore argue that the integration of transcription factor information might lead to a more precise prediction of true miRNA-target relationships, as our model would account for the underlying mechanisms of gene regulation more appropriately. In fact, several studies already dealt with the integration of both miRNA and TF information for the construction of regulatory networks where mRNA expression was considered to describe the TF activity [157, 10]. The aim of these studies was generally to analyze network properties of the inferred target networks and to identify potential key regulators. A systematic analysis of explainable variance by miRNAs in comparison to TFs has not been done before.

We therefore aim to evaluate whether the addition of TF information can enhance the quality of the resulting miRNA-target network. For this purpose, we extended our prior miRNA-target network obtained from TargetScan by TF-target interactions. We downloaded experimentally verified TF binding sites which were obtained by chromatin immunoprecipitation using specific antibodies for the transcription factors followed by sequencing of the precipitated DNA (ChIP-seq) [82]. This dataset was retrieved from the ENCODE database [35] and comprises 1,582,526 genomic binding sites for 161 TFs, which were investigated in 25 cell types. We overlapped these binding sites with the genomic positions of predicted human promoters determined by the Genomatix PromoterInspector [156] and finally obtained a total set of 971,933 TF-target interactions. We use respective mRNA expression profiles as a proxy for transcription factor activity. We then performed miRlastic on the newly generated target matrix consisting of both, miRNA-target and TF-target relationships and denote this approach as *miRlasticTF*. To account for the different regulatory properties of miRNAs and TFs, we redefined the optimization problem for the feature selection procedure (see Chapter 5.3.3) such that we only constrain the coefficients of the miRNAs to be negative. The coefficients of the TFs were not constrained, thus, can be either positive or negative. In other words, we model that TFs can

act either as activators or inhibitors of gene expression, whereas we only expect a repressive effect of miRNAs. As above, we analyzed the miRNA and mRNA data from the adipogenesis dataset. Here, we use the mRNA data to assign expression profiles to the TFs. In order to prevent a perfect fit in the regression model, we omit the case that TFs can regulate themselves.

The application of miRlasticTF on the given setup yielded a set of 149,815 evaluated TF-target interactions and a set of 1,944 miRNA-target interactions. If we compare this result with the miRNA-target network determined by miRlastic solely for miRNAs (Chapter 5.4), we observe that miRlasticTF yields only about half of the miRNA-target relationships as compared to miRlastic (Fig. 5.5A). Notably, more than 95% of the miRlasticTF miRNA-target relationships could also be detected by miRlastic.

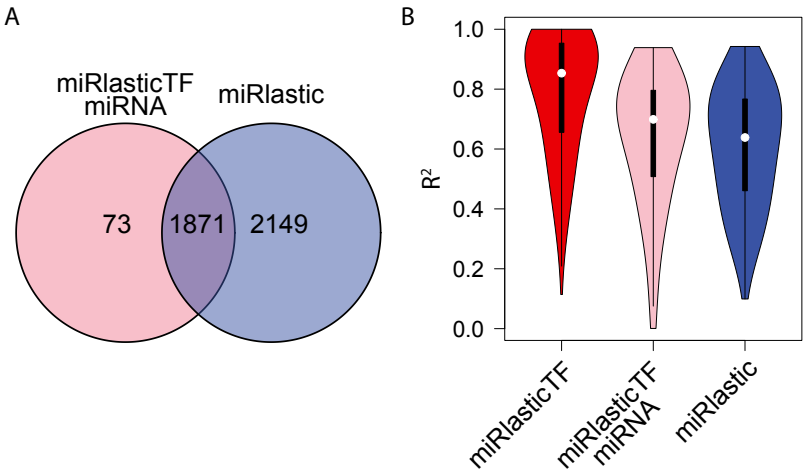


Figure 5.5: **(A)** Number of overlapping miRNA-target relationships yielded by miRlastic and miRlasticTF. More than 95% of the miRlasticTF relationships could also be determined by miRlastic, whereas miRlasticTF yields only about half of the miRlastic relationships. **(B)** Violin plot illustrating the kernel density estimation of the probability density function of explained variance R^2, where the median is indicated by a white dot. The value R^2 is assessed for all regression models of evaluated miRNA-target relationships by miRlastic (blue) and miRlasticTF (pink). The overall R^2 of miRlasticTF (red) indicates the amount of variance that could be explained in all regression models including both, miRNA and TFs.

To next dissect contributions of miRNA or TFs to their targets, we calculated the explained variance of all target mRNAs by the corresponding miRNAs and TFs, which were evaluated by miRlasticTF. We therefore calculated for each

mRNA the explained sum of squares (ESS) for the model, which solely included the identified miRNAs, as well as for the complete model including TFs. The ESS is defined as $\text{ESS} = \sum_{k=1}^{s} (\hat{y}_k - \bar{y})^2$, where $\hat{y}$ is the predicted response variable of length s and $\bar{y}$ the mean of the given response variable. To determine how much of the total variance could be explained, we compared the ESS to the total sum of squares (TSS), which is defined as $\sum_{k=1}^{s} (y_k - \bar{y})^2$. The amount of explained variance is expressed as $R^2 = \text{ESS/TSS}$.

By comparing the magnitude of R^2 for miRlasticTF with miRNAs only to the magnitude of R^2 for the overall miRlasticTF results (Fig. 5.5B), we observe an increase of R^2 of only about 0.15 if the TFs are included in the model. In addition, we can not determine whether increase of R^2 is only observable due to a naturally higher correlation among the genes or by a regulatory relationship. This correlation can be related to the fact that genes are often jointly regulated and that they were measured on the same platform. Since we furthermore know that transcription factor activity is predominantly regulated through post-translational modifications like phosphorylation [78], these results could be prone to errors. Interestingly, the magnitude of R^2 for miRlastic is equivalent to the magnitude of R^2 for miRlasticTF with miRNAs only. If we neglect transcription factors for the identification of miRNA-target relationships, we can thus still sufficiently well explain our regulatory model.

5.6 Evaluation using experimental data

In the previous section, we evaluated whether our miRlastic approach is reasonable if applied on miRNA data only, thereby neglecting the influence of transcription factors. The even more interesting aspect is the evaluation of the two results with regard to experimentally verified miRNA-target relationships. Therefore, we confirmed the interactions with experimentally validated interaction data retrieved from starBase [182] and compared the performance to other existing methods.

The aims of our validations are twofold: we want to demonstrate that the identified set of interactions using combined expression data is significantly enriched by experimentally validated interactions in comparison to TargetScan

only. Furthermore, we want to show that this enrichment is considerably higher for miRlastic as compared to other common methods. For this purpose, we selected all miRNA-mRNA interactions determined by HITS-CLIP or PAR-CLIP (high stringency) from the starBase database [182] which overlapped with our set of tested miRNAs and genes also predicted by TargetScan. In total, 4,039 previously validated target interactions between the overlapping set of 15 miRNAs and 2,627 genes were derived. We found that 2,049 out of 3,322 target interactions between these overlapping miRNAs and genes in our resulting network could be already experimentally verified. We then applied Fisher's exact test in order to determine whether this proportion of validated interactions in our network is significantly higher than the proportion of validated interactions from TargetScan. The test yielded a highly significant p-value of $p = 5.7 \times 10^{-21}$ (Fig. 5.6A).

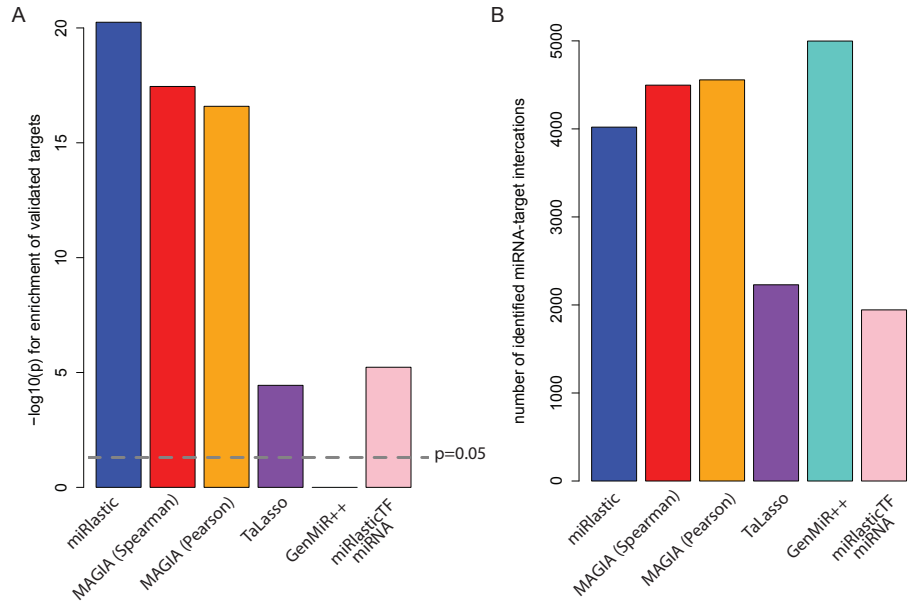


Figure 5.6: Validation of miRlastic using experimental data. MiRlastic (blue), MAGIA using Spearman (red) and Pearson (orange) correlation coefficient, TaLasso (purple), GenMiR++ (turquoise) as well as miRlasticTF (pink) were applied on the adipogenesis data using target predictions from TargetScan [99]. **(A)** The bars indicate the p-value resulting from the application of Fisher's exact test to determine the over-representation of experimentally validated target relationships from starBase [182] in comparison to the TargetScan predictions. **(B)** Total number of identified miRNA-target interactions for each of the applied methods.

MiRNA-target networks were also generated by applying correlation analysis (MAGIA [153]) and lasso (TaLasso [122]) as described above (Section 5.3.4). By applying Pearson correlation analysis, we identified 4,557 miRNA-mRNA relationships (Fig. 5.6B). The usage of the Spearman correlation coefficient yielded a network of 4,497 miRNA-mRNA relationships. The network, which was determined by lasso, contained 2,229 target interactions. In addition, we applied GenMiR++, which yielded 4,998 miRNA-mRNA relationships. We then again used Fisher's exact test to test whether the set of verified interactions are enriched in these networks. We obtained a p-value of $p = 2.6 \times 10^{-17}$ for the Pearson correlation network, $p = 3.5 \times 10^{-18}$ for the Spearman correlation network, $p = 3.6 \times 10^{-5}$ for the TaLasso network, $p = 1$ for GenMiR++ and $p = 5.9 \times 10^{-6}$ for miRlasticTF (Fig. 5.6A).

These results indicate that miRlastic is able to identify a higher fraction of previously validated target predictions as compared to the other methods. The two correlation approaches also performed well. The fraction of experimentally validated interactions in the lasso network was clearly lower. However, this fraction is still significantly higher as compared to TargetScan. In the case of GenMiR++, no significant overrepresentation of experimentally validated interactions could be observed. In addition, we show that the results of miRlasticTF are not as highly enriched as for miRlastic on miRNA only. We can therefore argue that the integration of TF information in combination with mRNA expression is not a suitable approach for the improvement of our approach. Furthermore, the degree of enrichment does not directly depend on the number of identified miRNA-target interactions (Fig. 5.6), as we observe a higher enrichment in miRlastic as for the MAGIA and GenMiR++ results even though the total number of interactions is less. On the other hand, the degree enrichment for miRlastic is also higher for miRlastic as for TaLasso and miRlasticTF, which yield a smaller number of target interactions.

5.7 Discussion and Conclusion

In this chapter, we introduced a method for the construction of miRNA-mRNA regulatory networks. These networks represent potential regulatory relation-

ships between miRNAs and genes based on sequence information and combined expression data. The aim of this approach is to validate *in silico* target predictions, which are known to be prone to a large number of false positives, and reduce them to a set of miRNA-mRNA interactions that can actually be explained by the expression data. Depending on the underlying experimental setup, we can therefore account for tissue- or condition-specific interactions, which can be used to investigate affected molecular processes with regard to post-transcriptional gene regulation.

Post-transcriptional gene regulation by miRNAs usually happens in a coordinated way, which results in the co-expression of functionally related miRNAs. We could show that co-expression between miRNAs, which are predicted to target a common gene, can actually be observed in the expression data. Hence, we optimized our method to account for correlated expression among miRNAs in order to capture the biological properties of miRNA regulation.

We evaluated our results by using mRNA expression data in addition to miRNA expression data. To fully unravel effective miRNA-mRNA interactions, we additionally need to consider the integration of protein expression data instead. This would allow to take both ways of miRNA regulation into account, mRNA degradation and translational repression. But even though the large-scale proteome measurement techniques have been improved remarkably during the last decade, the procedure for acquiring these measurements are still more laborious and less comprehensive than mRNA expression profiling experiments [4]. Since mammalian miRNAs are assumed to predominantly decrease the mRNA levels of their targets [60], the choice of mRNA expression measurements appears reasonable.

We could show that our approach performs well on both, synthetic and biological data and outperforms other methods, which are commonly used in miRNA research. Furthermore, we identified miRNA-mRNA relationships playing important roles in the context of adipogenesis, which is additional an indicator for valuable results.

Chapter 6

Functional characterization of miRNA-target networks

Even though many functional associations of miRNAs have been revealed, their full functional potential is still not exhausted. In the previous chapter we introduced miRlastic, which is able to infer a miRNA-mRNA regulatory network from target predictions and combined expression data (see Chapter 5) while the functional role of individual miRNAs was not yet elucidated.

In this chapter we introduce a local enrichment analysis (LEA) for networks generally consisting of a regulatory layer that is connected to set of genes. We want to use our approach to characterize this regulatory layer for which no functional annotation is available, whereas the genes can be mapped to functional groups from an ontology database. LEA is based on weighted bipartite networks, which can be derived e.g. from miRlastic (see Chapter 5). However, it is generally not restricted to this kind of application and may be applied on any bipartite network consisting of a node set that can be mapped to functional properties. Using the LEA approach, we can determine functional groups, which are locally enriched in the network, and identify these regions in order to allow for the inference of individual miRNA functions. We will apply LEA on the network, which was generated by the joint analysis of the miRNA and mRNA data from the adipogenesis study (see Section 5.4). We could identify processes that are specifically regulated by a subset of miRNAs. This enabled us to gain

insight into the functional role of miRNAs and also to determine joint functional properties of several miRNAs.

6.1 LEA: Local enrichment analysis

The primary goal of LEA is to identify regions within a miRNA-mRNA network which are strongly enriched for a certain biological process, thereby inferring information on the functional role of specific miRNAs. For this purpose, we can make use of functional gene annotations derived from databases such as GO [5] or KEGG [84]. These databases provide gene sets that are involved in specific processes or pathways denoted as functional groups. We assume that the genes in a locally enriched area are located in close proximity to genes assigned to the respective functional group. We thus use shortest paths as a basis to infer areas of local enrichment for a given functional group.

The concept of network proximity analysis has also already been used to identify regions in signaling pathways, which are specifically targeted by a set of miRNAs [89]. We want to translate this concept of network proximity to miRNA-target networks as follows: if we expect a term to be targeted by a special subset of miRNAs in the miRNA-target network, we assume a certain proximity for the genes assigned to that process. We measure this proximity by calculating the shortest paths between the genes in the network. This information can then be used to evaluate whether the given arrangement of nodes assigned to a certain process occurs by chance or not. The use of weighted edges for the calculation of the shortest paths furthermore accounts for the strengths of miRNA-mRNA relationships, which emphasizes the relevance of strongly regulated genes associated with the process.

Current methods for the functional analysis of miRNA-target networks primarily account for the whole set of regulated targets [89, 109]. Other methods partition the network into functional modules, which are then tested for functional enrichment [19, 133]. However, these methods demand for an *a priori* specification of module numbers and do not take the whole network structure into account. In addition, the strength of regulation is not considered for the determination of the modules.

6.1.1 Shortest distances between targets

The input of LEA is a directed bipartite graph $G = (V^{miR}, V^{mR}, E)$. The set of edges $E = \{e_1, \ldots, e_z\}$ corresponds to the potential regulatory relationships between the nodes from V^{miR} and V^{mR} as $e_l = (v_u^{miR}, v_w^{mR}) \in E$ with $l \in \{1, \ldots, z\}$. The edges and edge weights are represented as a matrix $\mathbf{W} = (w_{ij})$ with z non-zero entries. In the case of a miRlastic network, this weight corresponds to the scaled negative regression coefficient obtained by the elastic net approach. Hence, all non-zero entries of $\mathbf{W}$ must be smaller than zero.

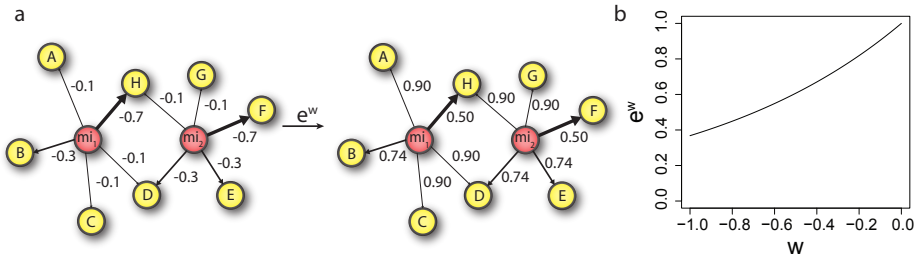


Figure 6.1: Weight transformation of miRNA-mRNA networks. (**a**) The negative edge weights w are indicated by the edge width and correspond to the regression coefficients determined by miRlastic. We transform these weights by e^w. (**b**) The transformation has three effects: edge weights become positive, negative coefficients become small and vice versa and highly negative edge weights are restraint to prevent a biased calculation of shortest paths.

The edge weight w_{ij} of the edge (v_i^{miR}, v_j^{mR}) corresponds to the strength of regulation between miRNA i and gene j (Fig. 6.1a). For easier interpretation of edge weights, we initially transform the non-zero entries of $\mathbf{W}$ such that we obtain a new matrix $\tilde{\mathbf{W}} = (\tilde{w}_{ij})$ with

$$\tilde{w}_{ij} = \begin{cases} e^{w_{ij}} & \text{if } w_{ij} < 0 \\ 0 & \text{otherwise.} \end{cases}$$

This transformation yields positive edge weights where highly negative coefficients become small and vice versa (Fig. 6.1b). In addition, we do not want to overestimate the influence of highly negative weights, which would bias the calculation of shortest path. We therefore restrain highly negative edge weights. The effect of transformation for the adipocyte miRNA-mRNA network gener-

ated by miRlastic (see Chapter 5.4) is illustrated in Fig. 6.2.

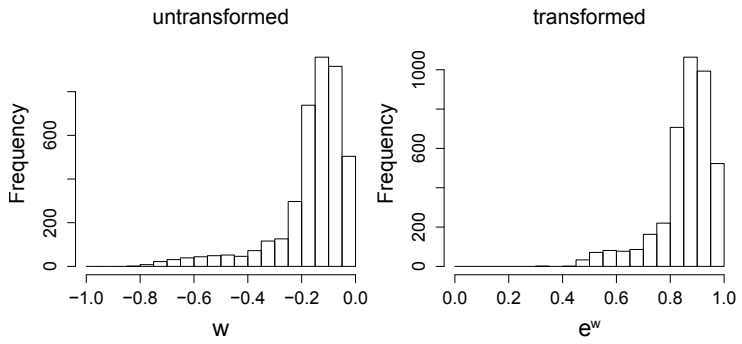


Figure 6.2: Histogram of edge weights for the adipocyte miRNA-mRNA network generated by miRlastic showing the effect of transformation. The x-axis is reversed and negative values become positive. Highly negative values are restrained after the transformation.

