Saarland University
Center for Bioinformatics
PhD Program in Bioinformatics

Integrated Analysis and Application Pipelines for Complex Disease Data

Dissertation
zur Erlangung des Grades
Doktor der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät III
Chemie, Pharmazie, Bio- und Werkstoffwissenschaften
der Universität des Saarlandes

von

Christian Spaniol

June 2015

Tag des Kolloquiums:

3. September 2015

Dekan:

Prof. Dr.-Ing Dirk Bähre

Prüfungsausschuss:

Prof. Dr. Uli Müller (Vorsitzender)
Prof. Dr. Volkhard Helms (1. Berichterstatter)
Prof. Dr. med. Matthias Riemenschneider (2. Berichterstatter)
Dr. Jessica Hoppstädter (Akad. Mitarbeiterin)

Abstract

The increasing amount of biological data available from high-throughput technologies poses great interdisciplinary challenges to research. Today, cost-efficient platforms generate manifold types of data and allow to build comprehensive resources that include but are not limited to genomics, proteomics, and metabolomics on a systemic scale. In order to adapt to this development in the post-wetlab analysis, computer scientists in computational biology work on methods and software frameworks that are able to account for data size and diversity, and allow to scrutinize data in respect to a specific context, such as the emergence of diseases.

Aiming for this, we first present a desktop software framework designed to integrate biological data that features a uniform interface to perform consecutive analysis steps managed by an automated task processing system. The extensibility of the platform based on a concise plugin interface was used for a study on breast cancer for which we developed a plugin to derive gene regulatory networks.

From this analysis, we derived a general approach to generate transcription factor-microRNA regulatory networks and built a webservice available for public use: TFmiR. Using differentially expressed sets of mRNAs and miRNAs, TFmiR generates a network with experimental or predicted evidence and provides downstream investigation, e.g. applying various network measures and overrepresentation analysis. Further in-depth analysis is provided with a motif search algorithm. For all motifs of particular interest, the software allows to investigate co-regulated and co-targeted subnetworks and calculates the functional similarity scores of the participating genes.

We investigated a comprehensive dataset on Alzheimer's disease that was provided by the neurological laboratory in Homburg. We conducted the individual analysis of the various types of data, followed by applying our approaches to build regulatory networks, and search for potential key drivers of the Alzheimer's disease. Moreover, we show a different strategy based on patient-similarity networks with the aim to find a descriptive combination of markers for AD spanning the multiple data sources.

Zusammenfassung

Biotechnologische Hochdurchsatzverfahren und die damit verbundene stetig anwachsende Menge an biologischen Daten stellen die Forschung vor ebenso wachsende Herausforderungen. Neue und kosteneffiziente Verfahren erlauben die Erstellung umfangreicher Datenbanken, die beispielsweise das vollständige Genom, Proteom, oder Metabolom eines Organismus oder Individuums enthalten können. Informatiker, Bioinformatiker, und Biologen arbeiten daher an Methoden und Softwareumgebungen um dieser Entwicklung nachzukommen und diese Daten trotz ihres Umfangs und Vielfältigkeit einheitlich erfassen zu können. Dabei gilt besonderes Interesse der Notwendigkeit, diese Daten im Hinblick auf ihre Bedeutung in bestimmten Kontexten zu untersuchen, wie zum Beispiel im Zusammenhang mit Krankheiten.

Mit diesem Ziel vor Augen zeigen wir zunächst die Softwareumgebung Mebitoo, die wir zur Integration und automatisierten Analyse von biologischen Daten entwickelten. Mit einer Erweiterung der Software zur Erstellung regulatorischer Netzwerke zeigen wir die vielfältige Einsetzbarkeit der Platform am Beispiel von Daten zu Brustkarzinomen.

Aufbauend darauf entwickelten wir eine allgemeine Strategie zur Erstellung regulatorischer Netzwerke, die auf differentiell exprimierten Genen und microRNAs basiert. Wir stellten einen Webservice zur Verfügung, der durch die Einbindung verschiedener Datenbanken zu experimentell bestimmten oder *in silico* berechneten mutmaßlichen Interaktionen ein regulatorisches Netzwerk, wahlweise im Hinblick auf eine mögliche Krankheit, erstellt und untersucht. Die bereitgestellten Analysen umfassen Methoden zur generellen Netzwerkevaluierung, sowie aufwändigere Algorithmen zur Bestimmung von Netzwerkmotiven und deren Subnetzen, und die Untersuchung auf deren Funktionalität.