A path between two gene nodes v_a^{mR} and v_b^{mR} is defined as $P(a,b) = (v_1^{mR}, \ldots, v_p^{mR})$ with $v_1^{mR} = v_a^{mR}$ and $v_p^{mR} = v_b^{mR}$ such that there exists a miRNA node that is connected to both nodes v_k^{mR} and v_{k+1}^{mR} in $P(a,b)$ for all $1 \leq k < p$. The distance of the path between this two nodes v_a^{mR} and v_b^{mR} is defined as

$$d(a,b) = \sum_{k=1}^{p-1} \min_{i^*}(w_{i^*k} + w_{i^*k+1}),$$

with i^* denoting the miRNAs that target the mRNAs k and $k+1$. A path $P(a,b)$ between the mRNAs a and b is then called shortest path $P_{min}(a,b)$ if it minimizes the distance $d(a,b)$. The distance of the shortest path is then defined as $d_{min}(a,b)$ and denoted as *shortest distance*. The set of shortest distances from a mRNA a to a set of mRNAs S is defined as $D_{min}(a,S)$. In case of the example in Fig. 6.1a, the shortest distances between all nodes in the graph are listed in Tab. 6.1.

In order to calculate these shortest paths in a given miRNA-mRNA network, we use the implementation of *Dijkstra's algorithm* [31] in the *igraph* package [27] for *R* [137].

Table 6.1: Shortest paths between mRNA nodes in the given example (Fig. 6.1).

	A	B	C	D	E	F	G	H
A	0.00	1.64	1.80	1.80	3.04	2.80	3.20	1.40
B	1.64	0.00	1.64	1.64	2.88	2.64	3.04	1.24
C	1.80	1.64	0.00	1.80	3.04	2.80	3.20	1.40
D	1.80	1.64	1.80	0.00	1.48	1.24	1.64	1.40
E	3.04	2.88	3.04	1.48	0.00	1.24	1.64	1.64
F	2.80	2.64	2.80	1.24	1.24	0.00	1.40	1.40
G	3.20	3.04	3.20	1.64	1.64	1.40	0.00	1.80
H	1.40	1.24	1.40	1.40	1.64	1.40	1.80	0.00

6.1.2 Scoring local neighborhoods

Let $M_k = \{g_{1k}, \ldots, g_{mk}\}$ denote the functional group consisting of m annotated genes for a specific term k with $k \in \{1, \ldots, l\}$ retrieved from a database and let $M_{k,G} := M_k \cap V^{mR}$ be the set of genes associated with M_k that overlap with the genes contained in G. In order to determine the enrichment of $M_{k,G}$ around a certain mRNA j in G, we compare the distribution of the shortest distances $D_{min}(j, M_{k,G})$ to the distribution of shortest distances $D_{min}(j, G \setminus M_{k,G})$. Note that $D_{min}(j, M_{k,G})$ includes the shortest path distance to a node itself, which is defined as zero, if $v_j \in M_{k,G}$. We apply a left-tailed Wilcoxon rank-sum test, which yields a p-value indicating whether the values of $D_{min}(j, M_{k,G})$ are significantly shifted to lower values as compared to the values of $D_{min}(j, G \setminus M_{k,G})$. We then use these p-values to assess a score $S(v_j)$ in form of the negative logarithm to base ten:

$$S(v_j) := -\log_{10}(p_j),$$

which describes the enrichment of the term for the given functional group M_k around gene j.

Eventually, we want to characterize the importance of certain miRNAs in G. For this purpose, we can calculate a score $S_{miR}(v_i)$ for every miRNA node $v_i \in V^{miR}$ by considering the set $\mathcal{V}_i := \{v_j | (v_j \in V_{mR}) \wedge (\exists (v_i, v_j) \in E)\}$ of associated mRNA nodes. The score is then defined as

$$S_{miR}(v_i) = \frac{1}{|\mathcal{V}_i|} \sum_{v_j \in \mathcal{V}_i} S(v_j)$$

The algorithm for the calculation of the gene and miR scores is summarized in

Algorithm 6.1.

Input: Assignment matrix $\boldsymbol{W}$ for graph G (neg. weights), Functional group M_k.
Result: Vector of gene scores, vector of miR scores.
Transform each non-zero element w_{ij} of $\boldsymbol{W}$ to $w_{ij} = e^{w_{ij}}$;
n:=Number of miRNAs in G;
m:=Number of mRNAs in G;
Initialize $ShortestDists[m, m]$;
for $j_1 = 1 : m$ **do**
  for $j_2 = 1 : m$ **do**
    $ShortestDists[j_1, j_2] = d_{min}(j_1, j_2)$;
  end
end
Initialize $scores[m]$;
for $j = 1 : m$ **do**
  $distsToSet = ShortestDists[j, k^*]$ for $k^* \in M_k$;
  $bgDists = ShortestDists[j, k']$ for $k' \notin M_k$;
  $scores[j] = -\log_{10}(p.Wilcoxon.leftTailed(distsToSet, bgDists))$;
end
Initialize $mirScores[n]$;
for $i = 1 : n$ **do**
  $mirScores[i] = sum(scores[j^*])$ for
  $j^* = \{j | (v_j \in V_{mR}) \wedge (\exists (v_i, v_j) \in E\}$;
end
return *scores,miRscores*

Algorithm 6.1: Calculation of gene and miR scores using LEA for a given functional group M_k.

Consider for example the given miRNA-mRNA network G in Fig. 6.3a. Let the overlap between the functional group M_k of a term k with the genes in G be $M_{k,G} = \{\mathrm{A}, \mathrm{I}, \mathrm{J}, \mathrm{P}\}$ (diamond shape). We now want to infer the genes in the proximity of the genes in $M_{k,G}$. For this purpose, we calculate the set of shortest distances for node M (blue) the genes in $M_{k,G}$, which yields $D_{min}(\mathrm{M}, M_{k,G}) = \{1.01, 1.21, 1.00, 1.05\}$. In case of node B (purple), the set of shortest distances is $D_{min}(\mathrm{B}, M_{k,G}) = \{2.07, 2.16, 2.30, 3.11\}$. We can now compare these distributions of shortest distances to the respective distributions of shortest distances to genes, which are not in $M_{k,G}$, namely $D_{min}(\mathrm{M}, V^{mR} \setminus M_{k,G})$ and $D_{min}(\mathrm{B}, V^{mR} \setminus M_{k,G})$ (Fig. 6.3c). We observe that in case of node M, the distribution of shortest distances $D_{min}(\mathrm{M}, M_{k,G})$ tends to be shifted to lower values as compared to $D_{min}(\mathrm{M}, V^{mR} \setminus M_{k,G})$, which is not the case for node B. In order to statistically test for this shift to lower values, we apply a one-

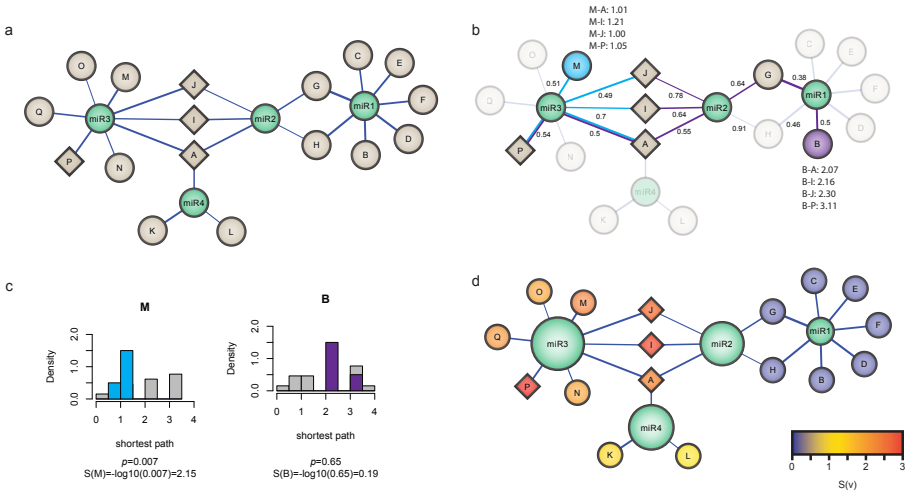


Figure 6.3: Node scoring in an example network. (**a**) We assume that four genes are assigned to a certain functional group (A, I, J, P; diamond shape). The edge weight indicates the strength of negative regulation. (**b**) The shortest path distances between these four nodes and the two nodes M (blue) and B (purple) are exemplary shown. The edge labels denote the weights after the transformation. (**c**) The distribution of shortest paths from node M to the nodes A, I, J and P is significantly shifted to lower values. No shift can be observed for node B. The p-values determined by Wilcoxon rank-sum test are converted to the node scores. (**d**) The scores of all nodes are indicated by the color. The size of the miRNA nodes corresponds to the miR score, which is the mean score of targeted nodes.

sided Wilcoxon rank-sum test and obtain a significant p-value in case of node M ($p = 7 \times 10^{-3}$), while the p-value for node B is not significant ($p = 6.5 \times 10^{-1}$). We can therefore conclude that node M is actually located in close proximity of the nodes associated with the term k, whereas node B is not.

Given the p-values from the Wilcoxon rank-sum test, we can calculate the scores of the two nodes M and B as $S(M) = -\log_{10}(7{\times}10^{-3}) = 2.15$ and $S(B) = -\log_{10}(6.5 \times 10^{-1}) = 0.19$, respectively. This score indicates the proximity of a gene to the genes in the functional group (Fig. 6.3d).

6.1.3 Identification of locally enriched functional groups

We are not only interested in finding particular overrepresented regions in our network for a preselected group of interest, but first of all in identifying these groups of interest. Hence, we aim to select functional groups, whose associated

genes are not equally distributed over the network but have the tendency to be located in close proximity. In other words, when genes are in a non-random close proximity in the miR-target network, we consider a functional group as enriched. For this purpose, we aim to identify terms whose functional groups are located in close proximity to each other and which are at the same time strongly regulated by the miRNAs in these areas.

Let again G be the network, which represents the regulatory relationships between miRNAs and mRNAs and $M_{k,G}$ the set of overlapping genes between the functional group M_k of term k and the genes in G. We now select all shortest distances $D_{min}(M_{k,G})$ between the nodes $v_j \in M_{k,G}$. By comparing the distribution of this set to the distribution of all other shortest distances $D_{min}(V_{mR} \setminus M_{k,G})$, we can determine whether the nodes of the functional group $M_{k,G}$ are located in a closer proximity compared to the other nodes in G. We are again able to quantify this shift by applying a one-sided Wilcoxon rank sum test, which yields a p-value for each given functional group. We can thus assess the enrichment of associated terms in a local area of the network. As we perform a repeated testing procedure, we have to correct the resulting p-values by a correction procedure such as the FDR or Bonferroni method (see Chapter 2.2.5). The algorithm for obtaining locally enriched functional groups is summarized in Algorithm 6.2.

Consider for example the network above (Fig. 6.3a) with the given arrangement of nodes $M_{k,G} = \{\mathrm{A}, \mathrm{I}, \mathrm{J}, \mathrm{P}\}$ assigned to a term k. If we compare the distribution of shortest path distances between these nodes (Fig. 6.4a), we can observe a shift to lower values. This test results in a p-value of $p = 1.3 \times 10^{-2}$.

Now consider another term $\tilde{k}$ with $M_{\tilde{k},G} = \{M, K, E\}$. Obviously, these nodes are equally distributed over the whole network (Fig. 6.3a). Therefore, clearly no shift to lower values can be observed as compared to the background distribution (Fig. 6.4b). Hence, we obtain a highly non-significant p-value by applying an one-sided Wilcoxon rank sum test ($p = 0.9$).

Input: Assignment matrix $\boldsymbol{W}$ for graph G (neg. weights), Set of functional groups M.
Result: Vector of p-values for local enrichment.
Transform each non-zero element w_{ij} of $\boldsymbol{W}$ to $w_{ij} = e^{w_{ij}}$;
m:=Number of mRNAs in G;
Initialize $ShortestDists[m, m]$;
for $j_1 = 1 : m$ **do**
  for $j_2 = 1 : m$ **do**
    $ShortestDists[j_1, j_2] = d_{min}(j_1, j_2)$;
  end
end
Initialize $pvalues[|M|]$;
for $k = 1 : |M|$ **do**
  $setDists = ShortestDists[k^*, k^*]$ for $k^* \in M_k$;
  $bgDists = ShortestDists[k', k']$ for $k'^* \notin M_k$;
  $pvalues[k] = p.Wilcoxon.leftTailed(setDists, bgDists)$;
end
$pvalues = p.adjust(pvalues)$;
return *pvalues*

Algorithm 6.2: Calculation of p-values to determine locally enriched functional groups using LEA.

6.2 Biological properties of microRNA-target network in adipogenesis

In the previous chapter we described the application of miRlastic for the construction of a miRNA-mRNA regulatory network that describes the relationships between miRNAs, which are altered by the adipocyte differentiation process, and potential target genes. We now aim to identify functional groups, which are specifically enriched in a certain area in our network. This allows us to infer functional properties of involved miRNAs. Therefore, we used the previously generated network (see Chapter 5.4) as input for our LEA approach. We used human pathway annotations from KEGG [84] as functional groups. Among this set of KEGG pathways, we selected only those that did not correspond to a disease or drug development group. In total, we obtained 135 functional groups, which we used for LEA. We applied LEA as described above, which resulted in a p-value for each group indicating the local enrichment of the respective pathway. We then applied the p-value correction by Benjamini and Hochberg [12] in order to correct for multiple testing and selected all pathways

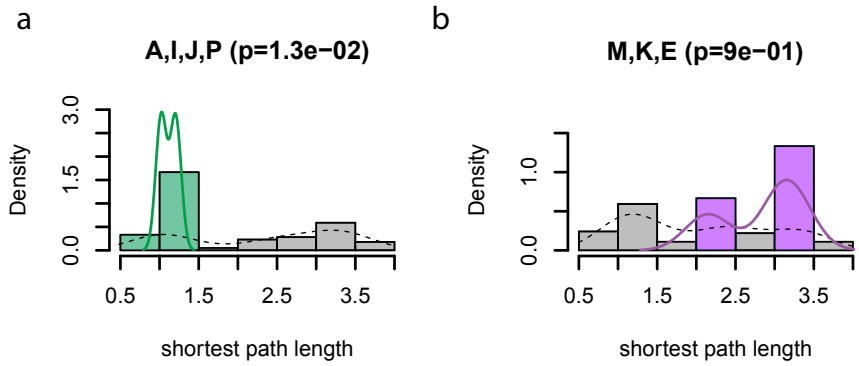


Figure 6.4: Shortest path distributions for nodes with high proximity (nodes A, I, J, P) and low proximity (nodes M, K, E). The selection of the nodes refers to the genes that are assigned to a certain functional group. The shortest path distribution of the high proximity nodes is significantly shifted to lower values compared to the background whereas the low proximity nodes are not shifted.

whose adjusted p-value was smaller than 0.05. We then finally obtained seven significantly locally enriched pathways (Tab. 6.2).

Table 6.2: Locally enriched pathways from KEGG in the adipogenesis network with adjusted p-value less than 0.05.

KEGG ID	p-value	adj. p-value	pathway name
hsa04010	1.00E-18	1.36E-16	MAPK signaling pathway
hsa04120	2.71E-10	3.63E-08	Ubiquitin mediated proteolysis
hsa04020	3.81E-09	5.07E-07	Calcium signaling pathway
hsa04380	1.49E-07	1.97E-05	Osteoclast differentiation
hsa04360	2.08E-07	2.73E-05	Axon guidance
hsa04512	5.01E-07	6.52E-05	ECM-receptor interaction
hsa04350	4.76E-05	6.14E-03	TGF-beta signaling pathway

For these seven pathways, we then calculated the scores $S(v^{mR})$ for the gene nodes v^{mR} in our network. We observed an unequal distribution over the network indicating the local enrichment for the respective pathway (Fig. 6.5).

We then used the scores of the gene nodes for calculating the miR scores $S_{miR}(v^{miR})$ for the miRNA nodes v^{miR}. Hence, we obtained a measure for the relevance of each miRNA for the respective pathway (Fig. 6.6).

By clustering the pathways and miRNAs with regard to the miR scores, we can primarily identify two sets of miRNAs with similar behavior and relevance to distinct sets of pathways. We obtain similar miR scores for the miRNAs of the miR-30 family across all pathways, which is what we would be expect by

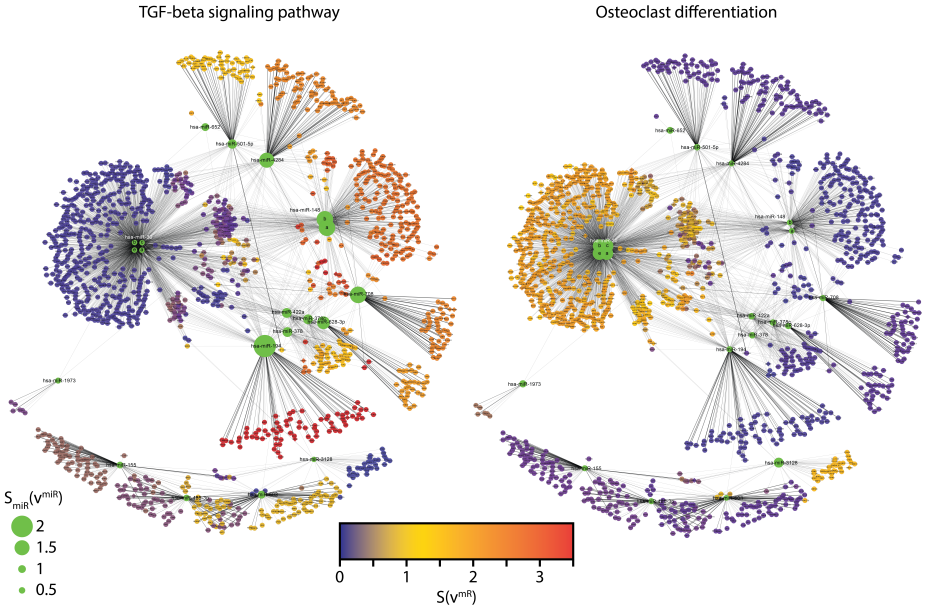


Figure 6.5: Local enrichment analysis for the TGF-beta signaling pathway and osteoclast diferentiation. The color of the mRNA nodes v^{mR} corresponds to their score $S(v^{mR})$. The size of the miRNA nodes v^{miR} corresponds to their miR score $S_{miR}(v^{miR})$. For osteoclast diferentiation, the enrichment can be primarily observed for the miR-30 family whereas the TGF-beta signaling pathway is mainly enriched for the miRNAs hsa-miR-194, hsa-miR-148a/b, hsa-miR-4284, hsa-miR-708 and hsa-miR-628-3p.

taking into account that all four miRNAs have the identical predicted target set and exhibit similar expression profiles. The same holds for the two miRNAs miR-148a/b. However, the miRNAs miR-194/4284/6283p and 708 also seem to have similar relevance across the pathways even though they are apparently not directly related to each other in terms of family or cluster membership. Hence, we are able to characterize coordinated regulation of biological processes for related miRNAs as well as for miRNAs, whose functional similarities are beyond the cluster or family membership.

The miRNAs of the miR-30 family show functional relevance for MAPK signaling, Ubiquitin mediated proteolysis, Osteoclast differentiation, Calcium signaling and Osteoclast differentiation. MAPK signaling [17] as well as calcium signaling [85] play important roles in adipogenesis. Also the suppressive effect of ubiquitin mediated proteolysis of PPARγ on adipocyte differentiation has already been pointed out [88]. Recent studies have already shown that the

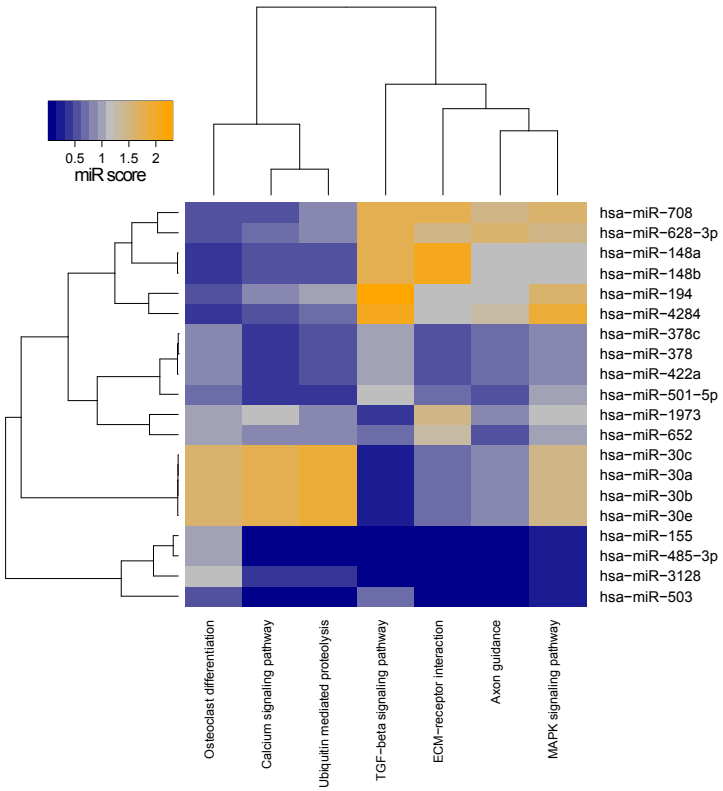


Figure 6.6: Heatmap of miR scores for each miRNA in the network indicating the functional role in the significantly locally enriched KEGG pathways.

miRNAs of the miR-30 family suppress the key players in adipogenesis, whose up-regulation is essential for the differentiation process [186], thereby negatively regulating osteoblast differentiation [179]. KEGG does not include a functional group for osteoblast differentiation but for osteoclast differentiation, which is indeed under strong influence of the miR-30 family. The regulation of osteoclast differentiation is in good agreement with previous findings as osteoclast differentiation directly depends on the regulation by osteoblasts [85] and as a functional role of miR-30 in osteoclast differentiation has also already been proposed [179]. Hence, we were able to reveal processes, which are relevant for adipogenesis and which are regulated by miR-30 family.

The two miRNAs from the miR-148 family in contrast exhibit a functional pattern, which is very distinct from the miR-30 family. The highest miR score for these two miRNAs is for extracellular matrix (ECM) receptor interaction

but they are also associated with TGF-β signaling. As discussed above, the TGF-β signaling pathway is a crucial process in adipogenesis [24]. Our results suggest that not only the miR-148 family act as regulator of this pathway, but also the miRNAs miR-194/4284/6283p and 708.

6.3 Discussion and Conclusion

In this chapter we introduced a local enrichment approach (LEA) for miRNA-mRNA networks, which allows for the functional characterization of miRNAs. LEA uses the network structure, which represents the regulatory relationships between miRNAs and target genes. It is based on the calculation of shortest paths between genes in the network, which are assigned to a certain functional group, thereby taking into account the strength of miRNA regulation. As a result, functional groups were identified, whose associated genes are locally enriched in the network. In addition, LEA identifies the enriched areas in the network and determines the miRNAs, which are associated with these areas.

LEA is especially designed for the output of miRlastic, but may also be adapted to networks generated by other approaches. The only requirement is that weighted bipartite graph is provided, where one node set can be mapped to functional properties. Another application may be the local enrichment analysis in a network consisting of genes and metabolites.

We showed that LEA generates good results when applied to the previously generated miRNA-mRNA network for the adipogenesis data. We could determine specific functional roles of miRNAs in processes, which are important for the adipocyte differentiation. We therefore gain functional insight into the regulatory mechanisms of miRNAs, which are affected during adipogenesis.

Chapter 7

MONA: Multilevel ontology analysis

In this chapter, we will introduce a novel approach called multilevel ontology analysis (MONA), which extends the basic idea of data integration from various sources towards a modular framework for functional analyses on multiple molecular levels. The goal of this approach is to provide a flexible method for gene set analysis, which is able to determine gene responses across multiple omics datasets.

The MONA model includes the associations between genes and functional categories, which commonly represent certain cellular functions and may be retrieved from databases like KEGG [84] or GO [5]. These categories, however, usually exhibit a high degree of redundancy. This holds especially for GO due to its hierarchical term structure. Our model-based approach accounts for this redundancy by inferring enriched functional categories simultaneously.

We will introduce the *cooperative* and the *inhibitory model*, which both represent special molecular relationships between the observed molecular levels. In general, MONA can be easily extended with regard to different kinds of molecular interactions between these levels in order to determine the hidden gene responses.

We will show the performance of MONA on both, synthetic and previously published real data. Finally, we will combine the cooperative and the inhibitory

model in order to account for the mRNA expression, DNA methylation and miRNA expression data from the adipogenesis dataset. Using this approach, we are able to reveal cellular processes that are affected on several molecular levels, which in turn allows us to gain deeper insight into the regulatory mechanisms that play a role during adipogenesis.

This chapter was published in parts in:

- **Steffen Sass***, Florian Buettner*, Nikola S. Mueller, and Fabian J. Theis. A modular framework for gene set analysis integrating multilevel omics data. *Nucleic Acids Res*, 41(21):9622–9633, Nov 2013.

$*$ = equal contributions

7.1 Model-based enrichment analysis

The ability of cells to adjust to given environmental or disease conditions is a result of their ability to perform specific biological functions and processes. These are in turn orchestrated by a tight regulation of gene responses across many molecular levels (Fig. 7.1). The gene product carrying out the biological function is a result of not only protein expression and activity, but also of gene expression on mRNA level, gene promoter methylation states and existing single nucleotide polymorphisms within the genome. Fine-tuning mechanisms of e.g. microRNA (miRNA) post-transcriptional modification of mRNAs also contribute to the joint gene responses of cells. Finally, protein phosphorylation controls the enzymatic activity of a gene product for example in signaling cascades [75].

Here, we propose a model-based method to reliably calculate interpretable probabilities for GO terms activity by integrating multi-level gene response data. We perform a multi-level ontology analysis (MONA) using a Bayesian approach with a computationally efficient method to approximate the marginal posteriors of ontology terms given lists of genes responding to experimental conditions. MONA is designed to easily handle any combination of molecular levels in a modular fashion. This is illustrated by a cooperative and an inhibitory model. We demonstrate that MONA outperforms existing methods by integrating multi-omics levels with appropriate biological models not only on synthetic

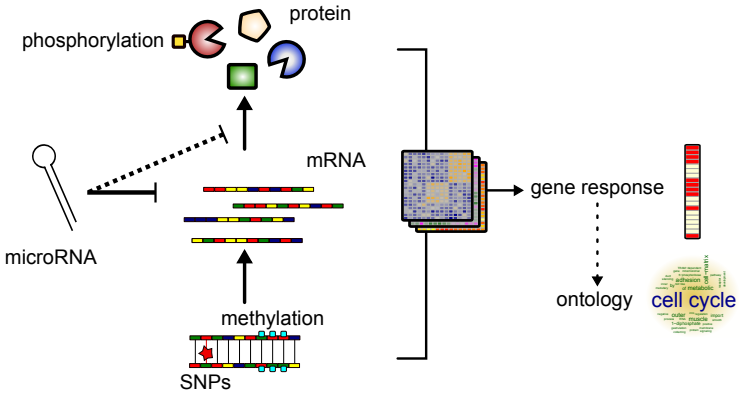


Figure 7.1: Multilevel gene responses. The signature of condition-specific changes in biological functions is captured in gene responses, which are measurable on many omics levels. When integrated across levels, organism-wide profiling provides a comprehensive and multilevel picture that most reliably describes active biological processes.

data but also on three integrative studies covering mRNA, protein, methylation states as well as post-transcriptional modifications by miRNA. The framework and inference method is flexible enough to easily allow for other data, underlying regulatory motifs or ontologies. For example, an extension to a cooperative three-level model is straight-forward.

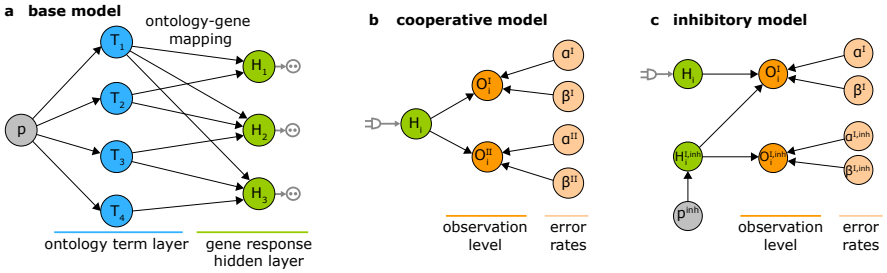


Figure 7.2: A modular approach for gene set enrichment analysis with multiple observed levels. **a.** In the base model terms T are connected to hidden gene products H. Each hidden gene product is observed in form of noisy measurements of one or several levels. **b,c.** Two examples for modules coupled to one hidden gene product depending on the biological relationship of the molecular levels analyzed. Each molecular level in the observation layer O has separate error rates. Noise of the measurements is represented by false positive and false negative rates α and β. Note that only the hidden gene products H_i are attached directly to an ontology term. The hidden inhibitor activity $H_i^{I,inh}$ is specific for a respective gene.

The base model can be represented by a Bayesian network with two layers

(Fig. 7.2a) as described previously [110, 9]: the (ontology) term layer consists of boolean nodes indicating whether a term is active or not. Each term (T) is connected to a set of hidden gene products (H) as defined by e.g. Gene Ontology (GO). This hidden (unobserved) layer of gene responses has to be introduced between the ontology and the layer of observed variables, for two reasons: First, measurement errors result in false positives (FP) and false negatives (FN) that have to be handled adequately. Second, incorrect or imprecise term-gene assignments may occur e.g. due to links inferred automatically by GO. Altogether, the hidden gene response layer also allows for a coherent integration of biological observations across multiple layers.

More formally, we define our base model (Fig. 7.2a) in form of conditional probability densities. These conditional densities are described in the following sections.

7.1.1 Terms

T_i are Bernoulli-distributed boolean random variables modelled with a prior probability p of being on. As we do not know p in advance, we place a Beta prior over p so that we can learn it from the data:

$$p \sim \text{Beta}(a, b) \tag{7.1}$$

with a and b being the shape parameters of the Beta-Distribution with probability density function

$$f(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, x^{a-1}(1-x)^{b-1},$$

where $\Gamma(z) = (z-1)!$. When a and b are set to 1, we have a uniform prior (i.e. before having seen the data we consider all possible values for p as equally likely). Prior knowledge on the distribution of p (e.g. if p is known to be small) can be included in form of different choices of a and b (e.g. $a = 1$ and $b = 5$ places most of the probability mass on values less than 0.5).

It is worth noting that the previously defined base model [9] slightly differs form our model: while we place a continuous prior on the probability for a term being on, they chose a restrictive, discrete prior which is defined by default as

$p \in \{1/N, \ldots, 20/N\}$ with N being the number of terms.

7.1.2 Hidden nodes

The nodes H_i represent the underlying hidden response of each individual gene. They are modelled as boolean variables, which are deterministically defined such that $H_i = 1$ if at least one term to which H_i is annotated is on; otherwise $H_i = 0$. If we define $T(H_i)$ to denote the set of terms to which gene H_i is annotated, then we can write:

$$P(H_i|T) = \begin{cases} 1 & \text{if } \exists\ T_j \in T(H_i) : T_j = 1 \\ 0 & \text{otherwise} \end{cases} \tag{7.2}$$

7.1.3 Modular framework to integrate multilevel observations

Depending on the number of observed levels (e.g. mRNA, protein and methylation) and their relation to each other, the observed nodes O_i are connected to hidden gene responses H_i. With MONA we present a general framework allowing for an easy integration of arbitrary molecular levels. We illustrate our novel approach by describing three different models in detail.

Single-level Model

In this scenario, measurements are only available for one level (e.g. mRNA expression). Consequently, each observation is connected to exactly one hidden node representing its respective gene product (this can be interpreted as a special case of figure 7.2b with only one observed level O_i^I).

Observations O_i^I are observed with false positive and false negative rates α^I and β^I; similar to p we place (usually uniform) Beta priors on α^I and β^I as we usually do not know these rates in advance and want to infer them from the

data.

$$P(O_i^I = 1|H_i) = \begin{cases} 1-\alpha^I & \text{if } H_i = 1 \quad \text{(true positive: TP)} \\ \alpha^I & \text{if } H_i = 0 \quad \text{(false positive: FP)} \end{cases} \tag{7.3}$$

$$P(O_i^I = 0|H_i) = \begin{cases} 1-\beta^I & \text{if } H_i = 0 \quad \text{(true negative: TN)} \\ \beta^I & \text{if } H_i = 1 \quad \text{(false negative: FN)} \end{cases} \tag{7.4}$$

Cooperative Model

The cooperative model accounts for studies where measurements of two (or more) different levels are available, which may be regarded as independent noisy observations (e.g. mRNA and protein) of an underlying common gene response. In contrast to the single-level model, an additional level is observed, which is modelled as independent observation O_i^{II} of gene product with separate false positive and false negative rates α^{II} and β^{II} (Fig. 7.2b). Again we place Beta priors on α^{II} and β^{II}. For each additional levels, error rates are defined accordingly.

$$P(O_i^{II} = 1|H_i) = \begin{cases} 1-\alpha^{II} & \text{if } H_i = 1 \\ \alpha^{II} & \text{if } H_i = 0 \end{cases} \tag{7.5}$$

$$P(O_i^{II} = 0|H_i) = \begin{cases} 1-\beta^{II} & \text{if } H_i = 0 \\ \beta^{II} & \text{if } H_i = 1 \end{cases} \tag{7.6}$$

Inhibitory Model

The inhibitory model is applicable when two levels are measured, but they could not be interpreted as independent measurements of the hidden gene function (Fig. 7.2c). A prominent example is the post-transcriptional modulation of mRNA expression by miRNAs. We introduce an additional hidden variable $H_i^{I,inh}$ to the model for each respective gene response H. $H_i^{I,inh}$ is a boolean random variable which represents the true underlying state of the inhibitor: If the inhibitor is active, $H_i^{I,inh} = 1$, otherwise $H_i^{I,inh} = 0$. $H_i^{I,inh}$ is modelled to be active with prior probability p^{inh} ($P(H_i^{I,inh} = 1) = p^{inh}$). $H_i^{I,inh}$ is observed in form of $O_i^{I,inh}$ with own false positive and false negative rates $\alpha^{I,inh}$ and

$\beta^{I,inh}$:

$$P(O_i^{I,inh} = 1|H_i^{I,inh}) = \begin{cases} 1-\alpha^{I,inh} & \text{if } H_i^{I,inh} = 1 \\ \alpha^{I,inh} & \text{if } H_i^{I,inh} = 0 \end{cases} \tag{7.7}$$

$$P(O_i^{I,inh} = 0|H_i^{I,inh}) = \begin{cases} 1-\beta^{I,inh} & \text{if } H_i^{I,inh} = 0 \\ \beta^{I,inh} & \text{if } H_i^{I,inh} = 1 \end{cases} \tag{7.8}$$

The second observable in the model is the inhibited level (O_i^I). As opposed to the cooperative model, the conditional probability density does not only depend on H_i, but also on $H_i^{I,inh}$:

$$P(O_i^I = 1|H_i^{I,inh}, H_i) = \begin{cases} 1-\alpha^I & \text{if } (H_i^{I,inh} = 0 \wedge H_i = 1) \\ & \vee (H_i^{I,inh} = 1 \wedge H_i = 0) \quad \text{(TP)} \\ \alpha^I & \text{if } (H_i^{I,inh} = 1 \wedge H_i = 1) \\ & \vee (H_i^{I,inh} = 0 \wedge H_i = 0) \quad \text{(FP)} \end{cases} \tag{7.9}$$

$$P(O_i^I = 0|H_i^{I,inh}, H_i) = \begin{cases} 1-\beta^I & \text{if } (H_i^{I,inh} = 1 \wedge H_i = 1) \\ & \vee (H_i^{I,inh} = 0 \wedge H_i = 0) \quad \text{(TN)} \\ \beta^I & \text{if } (H_i^{I,inh} = 0 \wedge H_i = 1) \\ & \vee (H_i^{I,inh} = 1 \wedge H_i = 0) \quad \text{(FN)} \end{cases} \tag{7.10}$$

This reflects the interaction between the two levels: true gene response can either be explained by uninhibited first level or if the inhibitor is active without the first level being active.

For inference a variety of techniques exist. Lu et al. [110] proposed a maximum-likelihood approach (analyzing only a single level), where the likelihood $L(T_{\text{active}}|\boldsymbol{D}, \boldsymbol{\theta})$ is maximized with respect to the set of active GO terms T_{active}, given the observed data $\boldsymbol{D}$ and a set of parameters $\boldsymbol{\theta}$. A drawback of the maximum likelihood method is that no distribution is inferred and only one local maximum is found, ignoring alternative solutions. A more robust approach then used Markov Chain Monte Carlo (MCMC) methods to estimate the marginal posterior probabilities $P(T|\boldsymbol{D})$ of being active [9]. The marginal posterior is calculated by using a Metropolis-Hastings algorithm to sample from the joint posterior distribution $P(T, \boldsymbol{\theta}|\boldsymbol{D})$. Such MCMC approaches asymptotically provide a random sampler of a target distribution when being run long enough. Consequently, they are a family of algorithms commonly used for in-

ferring posterior distributions of Bayesian networks, which cannot be analyzed analytically. However, major drawbacks are comparatively long run times and for every model definition (e.g. if a another level is measured) a new custom sampler has to be implemented which can be very time-consuming and requires expert knowledge.

In order to overcome the drawbacks of existing methods, we use computationally efficient approximate methods [15] to approximate the marginal posterior.

The marginal posteriors of interest were approximated using the EP algorithm [120], which is described above. These marginal posterior probabilities $P(T|\boldsymbol{D})$ (in the following simply referred to as term probability) can be interpreted as the outcome of the MONA algorithm in form of the probabilities for each term to be active as best explained by the data.

The posterior of the model factorizes as: $p(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{1}{p(\boldsymbol{D})} \prod_i f_i(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are all parameters of the model and f_i functions as defined in the model specifications while depending on the specific generative model definition. For example, for the cooperative model $\theta = \{p, T, H, \alpha^{I/II}, \beta^{I/II}\}$ such that

$$p(T, H, p, \alpha, \beta|\boldsymbol{D}) = \frac{p(T|p)p(\boldsymbol{D}|H, \alpha, \beta)p(H|T)p(\alpha)p(\beta)p(p)}{p(\boldsymbol{D})} \tag{7.11}$$

with the individual factors as defined in equations 7.1 to 7.6.

7.2 Implementation

We use probabilistic programming to perform the inference within the Infer.NET framework [119]. Infer.NET is a framework allowing for Bayesian inference in graphical models, which has been used successfully in the bioinformatics community in recent years [129, 160]. The approximation of the marginal posterior is performed by the infer.NET inference engine. The main advantage is that it is straight-forward to specify different models of gene responses given a common base model. Thus, changing model specification and adding additional level only requires few lines of code resulting in a fast and flexible framework for Bayesian GO analysis. We provide an implementation of MONA for the

cooperative and the inhibitory model together with a graphical user interface in form of a .NET application. The user has to provide a list of zeros and ones corresponding to the set of altered genes out of a total set of measured genes for each level. According to this list, an assignment matrix has to be provided that maps the gene at each position to a set of terms. This tool can also be used via the command line, which allows for a flexible integration into multi-omics analysis pipelines. Using the Mono ASP.NET framework, it can be used on any operating system.

7.3 Evaluation

7.3.1 Synthetic data

Realistic synthetic data generated for the single-level and the cooperative model were sampled from genome-wide yeast genes mapped to GO [5] (retrieved Oct. 2012). We used the Bioconductor package org.Sc.sgd.db which annotated 3890 terms to 6396 genes. Realistic data for the inhibitory model was generated by sampling from hgu133plus2.db,for Affymetrix human genome annotations where 14740 genes are annotated with 10944 terms. We randomly selected 3 to 6 independent terms to be active in each data-set. We sampled the corresponding observed level according to the single level, cooperative and the inhibitory model respectively. This was done for a range of different parameter values of $\alpha^{I/II}$, $\beta^{I/II}$ and p^{inh}. For the single/cooperative and the inhibitory model, we generated 600 and 400 synthetic datasets with different levels of observation noise respectively. More specifically, for the single-level model and the cooperative model we chose 3 different settings: $\alpha^{I/II} = 0.25$ and $\beta^{I/II} = 0.25$; $\alpha^{I/II} = 0.25$ and $\beta^{I/II} = 0.4$; $\alpha^{I/II} = 0.1$ and $\beta^{I/II} = 0.4$. The inhibitory model was evaluated for four different levels of observation noise and miRNA activation: $\alpha^{I/I,inh} = 0.25$, $\beta^{I/I,inh} = 0.25$ and $p^{inh} = 0.25$; $\alpha^{I/I,inh} = 0.25$, $\beta^{I/I,inh} = 0.25$ and $p^{inh} = 0.4$; $\alpha^{I/I,inh} = 0.1$, $\beta^{I/I,inh} = 0.4$ and $p^{inh} = 0.25$; $\alpha^{I/I,inh} = 0.1$, $\beta^{I/I,inh} = 0.4$ and $p^{inh} = 0.1$.

We compared results of MONA to related approaches for GO enrichment analysis, all suited for analysing single-level data. We quantified the statistical significance of differences in predictive power between the following ap-

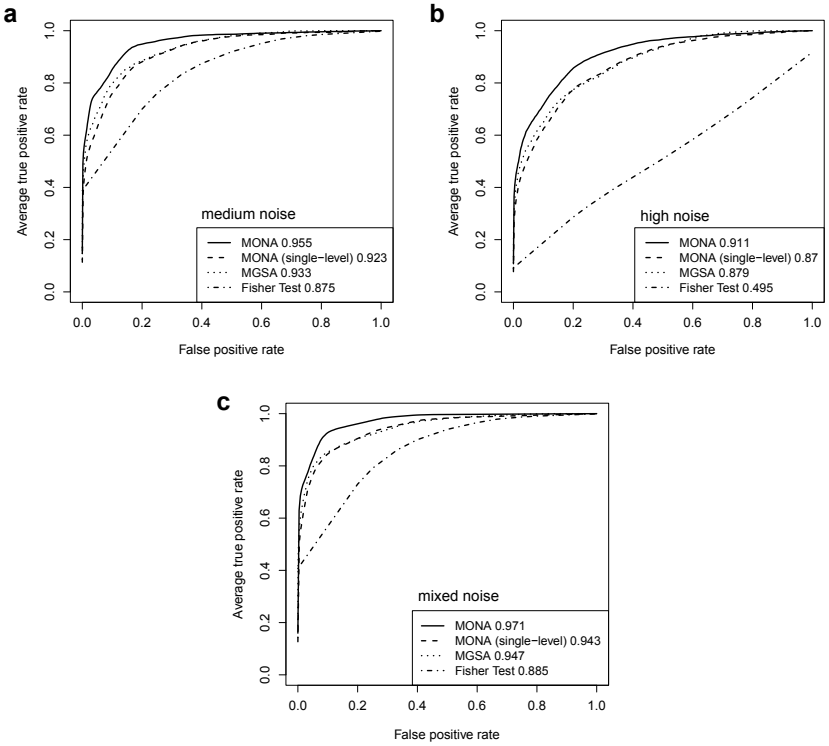


Figure 7.3: Performance of the cooperative model on synthetic data for 3 different levels of noise: (**a**) medium noise ($\alpha^{I/II} = 0.25$, $\beta^{I/II} = 0.25$), (**b**) high noise ($\alpha^{I/II} = 0.25$, $\beta^{I/II} = 0.4$) and (**c**) mixed noise ($\alpha^{I/II} = 0.1$, $\beta^{I/II} = 0.4$). AUC values are listed in the respecitve figure legends. With MONA the inference is based on 2 levels, all other algorithms are based on one level only.

proaches: inferring active GO terms based on i) one level only with MGSA, ii) one level-model of MONA and iii) multi-level integrative method MONA. Therefore, we performed an receiver-operating-characteristic (ROC) analysis of each synthetic dataset and quantified the statistical significance between two different approaches by performing a paired t-test (Bonferroni corrected) between the respective area-under-the-curve (AUC) values.

Although, most similar to MONA, MGSA [9] can only be applied to individual molecular levels. As MGSA is an MCMC sampling scheme for inferring marginal posteriors for the single-level model and converges to the exact solution when run long enough, we used the solutions provided by the MCMC sampling as gold standard for the single-level model. To illustrate benefits over the commonly used Fisher's exact test for GO enrichment, where each term is tested independently, we also tested the null-hypothesis of a term being off for

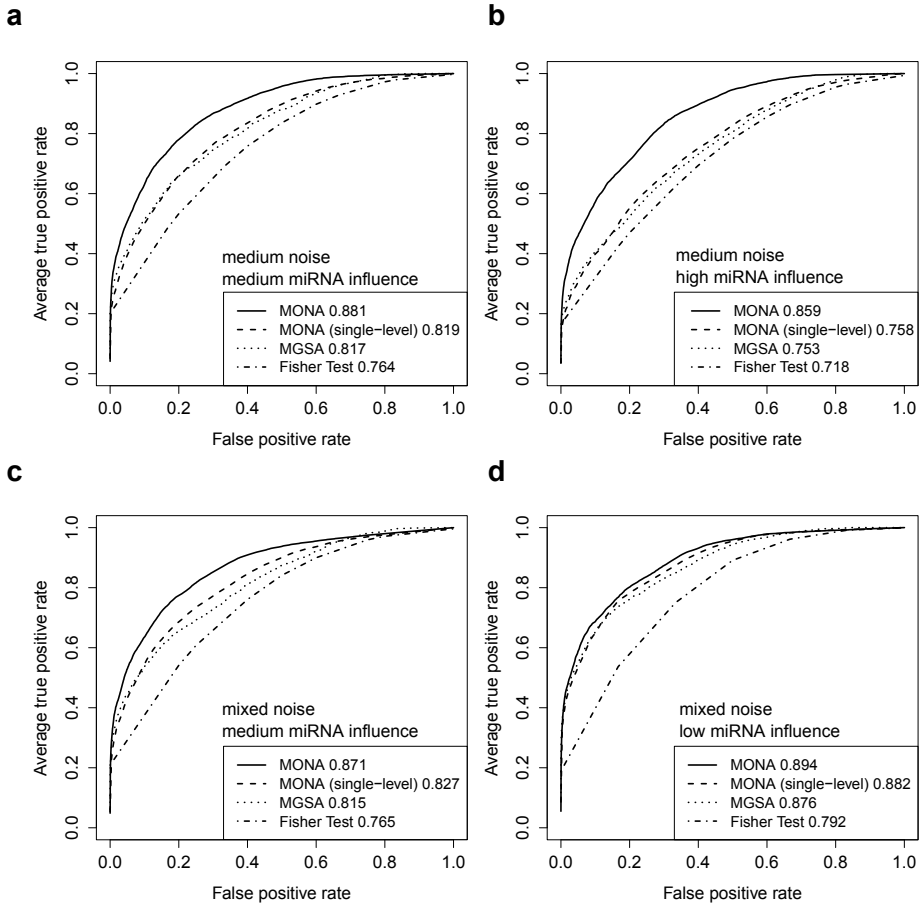


Figure 7.4: Performance of the inhibitory model on synthetic data for 3 different levels of miRNA activation and 2 different noise levels: (**a**) medium noise levels, medium miRNA influence ($\alpha^{I/I,inh} = 0.25$, $\beta^{I/I,inh} = 0.25$, $p^{inh} = 0.25$), (**b**) medium noise levels, high miRNA influence ($\alpha^{I/I,inh} = 0.25$, $\beta^{I/I,inh} = 0.25$, $p^{inh} = 0.4$), (**c**) mixed noise levels, medium miRNA influence ($\alpha^{I/I,inh} = 0.1$, $\beta^{I/I,inh} = 0.4$, $p^{inh} = 0.25$) and (**d**) mixed noise levels, low miRNA influence ($\alpha^{I/I,inh} = 0.1$, $\beta^{I/I,inh} = 0.4$, $p^{inh} = 0.1$). AUC values are listed in the respective figure legends. With MONA the inference is based on 2 levels, all other algorithms are based on one level only.

all terms and calculated ROC curves based on the p-values for all datasets.

For the single-level model as well as the cooperative model, we used uninformative priors for α, β and p in order to introduce as little bias as possible. However, when the marginals yielded an unrealistic value for p (i.e. more than 30% of terms being on) we repeated the inference with a weakly informative prior for p and set the shape parameters of the Beta distribution a and b to 1 and 5 respectively, placing most of the probability mass on values less than 0.5 (this was necessary in about 5% of the synthetic datasets). As we found that parameters p in the inhibitory model converged to unrealistic values more often, we always performed inference with weakly informative priors in this case.

We found that approximate inference with MONA in a single-level model yielded equally good results as the MCMC-based inference with MGSA (Fig. 7.3) for 3 different noise levels. AUC values for MGSA and the single-level model of MONA were 0.932, 0.878, 0.946 and 0.922, 0.87, 0.943 respectively. We used paired t-tests to test the null-hypothesis that both inference methods result in equal performance for a given observation error rate. Resulting p-values of 0.007, 0.14 and 1 indicate that only for error rate $\alpha = 0.25$ and $\beta = 0.25$ the difference in AUC was significant. However, in this case the mean difference in AUC of only 0.01 was rather small. This corresponds to an overall good quality of the EP approximation used by MONA compared to the exact inference method of the MGSA implementation.

AUC curves generated by MGSA do seem to differ systematically from the ROC curves generated using single-level MONA (Fig. 7.3): for all error rates, MGSA achieved higher true positive rates for low false positive rates. This is a consequence of systematic differences between the MCMC sampling approach and EP. For MGSA, the probability of a term being "on" is restricted to 20 discrete values between 0.0002 and 0.0051 so that all models with a higher value for p have a probability of 0. In contrast, for the EP algorithm a continuous Beta prior $(0, 1)$ is used.

Furthermore, the EP approximation is designed such that it prefers broad approximations and due to this zero-avoidance can assign non-zero probabilities to models which actually have a zero probability (this is the opposite behaviour of the MCMC sampling approach which assigns zero probability to all models

with $p > 0.0051$, some of which actually may have a non-zero probability). Consequently, MGSA should be used instead of using the approximate EP inference for a single level if only one level of observations is available.

When comparing the benefits of using integrated data information over individual data levels, the cooperative model yielded AUCs, which were significantly better than the performance of MGSA (p-values $< 10^{-12}$ in all settings). Similarly, in the inhibitory setting, MONA performed significantly better than MGSA (p-values $< 10^{-6}$) for low (10%), medium (25%) and high (40%) influence of miRNA activation (Fig. 7.4). As expected the benefit of including knowledge on the second level was greatest for the setting with high miRNA influence. In this setting also the benefit of the model-based single-level approach over the Fisher test, was smallest.

Run time

For evaluating run times, we applied MONA (here, the cooperative model), MONA on single level and MGSA on the synthetic data described above and repeated this procedure 10 times. MGSA took 192.59 seconds on average ($SD = 45.09s$) to compute the results, while MONA and single-level MONA took 8.45 and 6.96 seconds on average, respectively ($SD = 0.44s$; $SD = 0.36s$). MONA has a considerable gain of run time performance. Note, that MONA had only a slight increase in run time when a second level was introduced in the model.

7.3.2 Real data

The induction of environmental stress to an organism leads to changes on all molecular levels in order to cope with the new condition. An integrative study in yeast investigated changes in the proteome and transcriptome in response to an osmotic shock by NaCl [97]. The regulatory response was measured at different time points after NaCl treatment. We adopted the testing procedures for differential expression from the original study to calculate p-values of mRNAs and proteins [97]. We then considered mRNA and protein as responsive to osmotic stress if their calculated p-value was less than 0.05. In addition we applied a threshold of the absolute median fold change over time of > 0.5 and > 0.3 for mRNA and protein, respectively. Out of 5,916 genes and 2,207 proteins

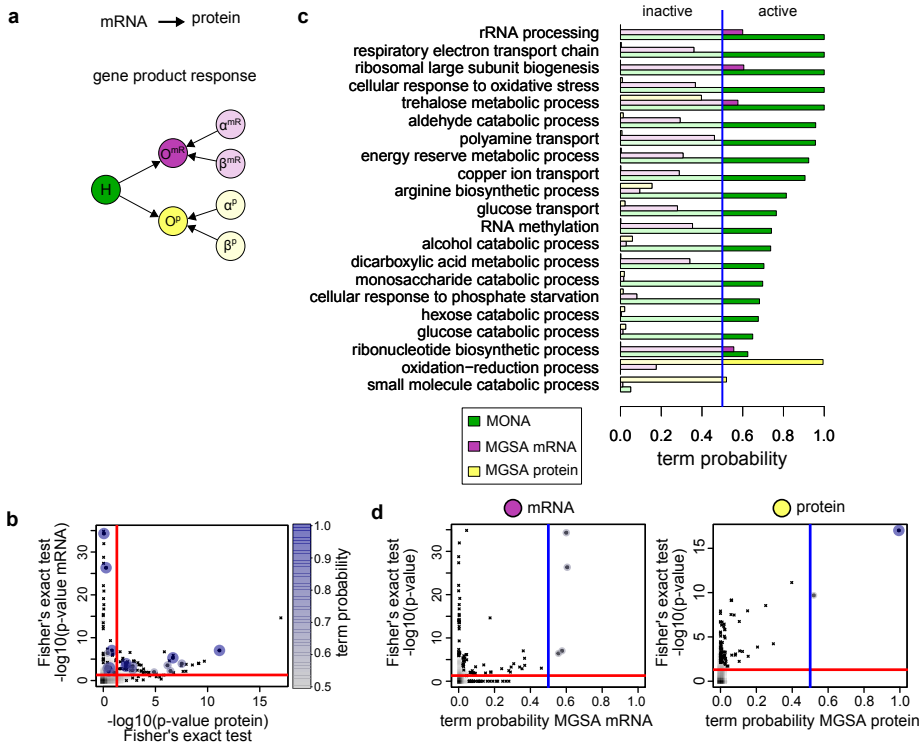


Figure 7.5: Analysis of mRNAs and proteins upon salt stress in yeast. **(a)** The cooperative model for mRNA (magenta) and protein (yellow) was used to specify the hidden gene response (green). **(b)** For each GO term, p-values of Fisher's exact test on mRNA and protein level are plotted against each other. Active terms resulting from MONA are marked as dots and are colour- and size-coded by its respective MONA term probability. **(c)** Probabilities of terms derived from MONA and MGSA on mRNA and protein level. **(d)** Term probabilities plotted against the p-values of Fisher's exact test for MGSA on mRNA and protein level. **(c-d)** Blue and red lines indicate probability of 0.5 and significance level of 0.05, respectively.

annotated to a GO term, 1,274 genes and 214 proteins were responding to osmotic shock.

The cooperative model is applicable to the present two-level study of gene and protein expressions (Fig. 7.5a). Here we assume that differential expression of a specific gene can be observed on both, mRNA and protein level. This was shown to hold especially for upregulated genes [97]. However, in practice it is possible that differential expression can only be observed in one of these levels due to measurement limitations or also biological reasons (imperfect correlation between mRNA and protein expression [29]). This is accounted for in

the generative model by introducing false positive and false negative rates (Fig. 7.2).

MONA yields probabilities for GO term for yeast response to osmotic shock, whereof we considered 19 GO terms to be active as their marginal posterior probability was greater than 0.5 (Fig. 7.5c). Amongst those terms, five terms had a probability of one to be active.

In order to investigate to what extent the probability of active terms depends on the cooperative influence of mRNA and protein activity, we first calculated p-values resulting from Fisher's exact test on mRNA and protein level separately (Fig. 7.5b). Most of the terms that were determined as active by MONA, were also significantly enriched among results of Fisher's exact test on both, mRNA and protein level. Expectedly, some terms were active with a high probability although they were only significant on mRNA level. This indicates that MONA uses the protein information to enhance the probability of certain terms but not necessarily dependent on it.

We next examine the biological relevance of active biological functions identified by MONA (Fig. 7.5c, green bars) starting with the most likely terms. The term *cellular response to oxidative stress* ($P = 1$) is consistent with the original study [97], which reported the general induction of stress response genes on both, mRNA and protein, levels. Typically there is a high overlap of genes for osmotic and oxidative stress [141], while the oxidative stress response is activated following the osmotic stress condition. A key gene known to be activated during this process is the oxidoreductase *GRE2* [141], which is also responding in the present study on both mRNA and protein level.

Another result of the original study was the induction of genes involved in trehalose metabolism [97], which was shown to be directly linked to the yeast stress response [68]. MONA identified the term *trehalose metabolic process* ($P = 1$) in good agreement with these findings. In the same context, MONA identified the following terms: *energy reserve metabolic process* ($P = 0.92$), *hexose catabolic process* ($P = 0.68$), *monosaccharide catabolic process* ($P = 0.70$), *glucose catabolic process* ($P = 0.65$), *alcohol catabolic process* ($P = 0.74$) and *glucose transport* ($P = 0.76$). In addition, the *respiratory electron transport chain* term ($P = 1$) is active under osmotic stress conditions arising also due to

the oxidative stress response. The activation of proteins involved in mitochondrial electron transport chain is crucial to counteract the production of reactive oxygen species upon salt stress [130]. The activity of *arginine biosynthetic process* ($P = 0.81$) is also in agreement with the literature, as it has been reported to be induced during oxidative stress [126]. Accordingly, the original study reported *amino acid biosynthesis* as being enriched in their analyses. Interestingly, MONA identified arginine as a more a specific amino acid to be active, which offers a more detailed insight to yeast stress response to an osmotic shock.

We finally compare MONA results to MGSA on mRNA and protein level, where only four and two terms were active, respectively. Terms identified on mRNA level alone were also considered as active by MONA, but had always lower probabilities < 0.6 (Fig. 7.5c, purple bars) and were also significantly enriched among the results of Fisher's exact test (Fig. 7.5d).

One of the two terms identified on protein level by MGSA (Fig. 7.5c, yellow bars) is *oxidation reduction process* which was also identified by mRNA MGSA ($P = 0.99$) and MONA. The other active term is *small molecule catabolic process* ($P = 0.52$). Interestingly MONA is able to identify the more specific child-term *respiratory electron transport chain*, which we have shown to be in agreement with literature. Both terms were also highly enriched at Fisher's exact test on protein level (Fig. 7.5d).

7.4 Analysis of multilevel gene responses during adipogenesis

In this section, we will show the results of MONA applied on the adipogenesis data. We will define a MONA model that is able to integrate mRNA, methylation and miRNA data simultaneously. Using this approach, we aim to identify processes, which are affected during adipogenesis and which may exhibit molecular changes on different levels.

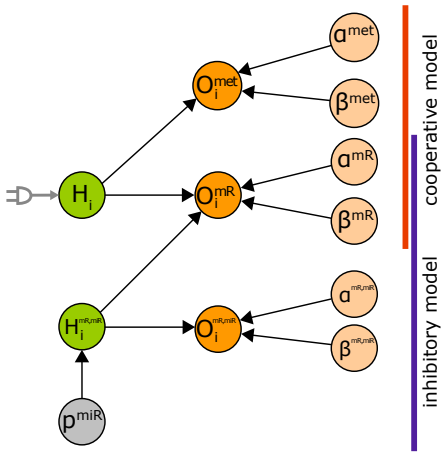


Figure 7.6: Three-level MONA model consisting of a cooperative part (mRNA + methylation; red) and an inhibitory part (mRNA + miRNA; blue). The hidden gene response is modeled by the independent observation of methylation and mRNA changes, whereas the mRNA response

7.5 Multilevel ontology analysis on adipogenesis data

In order to set up an appropriate framework for the integrative gene set analysis of mRNA, methylation and miRNA data, we combine the cooperative and the inhibitory model, which were described above (see Section 7.1.3). We obtained independent observations of mRNA and methylation response and can model them in a cooperative fashion (Fig. 7.6; cooperative model). The regulatory influence of miRNAs on the gene expression was determined by the miRlastic approach by integrating miRNA and mRNA expression measurements (Chapter 5.4). We do not regard miRNA regulation as an independent observation of gene response and model it via the inhibitory model (Fig. 7.6; inhibitory model). We refer to this appraoch as *three-level MONA*. The implementation of three-level MONA was performed in collaboration with Melanie Kopp in the course of the work for her master thesis, which is currently ongoing.

Initially, we assigned the differentially methylated CpG sites to associated genes according to the annotation of the chip manufacturer. We primarily observed hypomethylation of CpG sites during adipogenesis (Fig. 3.1) and therefore only selected CpG sites out of the set of differentially methylated ones,

which are hypomethylated during adipogenesis. Since hypomethylation of the DNA furthermore usually has an activating influence on the activity of associated genes [83], we selected only genes whose mRNA expression was up-regulated during adipogenesis. In addition, a gene was considered to be miRNA regulated, if it was targeted by a miRNA in the miRlastic network that was down-regulated during adipogenesis (Fig. 3.2).

In order to assign the genes to functional categories, we retrieved the human pathway annotations from *WikiPathways* [87]. We obtained a set of 2,978 genes that could be mapped to a certain functional category from Wikipathways of which we determined 352 genes whose mRNA was up-regulated, 19 genes with an associated hypomethylated CpG site and 384 genes which were considered to be targeted by an up-regulated miRNA. Only functional categories were considered that consisted of more than 10 assigned genes. In total, 218 functional categories were used as input for MONA together with the observed multi-level gene responses. We finally obtained a set of 14 functional categories with a term probability > 0.5, which we consider as active terms (Tab. 7.1).

In order to quantify the impact of each individual molecular level to the outcome of three-level MONA, we performed MONA runs for mRNA, methylation and miRNA data separately using the single-level model. The term probabilities of active terms determined by each of these runs in comparison to the outcome of three-level MONA are illustrated in Fig. 7.7.

We are able to reveal functional categories, which mainly correspond to the energy household machinery like *the citric acid (TCA) cycle and respiratory electron transport*, *Fatty Acid Biosynthesis*, *fatty acid, triacylglycerol, and ketone body metabolism* and *Mitochondrial LC-Fatty Acid Beta-Oxidation*. This is in accordance to the biological background of adipogenesis, since these processes are altered in adipocytes to become equipped with the ability for lipogenesis and lipolysis [113].

Most interestingly, we identified the functional category *Adipogenesis*, which actually comprises the whole set of molecular factors, which are known to play major roles in adipogenesis. It is worth noting that we could determine this functional category only by using the three-level model and not by any single model run. We observe indeed gene responses on all three molecular levels,

Table 7.1: Functional categories from WikiPathways with term probability > 0.5 obtained by three-level MONA. The term probability for each functional category is specified in the column "prob". The total amount of genes assigned to a functional category is listed under "sum". "mR" indicates the amount of genes with differentially up-regulated mRNA, "met" refers to the amount of genes with hypomethylated CpG site and "miR" to the amount of miRNA regulated genes assigned to the respective functional category.

ID	prob	name	sum	mR	met	miR
WP1817	1.00	Fatty acid, triacylglycerol, and ketone body metabolism	42	28	0	1
WP197	1.00	Cholesterol Biosynthesis	15	11	0	0
WP357	1.00	Fatty Acid Biosynthesis	21	18	0	0
WP2766	1.00	The citric acid (TCA) cycle and respiratory electron transport	29	11	0	0
WP465	1.00	Tryptophan metabolism	29	14	0	1
WP236	1.00	Adipogenesis	89	25	4	20
WP716	1.00	Vitamin A and Carotenoid Metabolism	21	9	0	3
WP325	0.97	Triacylglyceride Synthesis	18	9	1	1
WP368	0.92	Mitochondrial LC-Fatty Acid Beta-Oxidation	16	11	0	1
WP2788	0.84	Sphingolipid metabolism	25	6	0	1
WP24	0.61	Peptide GPCRs	24	6	0	0
WP100	0.60	Glutathione metabolism	16	4	0	0
WP241	0.56	One Carbon Metabolism	19	4	2	2
WP49	0.50	IL-2 Signaling Pathway	24	6	0	8

mRNA expression, DNA methylation as well miRNA regulation. However, these changes are not pronounced enough to yield any enrichment individually. The term actually becomes active only through the combination of the data from the individual levels.

We observe that the result of MONA using the three-level model is primarily dominated by the mRNA data. Most of the arising terms can also be determined by using the single model only. This is actually not surprising if we consider that mRNA expression measurements are a very comprehensive and meaningful data type for describing the gene activity. On the other hand, we do not observe any enrichment for the methylation data only. This indicates that there are no systematic changes in the methylation pattern that influence a process as a whole but rather affect only a small fraction of genes. On miRNA level, we primarily observe the activity of signaling pathways. This is also what we would expect, since we know that the most important role of miRNAs is the

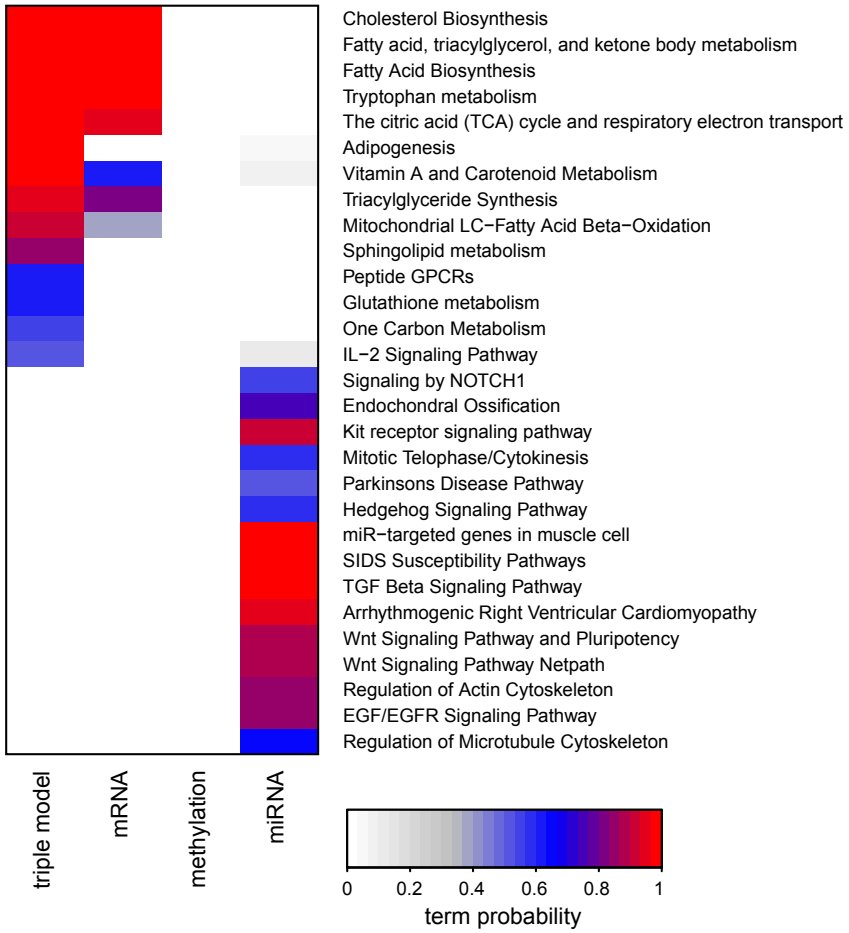


Figure 7.7: Comparison of term probabilities obtained by three-level MONA with results from single level runs. The color indicates the probability of the functional category in the respective run. No active terms could be obtained by applying single-level MONA on the methylation data individually.

regulation of signaling cascades [30]. However, these pathways do not seem to be affected on mRNA and methylation level. Hence, these terms are not considered as active by three-level MONA. As we model the miRNA influences via the inhibitory model, we do not account for the targeted genes in the same way as for the single level on miRNA data alone. In the three-level model, we rather aim to highlight the combinatorial effect of the levels and can thus consider terms to be specific for a certain level if they do not appear in the three-level model results. A good indicator for a miRNA-specific result on the miRNA single levels is that we obtain the term *miRtargeted genes in muscle cell*. We

could thus conclude that the changes in the regulation of signaling pathways are primarily carried out by miRNA activity. Interestingly, we also observe the *TGF Beta Signaling Pathway* as affected on miRNA level, which we already found in the LEA results (see Chapter 6.2). Note that we use LEA for the identification of *locally enriched* terms in the miRNA-target network, whereas we investigate *global effects* by applying the single-level MONA on the miRNA data. But as we observe a large fraction of miRNAs that cooperatively regulating this process (Fig. 6.6), it makes sense that this pathway is also determined to be globally affected.

7.6 Discussion and Conclusion

In this chapter, we introduced a modular framework for gene gene set analysis integrating multilevel omics data, which is named *multilevel ontology analyis* (MONA). The goal of this approach is the model-based identification of cellular processes through the assessment of gene responses across multiple large-scale molecular profiling experiments.

It is well known that a set of cellular processes is differently active among cells in different conditions. These conditions can be induced by an external stimulus but can also arise from different cell types or tissues. The activation of a certain cellular process in turn, implies the induction of a specific set of genes. We therefore expect that if a cellular process is active, the corresponding genes also respond to the condition. However, gene response is an abstract term, and we may observe it quite differently on different levels (e.g. mRNA, protein, methylation). Hence, we integrate gene response as latent variable in multi-omics observations. This concept is represented as a Bayesian network in MONA (Fig. 7.2).

Similar to common gene set analysis methods, MONA is applied to a given set of functional categories that may be retrieved from databases like KEGG [84] or GO [5]. However, the model-based approach has the great advantage over ordinary enrichment methods that it infers the activity of cellular processes across all given terms simultaneously, thereby accounting for term redundancies and related multiple testing problems [110]. Our method therefore combines

two benefits: the appropriate handling of integrated data from any omics level, while in parallel coping with term redundancies.

The models introduced in this chapter plugged to the base model are only a subset of possible models. For example, methylation, mRNA and protein levels can be inferred simultaneously using a cooperative model with three observations. In addition, the design allows us to implement additional models to simultaneously capture different molecular levels (Fig. 7.1). For example, when measuring proteins and the metabolome of cells, we may introduce a third "activating" model, where e.g. an existing metabolite may have an activating (unlike an inhibiting) effect on a proteins activity. Protein phosphorylation levels may also serve as activating evidence of a proteins function. Even complex gene interactions may be a basis for a model that could be plugged to the hidden gene response. The development of more and more powerful techniques for the inference of gene interactions [136] leads to a comprehensive and reliable knowledge of gene regulation and may improve the outcome of the MONA algorithm. Another improvement could also be achieved by introducing a weighted variant of MONA. Here, the magnitude of the fold change between different conditions could be considered in order to infer the hidden gene response.

We evaluated our approach on both, synthetic and real data in order to emphasize the advantages of combining data from multiple levels over individual analyses. We can show that MONA outperforms single-level approaches in terms of arising false positive and false negative terms when applied on synthetic data. Furthermore, we determined a more reliable set of functional categories from the application on previously published data.

After thorough validation of our method, we combined the cooperative and the inhibitory model to set up an appropriate model for the application of MONA on the adipogenesis dataset, which comprises mRNA expression, DNA methylation and miRNA expression data. Initially, we assessed gene responses on mRNA and methylation level using standard methods. In addition, we included the previously generated miRlastic results in order to obtain a set of genes, which is considered to be regulated by adipogenesis-associated miRNAs. We showed that even though the results of the multilevel gene set analysis is dominated by the response on mRNA level, the most comprehensive and mean-

ingful functional category can only be obtained by the combinatorial effect of multiple molecular levels.

Chapter 8

RAMONA: Remotely accessible multilevel ontology analysis

In the previous section, we introduced MONA, a model-based enrichment analysis approach for the integration of omics data from multiple molecular levels. In this chapter, we introduce the remotely accessible multilevel ontology analysis (RAMONA), which is a web interface for MONA. By providing an easily accessible implementation of MONA in combination with a comprehensive database structure, we aim to facilitate the application of the MONA approach for any applied researcher dealing with combined large-scale omics data. In addition, RAMONA enhances the MONA output with valuable information. We will describe the implementation of RAMONA and how it can be used to perform gene set analysis on a dataset derived from either one or two omics levels. We will then introduce the structure of the database, which provides ontology and mapping information for processing the input data. The output of RAMONA, which we will show subsequently, provides functional insight into the activity of biological processes across the given molecular levels. Finally, we will present the application of RAMONA on the adipogenesis data, thereby showing how the output of MONA can be used to visualize areas in signaling pathways,

which are affected on more then one molecular level. RAMONA is available at http://icb.helmholtz-muenchen.de/ramona. This chapter was published in parts in:

- **Steffen Sass**, Florian Buettner, Nikola S. Mueller, and Fabian J. Theis. RAMONA: a web application for gene set analysis on multilevel omics data. *Bioinformatics*, under review.

Preliminary work on the web interface has been done by Benedikt Rauscher and Michael Schollerer during a practical course, which was supervised during this work.

8.1 RAMONA: A web application for multilevel ontology analysis

Decreasing costs of large-scale molecular profiling studies, such as transcriptomics or proteomics, allow for the joint analysis of several molecular levels in parallel. The crucial step in the analysis of such diverse data is to combine the different levels such that a comprehensive insight in the response of genes to these conditions can be assessed. This in turn can be directly linked to the underlying biological processes affecting the activity of genes on several molecular levels. However, these kind of analyses are not straight-forward and often the molecular levels are treated as independent in order to allow for the use of single-omics analysis techniques.

In practice, gene response is initially determined by using statistical methods. Among the resulting set of altered genes one usually searches for overrepresented biological processes by applying common gene set enrichment methods [18, 166] that incorporate functional annotations from databases like Gene Ontology (GO) [5] or the Kyoto Encyclopedia of Genes and Genomes (KEGG) [84]. Even though there exists a multitude of easy-to-use web-based enrichment tools [73, 187], they are only capable of analyzing a single molecular level. Furthermore, no web tool is available that properly deals with term redundancies appearing frequently e.g. due to the tree structure of GO.

In order to provide a powerful method to integrate multi-level gene response

data for the determination of altered biological processes, we recently introduced the *multilevel ontology analysis (MONA)* [154]. MONA is a model-based Bayesian method, which is able to integrate data sets from multiple molecular levels by simultaneously dealing with term redundancies and related multiple testing problems.
However, the usage of the standalone MONA application can be a cumbersome process, since the user has to specify the data structure of the activated genes and their term annotations by himself and lacks a comprehensive visualisation of the results. Furthermore, it can only be run on Windows machines as it depends on the .NET library.
Here we introduce a web-based implementation of MONA, called *remotely accessible MONA (RAMONA)*, which is designed with the focus on practical usability for any applied researcher. It offers three models to analyze most common experimental setups. The web interface is capable of processing many given gene identifiers as well as of automatically mapping them to widely used ontologies derived from GO and KEGG. The detailed output of RAMONA includes an interactive visualisation of the inferred active terms in the context of their respective pathways or ontology hierarchy. This provides functional insight into the activity of biological processes and the role of associated genes responding to the given conditions by providing relevant details on the resulting processes.

8.1.1 Implementation

RAMONA is a web-based application whose interface is implemented in the Mono ASP.NET framework. The underlying MONA application is written in C# and is based on the Infer.NET framework [119].

MONA currently provides three models of molecular interactions (Fig. 8.1). The single level model, which can be used when measurements are only available on a single level. This corresponds to the principle of the Model-based Gene Set Analysis (MGSA) [9]. The cooperative model accounts for studies where measurements of two different levels are available, which may be regarded as independent noisy observations (e.g. mRNA and protein) of an underlying common gene response. The inhibitory model is applicable when two species are measured, but they could not be interpreted as independent measurements

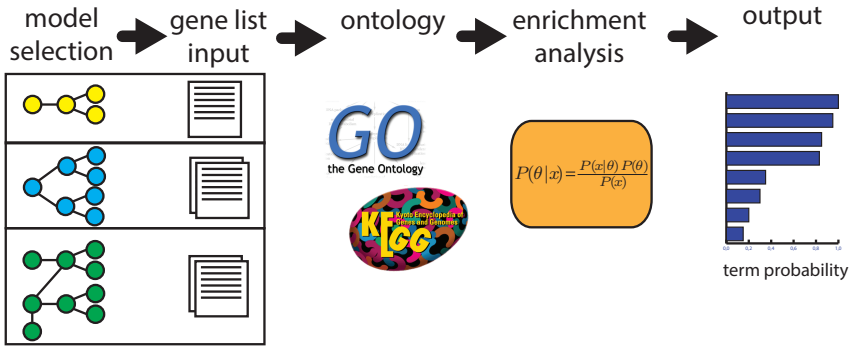


Figure 8.1: RAMONA workflow. The user has to specify the input according to the selected model. He can choose between GO and KEGG as ontologies. Using a Bayesian modelling approach, the tool is able to infer non-redundant enriched terms among the given gene lists.

of the hidden gene function. A prominent example is the post-transcriptional modulation of mRNA expression by miRNAs.

Given the user input, the MONA algorithm infers the marginal posterior probability of the term activity using a Bayesian network as described in [154]. The user has to specify the input of RAMONA according to the selected model. In general, this must be a set of genes that show a special behavior like the response to a certain condition and a set of measured genes which is referred to as background. A typical example for an input would be two lists of differentially expressed genes between two conditions for both, mRNA and protein level. For the cooperative model two lists of differentially expressed genes together with a background of all measured genes have to be provided. The probabilistic nature of RAMONA allows for the analysis of experiments, where different numbers of genes are measured (e.g. usually the case for mRNA and protein data). For the inhibitory model, a set of inhibited genes has to specified in addition to the responding genes and background. All these sets can be provided by text field input or text file upload. The user can manipulate the shape of all priors via the expert settings in order to e.g. encourage a sparser result; uniform priors are used as default settings for the single and cooperative case. In case of the inhibitory model, weakly informative priors are used as discussed previously [154].

RAMONA supports a variety of common gene identifiers for several organ-

isms that are mapped to specific terms. These terms include biological processes, molecular functions and cellular components from GO [5] as well as pathways from KEGG [84].

The actual MONA process runs in a background thread on the web server with runtime depending on the size of the input and the selected ontology. For common setups RAMONA runs not longer than one minute. In addition to the term probabilities provided by the model-based enrichment analysis, p-values for enrichment of the individual terms are calculated by using Fisher's exact test on each molecular level separately when the cooperative model is chosen.

8.1.2 Database structure

In order to allow for a flexible usage of RAMONA, we provide a MySQL database, which is storing several types of identifiers from different species as well as functional categories from GO or KEGG. The structure of the database is designed in way that allows for the easy retrieval and mapping of information on genes and ontologies (Fig. 8.2).

The central table of our database is the "gene" table, which stores all known genes for each species. The gene information is retrieved from the National Center for Biotechnology Information (NCBI) Gene database [114] and is indexed by the corresponding Entrez identifier. These genes are mapped to terms via the "gene2term" table in a many-to-many fashion. The terms are retrieved from GO and KEGG, respectively, which correspond to an "ontology". Information on ontologies is stored in a separate table. In order to allow for the input of several types of gene identifiers, we added another table called "identifier", which stores all identifiers that can be mapped to Entrez gene identifiers. Since the mapping of foreign identifiers to Entrez genes is not always unique, we also map them in a many-to-many fashion via the "identifier2gene" table. Besides Entrez gene ids, we currently support Ensembl ids [40], HGNC gene symbols [53] and Uniprot ids [168].

8.1.3 Output format

The output of RAMONA consists of three parts: a plot panel, a table and a panel for further term information (Fig. 8.3). If the cooperative model was

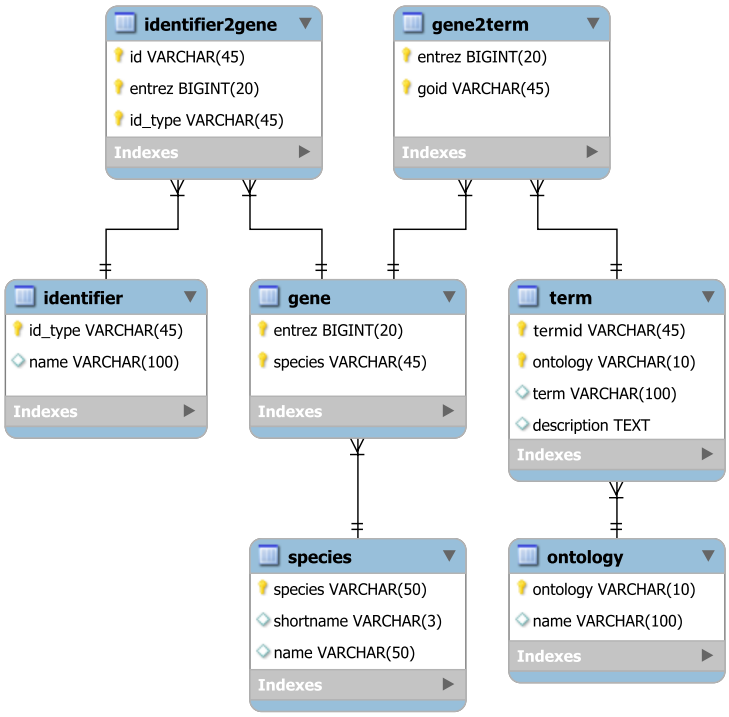
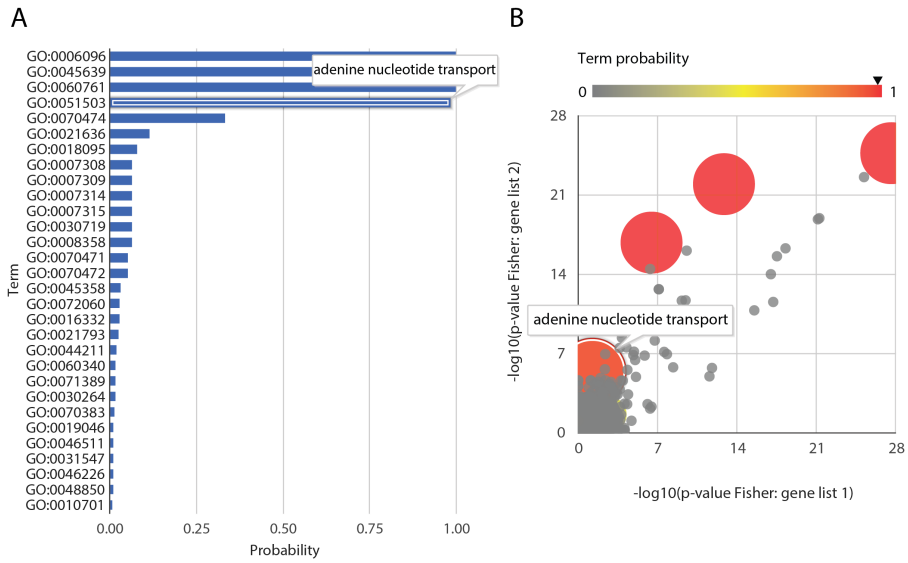


Figure 8.2: RAMONA database structure. Genes are linked to terms and foreign identifiers in a many-to-many fashion. The structure allows to map from several identifiers to the internal gene representation. In addition, different species and ontologies are supported.

chosen, the user can switch between a bar plot (Fig. 8.3A) and a scatter plot (Fig. 8.3B) to illustrate the results of RAMONA. Otherwise, only the bar plot is shown, which displays the term probabilities for the top 30 terms. The scatter plot displays the p-values of the Fisher tests, which are performed on the two input lists individually, in comparison to the term probabilities. This representation allows the user to determine the effect of the two individual input gene lists on the RAMONA outcome. In addition, it uncovers the redundancies that arise from the traditional gene set analyses and which do not appear in the RAMONA results. The table (Fig. 8.3C) provides an overview of all relevant information on the terms, namely the number of assigned genes and the number of altered genes in the given gene list(s). Additionally, the percentage of assigned genes is shown which were missing in the smaller background set.

By selecting a term, either in the bar plot, scatter plot or table, detailed information for the respective term can be displayed (Fig. 8.3D). This includes



C

ID	Probability	Term	# Background	# Gene list 1	# Gene list 2	not measured
GO:0006096	1	glycolysis	63	40	38	0.0%
GO:0045639	1	positive regulation of myeloid cell differentiation	64	27	36	0.0%
GO:0060761	1	negative regulation of response to cytokine stimulus	31	13	22	0.0%
GO:0051503	0.98	adenine nucleotide transport	5	2	5	0.0%

D

adenine nucleotide transport

The directed movement of adenine nucleotides, ATP, ADP, and/or AMP, into, out of or within a cell, or between cells, by means of some agent such as a transporter or pore.

Genes altered in list 1 (2 / 5)
CALHM1, SLC25A17

Genes altered in list 2 (5 / 5)
CALHM1, GJA1, SLC25A17, SLC25A42, SLC35B2

0,0% of genes in the smaller background set were missing.

Term probability: 0.98
p-value gene list 1: 5.63e-2 (Bonferroni adj. p=1.00e+0)
p-value gene list 2: 4.04e-6 (Bonferroni adj. p=4.96e-2)

Figure 8.3: The RAMONA output. **A** Resulting term probabilities are shown in a bar plot. **B** If the cooperative model is chosen, a scatter plot can be displayed that shows the p-values for each term determined by traditional gene set analysis (Fisher's exact test) on the two input gene lists individually. The color and size of the points correspond to the RAMONA term probability. **C** The tabular representation gives an overview of all relevant term information. **D** Additional information can be obtained by clicking on the terms in any of the three panels. This information includes the set of altered genes for each level as well as the decision whether a term is active (red) or not (black) for RAMONA or Fisher's exact test.

for each molecular level a list of regulated genes assigned to this term as well as the percentage of missing genes in case of the cooperative model. Furthermore, a link to the term database is provided, which allows for a graphical mapping

of the results. In case of KEGG, the respective pathway is displayed and the regulated genes are marked by a color for each molecular level. If GO was selected, the GO tree will be shown illustrating the term hierarchy including all active terms (probability >0.5).

8.2 Adipogenesis pathway analysis

We applied RAMONA on the mRNA expression and DNA methylation data from the adipogenesis dataset to demonstrate the functionality of our method. As a basis, we used the gene lists, which were generated by the statistical analysis (see Section 7.4). As above, we selected only genes whose mRNA was up-regulated during adipogenesis as well as genes with an associated hypomethylated CpG site. In total, we used the Entrez identifiers of 384 genes with up-regulated mRNA and 19 genes with hypomethylatd CpG sites as input for RAMONA. We chose the human genome as background set for both input lists and the cooperative model as input model. In addition, we placed an informative prior on p in order to put the probability mass on lower values (see Section 7.1).

In Chapter 7, we aimed to identify affected pathways using triple level MONA and WikiPathways. Here, we use the cooperative model and KEGG pathways as ontology. The miRNA information is therefore not included. As a result of the combined input data, we obtained 10 pathways, which were considered as active (term probability > 0.5), namely *steroid biosynthesis*, *valine, leucine and isoleucine degradation*, *biosynthesis of unsaturated fatty acids*, *PPAR signaling pathway*, *vitamin digestion and absorption*, *Tyrosine metabolism*, *Glyoxylate and dicarboxylate metabolism*, *Propanoate metabolism*, *citrate cycle (TCA cycle)*, *Terpenoid backbone biosynthesis* and *Pyruvate metabolism*. The PPAR signaling pathway, which includes the activation of PPARγ, is the key mechanism in adipogenesis [145]. Hence, we selected this pathway for visualization within the RAMONA framework (Fig. 8.4).

We observe that almost all targets of PPARγ, as well as PPARγ itself, are up-regulated in adipocytes on mRNA level (blue nodes). Clearly affected are the factors associated with the processes lipogenesis, fatty acid transport

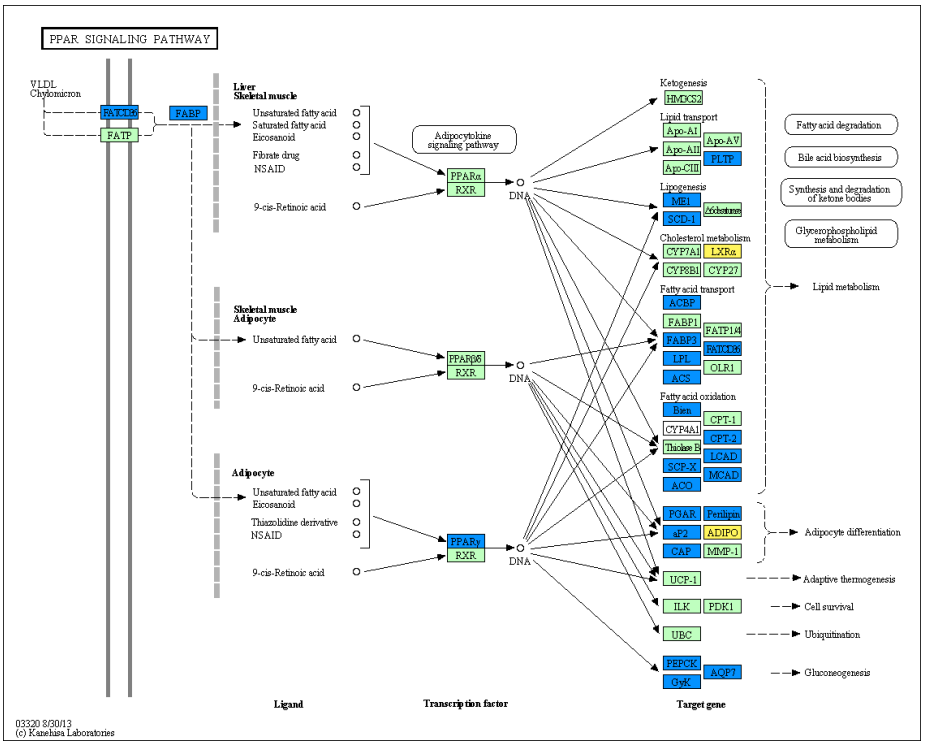


Figure 8.4: PPAR signaling network from KEGG [84], which was considered as active by RAMONA on the two sets consisting of genes with differentially up-regulated mRNA and hypomethylated promoters, respectively. The gene nodes in the KEGG pathway are colored by RAMONA according to the presence in the two gene sets. Blue nodes correspond to genes, which are contained in the list of genes with differentially up-regulated mRNA. Yellow nodes correspond to genes, which are part of both input gene lists. Green nodes are not part of any input gene list. White notes are not part of the human pathway.

and oxidation as well as gluconeogenesis. Interestingly, almost all of the factors, which are associated with adipocyte differentiation, are up-regulated, too. Among these factors, we observe that adiponection (ADIPO) is not only altered on mRNA level, but also on methylation level (yellow nodes). Hence, we can assume a regulatory effect of the DNA methylation on the mRNA expression of the adiponectin gene. Recently, a relationship between reduced methylation of the adiponectin promoter and the combined presence of obesity and insulin resistance has been reported [47]. This supports our finding that changes in the methylation pattern could actually regulate the activity of adiponectin, thereby playing an important role in adipocyte differentiation. Abnormal methylation patterns could then be a reason for irregular adipogenesis, which in turn may

cause metabolic disorders like diabetes. We furthermore observe mRNA and methylation changes for the LXRα gene. Differences in the methylation status of the LXRα genes have also already been reported to be associated with diabetes in rats [22].

8.3 Discussion and Conclusion

In this chapter, we introduced RAMONA, which is a web application for the increased usability of the MONA framework. We integrated the single, cooperative and inhibitory model in the RAMONA environment. RAMONA is built upon a database, which is used for the mapping of gene identifiers and for the association between genes and functional categories. In addition, it provides an enhanced output of the MONA results, which includes the coloring of KEGG pathways as well as the generation of term trees for GO.

The database structure of RAMONA is designed in a way that allows for an easy extension of the provided ontologies and identifiers. We therefore aim to successively increase the amount of ontologies for example by functional categories from WikiPathways [87]. In addition, the implementation of further models could be considered.

We showed that the application of RAMONA on the mRNA expression and DNA methylation data from the adipogenesis dataset allows us to gain functional insight into the regulatory mechanism of adipocyte differentiation. We could boil down the observed changes on mRNA and methylation level to a set of active pathways and furthermore identify two genes, which are supposed to be regulated by methylation changes. These genes are already known to play a role in insulin resistance and might be potential subjects for further investigations with regard to epigenetic regulation of adipogenesis.

Chapter 9

Summary & Outlook

The advances in large-scale experimental techniques for assessing molecular profiles in a comprehensive fashion enabled us to investigate molecular interactions not only on a small well-defined scale, but also on a system-wide level. The ability to measure various types of molecular features like RNA and protein expression, DNA methylation or metabolite concentration furthermore allows us to model complex regulatory relationships across different levels of gene activity. Hence, we can gain a comprehensive understanding of gene regulatory mechanisms in biological processes, which may be crucial for the investigation of complex diseases. These processes include the differentiation of preadipocytes into mature adipocytes, which are important players in the human energy homeostasis. It has been shown that adipocyte dysfunctions are directly linked to diseases such as type 2 diabetes mellitus or cardiovascular diseases [59]. The profound knowledge of complex molecular interactions throughout this process is thus necessary to improve the understanding of such multifactorial diseases. Even though the adipocyte differentiation process has been subject of many studies already, its molecular properties are still not fully understood [145].

In this thesis, we investigated this process by integrating multilevel large-scale molecular profiling data. We aimed to identify regulatory interactions between different molecular features, which are relevant among the different mechanisms of adipogenesis. The underlying study incorporated large-scale measurements of mRNA and miRNA expression as well as DNA methylation, which were

assessed prior and after the differentiation process of human adipocytes. In general, these datasets comprise a large amount of data whose joint analysis is not straight-forward. In order to build up an appropriate analysis pipeline for these data, we introduced novel analysis techniques for the integration of multilevel large-scale molecular profiling data together with prior knowledge from different sources.

9.1 Summary

One important aspect in the differentiation process of adipocytes is the post-transcriptional gene regulation through miRNAs [188]. The interplay between miRNAs and mRNAs itself is a complex process, which is under the control of further regulatory levels. We therefore initially aimed to reveal properties of miRNA regulation itself. We could show that miRNA regulation often happens in a coordinated fashion, which leads to an increased coexpression among miRNAs that target proteins from a common protein complex (Chapter 4). In addition, we showed that targets of clustered miRNAs tend to be connected more densely in a protein-protein interaction network.

By considering our findings on coordinated miRNA regulation, we proposed a novel method called *miRlastic* (Chapter 5) to determine regulatory relationships between miRNAs and targets, which are relevant for the adipocyte differentiation process, on the basis of combined miRNA and mRNA expression data and *in silico* target predictions. We used a multiple regression analysis approach with elastic net penalty, which accounts for the co-expression among miRNAs arising due to coordinated gene regulation. We could show that our approach outperforms other common methods in the identification of previously validated target interactions and in terms of arising false positives and false negatives on simulated data.

Based on the inferred miRNA-mRNA regulatory relationships, we aimed to determine whether adipogenesis-associated miRNAs play individual roles in certain biological processes. For this purpose, we introduced a method for local enrichment analysis within miRNA-target networks using functional gene annotations (Chapter 6). As we are able to assess the interaction strength between

miRNAs and mRNAs, we included this information by calculating the shortest paths among the genes in the network. We could identify certain functional categories in our previously generated miRNA-mRNA network, which are significantly locally enriched. We then assigned the enriched regions in our network to the corresponding miRNAs. By doing so, we could show that certain groups of miRNAs tend to act together individually in regulating essential processes of adipogenesis. Especially the miR-30 family, which is up-regulated during adipogenesis, exhibited a conspicuous functional role in adipogenesis, which has been also already suggested in the literature [186, 173].

As a next step, we introduced a multilevel enrichment analysis, which allowed us to identify processes, which are affected during adipogenesis across different molecular levels (Chapter 7). Since we may describe alterations of gene activity not only on each of these levels separately, our approach combines them in order to provide a better functional insight into the underlying molecular mechanisms. Using this information, we could in turn infer the individual effect of changes in mRNA expression, DNA methylation and miRNA regulation on the respective process. We showed that the joint model-based enrichment analysis is actually able to determine relevant processes for adipogenesis, which we do not observe on any of the single levels. We designed our approach in a modular fashion, which means that it can be easily adapted to any given experimental setup in order to account for the corresponding molecular features. It can serve as a powerful tool for the joint analysis of molecular profiling data, which might be used by any applied researcher dealing with such kinds of data.

We thus introduced an implementation of our method in form of a web application to facilitate the usage of our approach (Chapter 8). Our web application can process several kinds of input gene identifiers for different species by including a database that holds information from various resources. The database is furthermore used to map the given input to functional categories. Whereas the user has to specify this information in the standalone version, the web application provides an automated mapping of the data. In addition, the enhanced output makes it possible to determine the combined effects of the gene responses across the given molecular levels. Using this approach, we were finally able not only to infer affected pathways on the combined mRNA and methylation lev-

els, but also to identify specific genes among these pathways, which might be epigenetically regulated factors in the adipocyte differentiation process.

9.2 Outlook

For the methods and findings presented in this thesis, several extension may be possible, which we will discuss in this section.

9.2.1 Methodological extensions

MiRlastic

- We tested whether we could improve our results by taking transcription factor activity into account, which we define as the corresponding mRNA expression. However, especially in mammalian systems, the activity of transcription factors is not only determined on transcriptional level, but rather post-translationally [78]. Instead of using the mRNA expression to assess the transcription factor activity, we could also consider the integration of further experimental data e.g. from phosphoproteomics experiments [144].
- Since the techniques for measuring large-scale protein profiles are constantly improving [4], we might consider the application of miRlastic not only on combined miRNA-mRNA data but also on miRNA-protein data. This would allow us to to take not only mRNA degradation by miRNAs into account, but also translational repression.
- Currently, miRlastic is implemented as a package for R [137]. To facilitate the usability of miRlastic, we could provide a web implementation together with an optimized visualization of the output.

LEA

- We designed LEA specifically for the application on weighted miRNA-target networks. However, the LEA approach could be also extended for the application on any network that contains nodes with annotated terms.

- LEA is currently available as implementation for R [137]. Since the outcome of LEA especially benefits from an appropriate visualization, we could link it directly to a visualization tool. For this purpose, we could consider the implementation of a plug-in for *Cytoscape* [158].

MONA

- The input of MONA currently consists of a boolean list, which indicates whether a gene is altered on a certain level or not. However, a continuous measure for this alteration would be more appropriate, since we may account for the fact that larger changes might have stronger effects than smaller ones. In addition, we do not necessarily depend on an arbitrary threshold for differential changes in gene activity. Our method could be improved by integrating the direction of these changes. By knowing whether the activity of a gene is increased or decreased, we could considerably improve the modeling of the interactions between different molecular levels.
- The development of more and more powerful techniques for the identification of gene interactions allows us to determine regulatory relationships between the genes in our model [136]. By integrating these interactions into the MONA model, we could further improve the inference of term activity.
- We could further extend the joint analysis of multiple datasets towards the integration of additional molecular levels that can describe the activity of genes such as SNPs, metabolites or post-translational modifications. For this purpose, we could adjust the MONA model for example to account for enzymatic activity of genes.

RAMONA

- The RAMONA database currently includes functional categories from KEGG [84] and GO [5]. We could extend the set of functional annotations by additional ontologies which may be retrieved from resources such as WikiPathways [87] or REACTOME [26]. We could furthermore add additional identifiers and species.

- The output of RAMONA could be further enhanced in order to provide insight into the contribution of the individual molecular levels to the resulting term activity. This could be achieved by an additional plot that highlights the outcome of the enrichment analysis on the single levels. By showing the results of Fisher's exact test on the single levels, we could furthermore point out the reduction of redundancy by using our model-based approach.

9.2.2 Follow-up studies on adipogenesis

- In the underlying study we focused on the molecular changes in adipogenesis. However, further phenotypic information could be included to enhance the results such as the insulin sensitivity of the probands. This could provide novel insights into the onset of type 2 diabetes.

- By integrating mRNA and miRNA data we were able to determine potential miRNA-target relationships for altered miRNAs. To validate our findings, experimental proof for the differential expression of the miRNAs as well as for the miRNA-target interactions should be provided. To experimentally evaluate the functional effect of the miRNAs, transfection or knock-down experiments would be necessary.

- Using three-level MONA, we find that all three molecular levels, mRNA expression, DNA methylation and miRNA regulation, jointly influence the process of adipogenesis. It would be worthwhile to further elucidate the regulatory interplay between the associated genes, especially with regard to epigenetic regulation.

9.3 Conclusion

In this thesis, we addressed the problem of data integration from different molecular profiling experiments in order to provide useful tools for the identification of functional and regulatory relationships between molecular features. Using our methods, we were able to reveal molecular mechanisms, which are involved in the differentiation of adipocytes. On the one hand, we could confirm previ-

ously gathered knowledge on adipocyte differentiation, but on the other hand also reveal novel insights. In conclusion, we showed that our proposed methods are able to provide valuable results when applied on combined data from different molecular levels. We made user-friendly implementations of our methods available for the research community in order to support the data analysis of future studies.

Bibliography

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, 19(6):716–723, Dec 1974.

[2] A. Alexa, J. Rahnenfhrer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, 22(13):1600–1607, Jul 2006.

[3] F. Allantaz, D. T. Cheng, T. Bergauer, P. Ravindran, M. F. Rossier, M. Ebeling, L. Badi, B. Reis, H. Bitter, M. D'Asaro, A. Chiappe, S. Sridhar, G. D. Pacheco, M. E. Burczynski, D. Hochstrasser, J. Vonderscher, and T. Matthes. Expression profiling of human immune cell subsets identifies mirna-mrna regulatory relationships correlated with cell type specific expression. *PLoS One*, 7(1):e29979, 2012.

[4] A. F. M. Altelaar, J. Munoz, and A. J. R. Heck. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics*, 14(1):35–48, Dec 2012.

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

[6] D. Baek, J. Villn, C. Shin, F. D. Camargo, S. P. Gygi, and D. P. Bartel. The impact of micrornas on protein output. *Nature*, 455(7209):64–71, Sep 2008.

[7] T. Bammler, R. P. Beyer, S. Bhattacharya, G. A. Boorman, A. Boyles, B. U. Bradford, R. E. Bumgarner, P. R. Bushel, K. Chaturvedi, D. Choi, M. L. Cunningham, S. Deng, H. K. Dressman, R. D. Fannin, F. M. Farin, J. H. Freedman, R. C. Fry, A. Harper, M. C. Humble, P. Hurban, T. J. Kavanagh, W. K. Kaufmann, K. F. Kerr, L. Jing, J. A. Lapidus, M. R. Lasarev, J. Li, Y.-J. Li, E. K. Lobenhofer, X. Lu, R. L. Malek, S. Milton, S. R. Nagalla, J. P. O'malley, V. S. Palmer, P. Pattee, R. S. Paules, C. M. Perou, K. Phillips, L.-X. Qin, Y. Qiu, S. D. Quigley, M. Rodland, I. Rusyn, L. D. Samson, D. A. Schwartz, Y. Shi, J.-L. Shin, S. O. Sieber, S. Slifer, M. C. Speer, P. S. Spencer, D. I. Sproles, J. A. Swenberg, W. A. Suk, R. C. Sullivan, R. Tian, R. W. Tennant, S. A. Todd, C. J. Tucker, B. Van Houten, B. K. Weis, S. Xuan, H. Zarbl, and M. o. t. T. R. C. . Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods*, 2(5):351–356, May 2005.

[8] S. Baskerville and D. P. Bartel. Microarray profiling of micrornas reveals frequent coexpression with neighboring mirnas and host genes. *RNA*, 11(3):241–247, Mar 2005.

[9] S. Bauer, J. Gagneur, and P. N. Robinson. GOing bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Research*, 38(11):3523–3532, Jan. 2010.

[10] D. Beck, S. Ayers, J. Wen, M. B. Brandl, T. D. Pham, P. Webb, C.-C. Chang, and X. Zhou. Integrative analysis of next generation sequencing for small non-coding rnas and transcriptional regulation in myelodysplastic syndromes. *BMC Med Genomics*, 4(1):19, 2011.

[11] J. R. Beckwith. Regulation of the lac operon. recent studies on the regulation of lactose metabolism in escherichia coli support the operon model. *Science*, 156(3775):597–604, May 1967.

[12] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[13] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, and et al. High density DNA methylation array with single cpg site resolution. *Genomics*, 98(4):288–295, Oct 2011.

[14] M. Bibikova, J. Le, B. Barnes, S. Saedinia-Melnyk, L. Zhou, R. Shen, and K. L. Gunderson. Genome-wide DNA methylation profiling using infinium assay. *Epigenomics*, 1(1):177–200, Oct 2009.

[15] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[16] C. E. Bonferroni. *Il calcolo delle assicurazioni su gruppi di teste*. Tipografia del Senato, 1935.

[17] F. Bost, M. Aouadi, L. Caron, and B. Bintruy. The role of mapks in adipocyte differentiation and obesity. *Biochimie*, 87(1):51–56, Jan 2005.

[18] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. GO::TermFinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics (Oxford, England)*, 20(18):3710–3715, Dec. 2004.

[19] K. Bryan, M. Terrile, I. M. Bray, R. Domingo-Fernandz, K. M. Watters, J. Koster, R. Versteeg, and R. L. Stallings. Discovery and visualization of mirna-mrna functional modules within integrated data using bicluster analysis. *Nucleic Acids Res*, 42(3):e17, Feb 2014.

[20] F. M. Buffa, C. Camps, L. Winchester, C. E. Snell, H. E. Gee, H. Sheldon, M. Taylor, A. L. Harris, and J. Ragoussis. microrna-associated progression pathways and potential therapeutic targets identified by integrated mrna and microrna expression profiling in breast cancer. *Cancer Res.*, 71(17):5635–45, 2011.

[21] J. C. Caruso and N. Cliff. Empirical size, coverage, and power of confidence intervals for spearman's rho. *Educational and Psychological Measurement*, 57(4):637–654, Aug 1997.

[22] Y. Cheng, G. Liu, Q. Pan, S. Guo, and X. Yang. Elevated expression of liver x receptor alpha (LXRα) in myocardium of streptozotocin-induced diabetic rats. *Inflammation*, 34(6):698–706, Dec 2011.

[23] S. W. Chi, J. B. Zang, A. Mele, and R. B. Darnell. Argonaute hits-clip decodes microrna-mrna interaction maps. *Nature*, 460(7254):479–486, Jul 2009.

[24] L. Choy and R. Derynck. Transforming growth factor-beta inhibits adipocyte differentiation by smad3 interacting with ccaat/enhancer-binding protein (c/ebp) and repressing c/ebp transactivation function. *J Biol Chem*, 278(11):9609–9619, Mar 2003.

[25] J. Cox and M. Mann. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics*, 13(Suppl 16):S12, Nov. 2012.

[26] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, and et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(Database):D691–D697, Jan 2011.

[27] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[28] R. B. Darnell. Hits-clip: panoramic views of protein-rna regulation in living cells. *Wiley Interdiscip Rev RNA*, 1(2):266–286, 2010.

[29] L. M. F. de Godoy, J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Frhlich, T. C. Walther, and M. Mann. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, 455(7217):1251–1254, Oct 2008.

[30] G. Di Leva, M. Garofalo, and C. M. Croce. Micrornas in cancer. *Annu Rev Pathol*, 9:287–314, 2014.

[31] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, Dec 1959.

[32] T. A. Down, V. K. Rakyan, D. J. Turner, P. Flicek, H. Li, E. Kulesha, S. Grf, N. Johnson, J. Herrero, E. M. Tomazou, and et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotechnology*, 26(7):779–785, Jul 2008.

[33] P. Du, X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11(1):587, 2010.

[34] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[35] E.N.C.O.D.E Project Consortium, B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

[36] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks. Microrna targets in drosophila. *Genome Biol*, 5(1):R1, 2003.

[37] S. Falcon and R. Gentleman. Using gostats to test gene lists for go term association. *Bioinformatics*, 23(2):257–258, Jan 2007.

[38] R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87, Jan 1922.

[39] R. A. Fisher et al. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

[40] P. Flicek, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, and et al. Ensembl 2014. *Nucleic Acids Research*, 42(D1):D749–D755, Jan 2014.

[41] P. L. Flom and D. L. Cassell. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. In *NorthEast SAS Users Group Inc 20th Annual Conference: 11-14th November 2007; Baltimore, Maryland*, 2007.

[42] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, Dec 2007.

[43] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, 33(1):1–22, 2010.

[44] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel. Most mammalian mrnas are conserved targets of microrna. *Genome Res*, 19(1):92–105, Jan 2009.

[45] W. J. Fu. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397, Sep 1998.

[46] G. M. Furnival and R. W. Wilson. Regressions by leaps and bounds. *Technometrics*, 16(4):499–511, Nov 1974.

[47] M. C. García-Cardona, F. Huang, J. M. García-Vivas, C. López-Camarillo, B. E. Del Río Navarro, E. Navarro Olivos, E. Hong-Chong, F. Bolaños Jiménez, and L. A. Marchat. Dna methylation of leptin and adiponectin promoters in children is reduced by the combined presence of obesity and insulin resistance. *Int J Obes (Lond)*, Feb 2014.

[48] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[49] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(1):721–741, 1984.

[50] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.

[51] A. D. Goldberg, C. D. Allis, and E. Bernstein. Epigenetics: A landscape takes shape. *Cell*, 128(4):635–638, Feb 2007.

[52] M. G. Goll and T. H. Bestor. Eukaryotic cytosine methyltransferases. *Annual Review of Biochemistry*, 74(1):481–514, Jun 2005.

[53] K. A. Gray, L. C. Daugherty, S. M. Gordon, R. L. Seal, M. W. Wright, and E. A. Bruford. Genenames.org: the hgnc resources in 2013. *Nucleic Acids Res*, 41(Database issue):D545–D552, Jan 2013.

[54] A. Green, N. Christian Hirsch, and S. Krøger Pramming. The changing world demography of type 2 diabetes. *Diabetes/Metabolism Research and Reviews*, 19(1):3–7, Jan 2003.

[55] J. L. Griffin. The cinderella story of metabolic profiling: does metabolomics get to go to the functional genomics ball? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1465):147–161, Jan 2006.

[56] S. Griffiths-Jones. Rfam: an rna family database. *Nucleic Acids Research*, 31(1):439–441, Jan 2003.

[57] S. Griffiths-Jones. The microrna registry. *Nucleic Acids Res*, 32(Database issue):D109–D111, Jan 2004.

[58] S. Grossmann, S. Bauer, P. N. Robinson, and M. Vingron. Improved detection of overrepresentation of gene-ontology annotations with parent child analysis. *Bioinformatics (Oxford, England)*, 23(22):3024–3031, Nov. 2007.

[59] A. Guilherme, J. V. Virbasius, V. Puri, and M. P. Czech. Adipocyte dysfunctions linking obesity to insulin resistance and type 2 diabetes. *Nat Rev Mol Cell Biol*, 9(5):367–377, May 2008.

[60] H. Guo, N. T. Ingolia, J. S. Weissman, and D. P. Bartel. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–840, Aug. 2010.

[61] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, J. Ascano, M., A. C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, and T. Tuschl. Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. *Cell*, 141(1):129–41, 2010.

[62] D. W. Haslam and W. P. T. James. Obesity. *The Lancet*, 366(9492):1197–1209, Oct 2005.

[63] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. Springer, 2011.

[64] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[65] P. J. Havel. Update on adipocyte hormones: Regulation of energy balance and carbohydrate/lipid metabolism. *Diabetes*, 53(Supplement 1):S143–S151, Feb 2004.

[66] L. He and G. J. Hannon. Micrornas: small rnas with a big role in gene regulation. *Nat Rev Genet*, 5(7):522–531, Jul 2004.

[67] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[68] S. Hohmann. Osmotic stress signaling and osmoadaptation in yeasts. *Microbiology and Molecular Biology Reviews*, 66(2):300–372, June 2002.

[69] R. Holliday and J. E. Pugh. Dna modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, Jan 1975.

[70] J.-H. Hong. Taz, a transcriptional modulator of mesenchymal stem cell differentiation. *Science*, 309(5737):1074–1078, Aug 2005.

[71] S.-D. Hsu, F.-M. Lin, W.-Y. Wu, C. Liang, W.-C. Huang, W.-L. Chan, W.-T. Tsai, G.-Z. Chen, C.-J. Lee, C.-M. Chiu, C.-H. Chien, M.-C. Wu, C.-Y. Huang, A.-P. Tsou, and H.-D. Huang. mirtarbase: a database curates experimentally validated microrna-target interactions. *Nucleic Acids Res*, 39(Database issue):D163–D169, Jan 2011.

[72] D. W. Huang, B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler, H. C. Lane, and et al. David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*, 35(Web Server):W169–W175, May 2007.

[73] J. C. Huang, T. Babak, T. W. Corson, G. Chua, S. Khan, B. L. Gallie, T. R. Hughes, B. J. Blencowe, B. J. Frey, and Q. D. Morris. Using expression profiling data to identify human microrna targets. *Nat. Methods*, 4(12):1045–1049, 2007.

[74] T. Hunkapiller, R. J. Kaiser, B. F. Koop, and L. Hood. Large-scale and automated dna sequence determination. *Science*, 254(5028):59–67, Oct 1991.

[75] T. Hunter. Signaling–2000 and beyond. *Cell*, 100(1):113–127, Jan 2000.

[76] M. Inui, G. Martello, and S. Piccolo. Microrna control of signal transduction. *Nat Rev Mol Cell Biol*, 11(4):252–263, Apr 2010.

[77] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of affymetrix genechip probe level data. *Nucleic Acids Res*, 31(4):e15, Feb 2003.

[78] S. P. Jackson. Regulating transcription factor activity by phosphorylation. *Trends Cell Biol*, 2(4):104–108, Apr 1992.

[79] R. Jaenisch and A. Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33(3s):245–254, Mar 2003.

[80] A. W. James. Review of signaling pathways governing msc osteogenic and adipogenic differentiation. *Scientifica (Cairo)*, 2013:684736, 2013.

[81] R. Ji, Y. Cheng, J. Yue, J. Yang, X. Liu, H. Chen, D. B. Dean, and C. Zhang. Microrna expression signature and antisense-mediated depletion reveal an essential role of microrna in vascular neointimal lesion formation. *Circ Res*, 100(11):1579–1588, Jun 2007.

[82] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, Jun 2007.

[83] P. A. Jones. The role of dna methylation in mammalian epigenetics. *Science*, 293(5532):1068–1070, Aug 2001.

[84] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, Nov. 2011.

[85] T. Katagiri and N. Takahashi. Regulatory mechanisms of osteoblast and osteoclast differentiation. *Oral Diseases*, 8(3):147–159, May 2002.

[86] T. Kawamata and Y. Tomari. Making risc. *Trends Biochem Sci*, 35(7):368–376, Jul 2010.

[87] T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, and A. R. Pico. Wikipathways: building research communities on biological pathways. *Nucleic Acids Research*, 40(D1):D1301–D1307, Jan 2012.

[88] J.-H. Kim, K. W. Park, E.-W. Lee, W.-S. Jang, J. Seo, S. Shin, K.-A. Hwang, and J. Song. Suppression of ppar through mkrn1-mediated ubiquitination and degradation prevents adipocyte differentiation. *Cell Death Differ*, 21(4):594–603, Apr 2014.

[89] A. Kowarsch, M. Preusse, C. Marr, and F. J. Theis. miTALOS: analyzing the tissue-specific regulation of signaling pathways by human and mouse microRNAs. *RNA*, 17(5):809–819, May 2011.

[90] A. Kozomara and S. Griffiths-Jones. mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic Acids Res*, 42(1):D68–D73, Jan 2014.

[91] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of novel genes coding for small expressed rnas. *Science*, 294(5543):853–858, Oct 2001.

[92] M. Lagos-Quintana, R. Rauhut, J. Meyer, A. Borkhardt, and T. Tuschl. New micrornas from mouse and human. *RNA*, 9(2):175–179, Feb 2003.

[93] M. Lagos-Quintana, R. Rauhut, A. Yalcin, J. Meyer, W. Lendeckel, and T. Tuschl. Identification of tissue-specific micrornas from mouse. *Current Biology*, 12(9):735–739, Apr 2002.

[94] P. W. Laird. Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11(3):191, Mar 2010.

[95] P. Landgraf, M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A. O. Kamphorst, M. Landthaler, C. Lin, N. D. Socci, L. Hermida, V. Fulci, S. Chiaretti, R. Fo, J. Schliwka, U. Fuchs, A. Novosel, R.-U. Mller, B. Schermer, U. Bissels, J. Inman, Q. Phan, M. Chien, D. B. Weir, R. Choksi, G. De Vita, D. Frezzetti, H.-I. Trompeter, V. Hornung, G. Teng, G. Hartmann, M. Palkovits, R. Di Lauro, P. Wernet, G. Macino, C. E. Rogler, J. W. Nagle, J. Ju, F. N. Papavasiliou, T. Benzing, P. Lichter, W. Tam, M. J. Brownstein, A. Bosio, A. Borkhardt, J. J. Russo, C. Sander, M. Zavolan, and T. Tuschl. A mammalian microrna expression atlas based on small rna library sequencing. *Cell*, 129(7):1401–1414, Jun 2007.

[96] N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel. An abundant class of tiny rnas with probable regulatory roles in caenorhabditis elegans. *Science*, 294(5543):858–862, Oct 2001.

[97] M. V. Lee, S. E. Topper, S. L. Hubler, J. Hose, C. D. Wenger, J. J. Coon, and A. P. Gasch. A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Molecular Systems Biology*, 7(1):1–12, July 2011.

[98] R. C. Lee, R. L. Feinbaum, and V. Ambros. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell*, 75(5):843–854, Dec 1993.

[99] B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are MicroRNA targets. *Cell*, 120(1):15–20, Jan. 2005.

[100] B. P. Lewis, I.-h. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787–798, Dec 2003.

[101] W. Li, L. Chen, W. Li, X. Qu, W. He, Y. He, C. Feng, X. Jia, Y. Zhou, J. Lv, and et al. Unraveling the characteristics of microrna regulation in the developmental and aging process of the human brain. *BMC Med Genomics*, 6(1):55, 2013.

[102] X. Li, Q. Cui, C. Kao, G. J. Wang, and G. Balian. Lovastatin inhibits adipogenic and stimulates osteogenic differentiation by suppressing ppargamma2 and increasing cbfa1/runx2 expression in bone marrow mesenchymal cell cultures. *Bone*, 33(4):652–659, Oct 2003.

[103] X. Li, R. Gill, N. G. Cooper, J. K. Yoo, and S. Datta. Modeling microrna-mrna interactions using pls regression in human colon cancer. *BMC Med Genomics*, 4:44, 2011.

[104] L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson. Microarray analysis shows that some micrornas downregulate large numbers of target mrnas. *Nature*, 433(7027):769–773, Feb 2005.

[105] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, and et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, Nov 2009.

[106] J. Liu, M. A. Carmell, F. V. Rivas, C. G. Marsden, J. M. Thomson, J.-J. Song, S. M. Hammond, L. Joshua-Tor, and G. J. Hannon. Argonaute2 is the catalytic engine of mammalian rnai. *Science*, 305(5689):1437–1441, Sep 2004.

[107] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–1680, Dec 1996.

[108] D. Long, R. Lee, P. Williams, C. Y. Chan, V. Ambros, and Y. Ding. Potent effect of target structure on microrna function. *Nat Struct Mol Biol*, 14(4):287–294, Apr 2007.

[109] M. Lu, Q. Zhang, M. Deng, J. Miao, Y. Guo, W. Gao, and Q. Cui. An analysis of human microrna and disease associations. *PLoS One*, 3(10):e3420, 2008.

[110] Y. Lu, R. Rosenfeld, I. Simon, G. J. Nau, and Z. Bar-Joseph. A probabilistic generative model for GO enrichment analysis. *Nucleic Acids Research*, 36(17):e109, Oct. 2008.

[111] Y. M. Lu, Y. Zhou, W. B. Qu, M. H. Deng, and C. G. Zhang. A lasso regression model for the construction of microrna-target regulatory networks. *Bioinformatics*, 27(17):2406–2413, 2011.

[112] E. Lund, S. Gttinger, A. Calado, J. E. Dahlberg, and U. Kutay. Nuclear export of microrna precursors. *Science*, 303(5654):95–98, Jan 2004.

[113] O. A. MacDougald and M. D. Lane. Transcriptional regulation of gene expression during adipocyte differentiation. *Annual Review of Biochemistry*, 64(1):345–373, Jun 1995.

[114] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Research*, 39(Database):D52–D57, Jan 2011.

[115] J. H. Malone and B. Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, 9(1):34, 2011.

[116] M. Malumbres. mirnas versus oncogenes: the power of social networking. *Mol Syst Biol*, 8:569, 2012.

[117] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, Mar 1947.

[118] M. B. Miller and Y.-W. Tang. Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical Microbiology Reviews*, 22(4):611–633, Oct 2009.

[119] T. Minka, J. Winn, J. Guiver, and D. Knowles. Infer.NET 2.5, 2012. Microsoft Research Cambridge.

[120] T. P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[121] R. Mukhopadhyay, W. Yu, J. Whitehead, J. Xu, M. Lezcano, S. Pack, C. Kanduri, M. Kanduri, V. Ginjala, A. Vostrov, W. Quitschke, I. Chernukhin, E. Klenova, V. Lobanenkov, and R. Ohlsson. The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide. *Genome Res*, 14(8):1594–1602, Aug 2004.

[122] A. Muniategui, R. Nogales-Cadenas, M. Vazquez, L. A. X, X. Agirre, A. Luttun, F. Prosper, A. Pascual-Montano, and A. Rubio. Quantification of mirna-mrna interactions. *PLoS One*, 7(2):e30766, 2012.

[123] A. Muniategui, J. Pey, F. J. Planes, and A. Rubio. Joint analysis of mirna and mrna expression data. *Briefings in Bioinformatics*, 14(3):263–278, May 2013.

[124] R. M. Neal. Probabilistic inference using markov chain monte carlo methods. 1993.

[125] R. M. Neal. Slice sampling. *Annals of statistics*, 31:705–741, 2003.

[126] A. Nishimura, T. Kotani, Y. Sasano, and H. Takagi. An antioxidative mechanism mediated by the yeast n-acetyltransferase mpr1: oxidative stress-induced arginine synthesis and its physiological role. *FEMS yeast research*, 10(6):687–698, Sept. 2010.

[127] K. Okamura, M. D. Phillips, D. M. Tyler, H. Duan, Y.-t. Chou, and E. C. Lai. The regulatory activity of microrna* species has substantial influence on microrna and 3' utr evolution. *Nat Struct Mol Biol*, 15(4):354–363, Apr 2008.

[128] M. D. Paraskevopoulou, G. Georgakilas, N. Kostoulas, I. S. Vlachos, T. Vergoulis, M. Reczko, C. Filippidis, T. Dalamagas, and A. G. Hatzige-

orgiou. Diana-microt web server v5.0: service integration into mirna functional analysis workflows. *Nucleic Acids Res*, 41(Web Server issue):W169–W173, Jul 2013.

[129] L. Parts, O. Stegle, J. Winn, and R. Durbin. Joint genetic analysis of gene expression data with inferred cellular phenotypes. *PLoS Genet*, 7(1):e1001276, 01 2011.

[130] M. M. Pastor, M. Proft, and A. Pascual-Ahuir. Mitochondrial function is an inducible determinant of osmotic stress adaptation in yeast. *Journal of Biological Chemistry*, 284(44):30307–30317, Oct. 2009.

[131] K. Pearson. Contributions to the mathematical theory of evolution. iii. regression, heredity, and panmixia. *Proceedings of the Royal Society of London*, 59(353-358):69–71, 1895.

[132] A. C. Pease, D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes, and S. P. Fodor. Light-generated oligonucleotide arrays for rapid dna sequence analysis. *Proc Natl Acad Sci U S A*, 91(11):5022–5026, May 1994.

[133] X. Peng, Y. Li, K.-A. Walters, E. R. Rosenzweig, S. L. Lederer, L. D. Aicher, S. Proll, and M. G. Katze. Computational identification of hepatitis c virus associated microrna-mrna regulatory modules in human livers. *BMC Genomics*, 10:373, 2009.

[134] S. M. Peterson, J. A. Thompson, M. L. Ufkin, P. Sathyanarayana, L. Liaw, and C. B. Congdon. Common features of microrna target prediction tools. *Front Genet*, 5:23, 2014.

[135] A. J. Pratt and I. J. MacRae. The rna-induced silencing complex: a versatile gene-silencing machine. *J Biol Chem*, 284(27):17897–17901, Jul 2009.

[136] R. J. Prill, D. Marbach, J. Saez-Rodriguez, P. K. Sorger, L. G. Alexopoulos, X. Xue, N. D. Clarke, G. Altan-Bonnet, and G. Stolovitzky. Towards a rigorous assessment of systems biology models: the dream3 challenges. *PLoS One*, 5(2):e9202, 2010.

[137] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[138] B. H. Ramsahoye, D. Biniszkiewicz, F. Lyko, V. Clark, A. P. Bird, and R. Jaenisch. Non-cpg methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences*, 97(10):5237–5242, May 2000.

[139] T. Ravasi, H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin, P. Carninci, C. O. Daub, A. R. R. Forrest, J. Gough, S. Grimmond, J.-H. Han, T. Hashimoto, W. Hide, O. Hofmann, A. Kamburov, M. Kaur, H. Kawaji, A. Kubosaki, T. Lassmann, E. van Nimwegen, C. R. MacPherson, C. Ogawa, A. Radovanovic, A. Schwartz, R. D. Teasdale, J. Tegnr, B. Lenhard, S. A. Teichmann, T. Arakawa, N. Ninomiya, K. Murakami, M. Tagami, S. Fukuda, K. Imamura, C. Kai, R. Ishihara, Y. Kitazume, J. Kawai, D. A. Hume, T. Ideker, and Y. Hayashizaki. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752, Mar 2010.

[140] S. U. Raymond, S. Leeder, and H. M. Greenberg. Obesity and cardiovascular disease in developing countries: a growing problem and an economic threat. *Current Opinion in Clinical Nutrition and Metabolic Care*, 9(2):111–116, Mar 2006.

[141] M. Rep, M. Proft, F. Remize, M. Tams, R. Serrano, J. M. Thevelein, and S. Hohmann. The saccharomyces cerevisiae sko1p transcription factor mediates HOG pathway-dependent osmotic regulation of a set of genes encoding enzymes implicated in protection from oxidative damage. *Molecular Microbiology*, 40(5):10671083, 2001.

[142] P. H. Reyes-Herrera and E. Ficarra. One decade of development and evolution of microrna target prediction algorithms. *Genomics Proteomics Bioinformatics*, 10(5):254–263, Oct 2012.

[143] A. Rinck, M. Preusse, B. Laggerbauer, H. Lickert, S. Engelhardt, and F. J. Theis. The human transcriptome is enriched for mirna-binding sites

located in cooperativity-permitting distance. *RNA Biol*, 10(7):1125–1135, Jul 2013.

[144] M. Rodstein. Crime and the aged. 2. the criminals. *JAMA*, 234(6):639, Nov 1975.

[145] E. D. Rosen and O. A. MacDougald. Adipocyte differentiation from the inside out. *Nat Rev Mol Cell Biol*, 7(12):885–896, Dec 2006.

[146] E. D. Rosen, C. J. Walkey, P. Puigserver, and B. M. Spiegelman. Transcriptional regulation of adipogenesis. *Genes Dev*, 14(11):1293–1307, Jun 2000.

[147] S. E. Ross, R. L. Erickson, I. Gerin, P. M. DeRose, L. Bajnok, K. A. Longo, D. E. Misek, R. Kuick, S. M. Hanash, K. B. Atkins, S. M. Andresen, H. I. Nebb, L. Madsen, K. Kristiansen, and O. A. MacDougald. Microarray analyses during adipogenesis: understanding the effects of wnt signaling on adipogenesis and the roles of liver x receptor alpha in adipocyte metabolism. *Mol Cell Biol*, 22(16):5989–5999, Aug 2002.

[148] S. E. Ross, N. Hemati, K. A. Longo, C. N. Bennett, P. C. Lucas, R. L. Erickson, and O. A. MacDougald. Inhibition of adipogenesis by wnt signaling. *Science*, 289(5481):950–953, Aug 2000.

[149] H. Rottenberg, S. R. Caplan, and A. Essig, J. M. Stoichiometry and coupling: Theories of oxidative phosphorylation. *Nature*, 216(5115):610–611, Nov 1967.

[150] L. Rowen. Sequencing the human genome. *Science*, 278(5338):605–607, Oct 1997.

[151] A. Ruepp, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, M. Stransky, B. Waegele, T. Schmidt, O. N. Doudieu, V. Stumpflen, and et al. Corum: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Research*, 36(Database):D646–D650, Dec 2007.

[152] A. Ruepp, A. Kowarsch, D. Schmidl, F. Bruggenthin, B. Brauner, I. Dunger, G. Fobo, G. Frishman, C. Montrone, and F. J. Theis. Phe-

nomir: a knowledgebase for microrna expression in diseases and biological processes. *Genome Biology*, 11(1):R6, 2010.

[153] G. Sales, A. Coppe, A. Bisognin, M. Biasiolo, S. Bortoluzzi, and C. Romualdi. Magia, a web-based tool for mirna and genes integrated analysis. *Nucleic Acids Res*, 38(Web Server issue):W352–W359, Jul 2010.

[154] S. Sass, F. Buettner, N. S. Mueller, and F. J. Theis. A modular framework for gene set analysis integrating multilevel omics data. *Nucleic Acids Res*, 41(21):9622–9633, Nov 2013.

[155] S. Sass, S. Dietmann, U. C. Burk, S. Brabletz, D. Lutter, A. Kowarsch, K. F. Mayer, T. Brabletz, A. Ruepp, F. J. Theis, and Y. Wang. MicroRNAs coordinately regulate protein complexes. *BMC Systems Biology*, 5(1):136, Aug. 2011.

[156] M. Scherf, A. Klingenhoff, and T. Werner. Highly specific localization of promoter regions in large genomic sequences by promoterinspector: a novel context analysis approach. *Journal of Molecular Biology*, 297(3):599–606, Mar 2000.

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

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

[159] K. L. Sheaffer, R. Kim, R. Aoki, E. N. Elliott, J. Schug, L. Burger, D. Schubeler, and K. H. Kaestner. DNA methylation is required for the control of stem cell differentiation in the small intestine. *Genes & Development*, 28(6):652–664, Mar 2014.

[160] A. Simpson, V. Y. Tan, J. Winn, M. Svensn, C. M. Bishop, D. E. Heckerman, I. Buchan, and A. Custovic. Beyond atopy: multiple patterns of

sensitization in relation to asthma in a birth cohort study. *Am J Respir Crit Care Med*, 181(11):1200–1206, Jun 2010.

[161] P. J. Smith, L. S. Wise, R. Berkowitz, C. Wan, and C. S. Rubin. Insulin-like growth factor-i is an essential regulator of the differentiation of 3t3-l1 adipocytes. *J Biol Chem*, 263(19):9402–9408, Jul 1988.

[162] G. K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

[163] S. E. Spruill, J. Lu, S. Hardy, and B. Weir. Assessing sources of variability in microarray gene expression data. *Biotechniques*, 33(4):916–20, 922–3, Oct 2002.

[164] F. J. T. Staal, M. van der Burg, L. F. A. Wessels, B. H. Barendregt, M. R. M. Baert, C. M. M. van den Burg, C. Van Huffel, A. W. Langerak, V. H. J. van der Velden, M. J. T. Reinders, and et al. DNA microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-b acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers. *Leukemia*, 17(7):1324–1332, Jul 2003.

[165] J. M. Stephens. The fat controller: adipocyte development. *PLoS Biol*, 10(11):e1001436, 2012.

[166] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, Oct. 2005.

[167] D. Takai and P. A. Jones. Comprehensive analysis of cpg islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences*, 99(6):3740–3745, Mar 2002.

[168] The UniProt Consortium. Activities at the universal protein resource (uniprot). *Nucleic Acids Res*, 42(Database issue):D191–D198, Jan 2014.

[169] P. D. Thomas, H. Mi, G. E. Swan, C. Lerman, N. Benowitz, R. F. Tyndale, A. W. Bergen, D. V. Conti, P. o. N. A. , and T. Consortium. A systems biology network model for genetic association studies of nicotine addiction and treatment. *Pharmacogenet Genomics*, 19(7):538–551, Jul 2009.

[170] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[171] S. Vasudevan and J. A. Steitz. Au-rich-element-mediated upregulation of translation by fxr1 and argonaute 2. *Cell*, 128(6):1105–1118, Mar 2007.

[172] T. Vergoulis, I. S. Vlachos, P. Alexiou, G. Georgakilas, M. Maragkakis, M. Reczko, S. Gerangelos, N. Koziris, T. Dalamagas, and A. G. Hatzigeorgiou. Tarbase 6.0: capturing the exponential growth of mirna targets with experimental support. *Nucleic Acids Res*, 40(Database issue):D222–D229, Jan 2012.

[173] J. Wang, X. Guan, F. Guo, J. Zhou, A. Chang, B. Sun, Y. Cai, Z. Ma, C. Dai, X. Li, and et al. mir-30e reciprocally regulates the differentiation of adipocytes and osteoblasts by directly targeting low-density lipoprotein receptor-related protein 6. *Cell Death and Disease*, 4(10):e845, Oct 2013.

[174] R. Y.-H. Wang, C. W. Gehrke, and M. Ehrlich. Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. *Nucleic Acids Research*, 8(20):4777–4790, 1980.

[175] Y. P. Wang and K. B. Li. Correlation of expression profiles between microrna s and mrna targets using nci-60 data. *BMC Genomics*, 10, 2009.

[176] M. Weber, J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, and D. Schbeler. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics*, 37(8):853–862, Aug 2005.

[177] H. V. Westerhoff and B. O. Palsson. The evolution of molecular biology into systems biology. *Nature Biotechnology*, 22(10):1249–1252, Oct 2004.

[178] E. Wienholds and R. H. Plasterk. Microrna function in animal development. *FEBS Letters*, 579(26):5911–5922, Oct 2005.

[179] T. Wu, H. Zhou, Y. Hong, J. Li, X. Jiang, and H. Huang. mir-30 family members negatively regulate osteoblast differentiation. *Journal of Biological Chemistry*, 287(10):7503–7511, Mar 2012.

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

[181] H. Xie, B. Lim, and H. F. Lodish. Microrna induced during adipogenesis that accelerate fat cell development are downregulated in obesity. *Diabetes*, 58(5):1050–1057, May 2009.

[182] J.-H. Yang, J.-H. Li, P. Shao, H. Zhou, Y.-Q. Chen, and L.-H. Qu. starbase: a database for exploring microrna-mrna interaction maps from argonaute clip-seq and degradome-seq data. *Nucleic Acids Res*, 39(Database issue):D202–D209, Jan 2011.

[183] J.-S. Yang, M. D. Phillips, D. Betel, P. Mu, A. Ventura, A. C. Siepel, K. C. Chen, and E. C. Lai. Widespread regulatory activity of vertebrate microrna* species. *RNA*, 17(2):312–326, Feb 2011.

[184] R. A. Yates and A. B. Pardee. Control by uracil of formation of enzymes required for orotate synthesis. *J Biol Chem*, 227(2):677–692, Aug 1957.

[185] D. Yue, H. Liu, and Y. Huang. Survey of computational algorithms for microrna target prediction. *Curr Genomics*, 10(7):478–492, Nov 2009.

[186] L.-E. Zaragosi, B. Wdziekonski, K. Brigand, P. Villageois, B. Mari, R. Waldmann, C. Dani, and P. Barbry. Small rna sequencing reveals mir-642a-3p as a novel adipocyte-specific microrna and mir-30 as a key regulator of human adipogenesis. *Genome Biology*, 12(7):R64, 2011.

[187] B. Zhang, S. Kirov, and J. Snoddy. Webgestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res*, 33(Web Server issue):W741–W748, Jul 2005.

[188] R. Zhang, D. Wang, Z. Xia, C. Chen, P. Cheng, H. Xie, and X. Luo. The role of micrornas in adipocyte differentiation. *Front. Med.*, 7(2):223–230, Jun 2013.

[189] Q.-h. Zhao, S.-g. Wang, S.-x. Liu, J.-p. Li, Y.-x. Zhang, Z.-y. Sun, Q.-m. Fan, and J.-w. Tian. PPARγ forms a bridge between DNA methylation and histone acetylation at the C/EBPα gene promoter to regulate the balance between osteogenesis and adipogenesis of bone marrow stromal cells. *FEBS J*, 280(22):5801–5814, Nov 2013.

[190] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J Roy Stat Soc B*, 67:301–320, 2005.