Abschließend beschreiben wir die Untersuchung eines umfassenden Datensatzes zur Alzheimer'schen Krankheit, welcher vom neurologischen Labor der Universitätsklinik des Saarlandes zusammengestellt wurde. Die Daten umfassen Gen- und miRNA Expressionsprofile, Methylierung, Proteinlevelmessungen, und SNPs zu einer Kohorte von Alzheimerpatienten und Kontrollen. Wir untersuchten die Daten jeweils individuell und zeigten anschließend die Anwendung unserer Pipeline Identifikation von mutmaßlichen *Key Drivern.* Darüberhinaus verfolgten wir einen Ansatz, der auf Ähnlichkeitsnetzen für die jeweiligen Patienten beruht.

Acknowledgments

I like to thank all members of Professor Volkhard Helms' department for computational biology for offering me an excellent and always pleasant environment to work in.

Especially, I want to mention Jennifer Degač and Thorsten Will for being valuable partners in discussions about practical applications, theoretical problems, and utter nonsense alike.

I am thankful to Sabrina Pichler and Wei Gu who were a great support for my work on Alzheimer's disease.

Most notably, I thank my colleague and friend Mohamed Hamed Al Fahmy. He is an inspiring person to work with and his dedication was a "key driver" for many valuable projects.

Additionally, I am grateful to Professor Dr. Matthias Riemenschneider for giving me the ongoing opportunity to study the Alzheimer's Disease at his department.

I owe gratitude to Professor Dr. Volkhard Helms for offering me a position at his group and for his ongoing support, help and encouragement at all times.

Last, I thank my friends and family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the last century, biological technology evolved to create a critical amount of data that exceeded the capability to comprehend manually in magnitudes. Thus, a need for structured data storage established and drove the development of biological databases, starting with Margret Dayhoff and the *Atlas of Protein Sequence and Structure*. This lead to the Protein Information Resource database of protein sequences - today part of the Uniprot consortium [Dayhoff et al., 1976; Dayhoff, 1965]. Meanwhile, her group started the computational analysis of protein structures [Dayhoff, 1974, 1969]. With Dayhoff's field of research prevailing and gaining importance up to today, we regard this as the beginning of computational biology, and bioinformatics. Ever since, the rapidly increasing fund of biological data spans over a respectable amount of databases that hold DNA/RNA sequencing data, methylomes, genotypes, and protein structures.

When evaluating the human genome, more than 1 800 disease associated genes were discovered, enabling the development of a variety of genetic tests for certain conditions [Miklos and Rubin, 1996]. However, the knowledge of the genome helps to assess genetic risks but delivers not necessarily deterministic answers for the occurrence of many diseases. Epigenetic regulation mechanisms known as DNA methylation, histone modification, chromatin remodeling and noncoding RNAs, influence *which* genes actually are expressed and play a major role in cell differentiation, growth, and eventually in the pathological history of each individual [Bird, 1986]. For this reason, studies of expression data and epigenetic features together delivered valuable insights.

In order to enhance a comprehensive understanding, databases were designed to embrace a variety of multifaceted biological data with respect to a specific purpose, such as The Cancer Genome Atlas (TCGA), which contains gene expression profiling, copy number variation profiling, SNP genotyping, genome wide DNA methylation profiling, microRNA profiling, and exon sequencing.

The availability of such comprehensive datasets rose the question of how to contextualize each other, and lead to the efforts to employ integrated analysis on biological data.

1.1 Biological Data

The translation of biological features on molecular level to generate computationally assessable data poses various challenges itself. For example, there is no way to observe the DNA sequence of a human genome in its entirety of over 3 000 Mega-base pairs using a microscope. Thus research focused on methods to accomplish such tasks, starting with the first practical method presented by Sanger et al. [1977] who amplified DNA using a DNA polymerase with specific terminators for each of the four nucleotides, and enabled to determine the base pair sequence using gel electrophoresis. As one of many examples, DNA sequencing methods enabled structured assessment of biological data and today, various experimental protocols exist to obtain methylomes, expression profiles, DNA/RNA sequences, protein structures and many more.

In the following sections, we outline the methods applicable to obtain the data we studied in the scope of this work.

1.1.1 Enzyme Linked Immunosorbent Assay (ELISA)

The Enzyme Linked Immunosorbent Assay (ELISA) is one of the most established methods to detect the presence of a certain substance in a sample, and its variations share the same concept.

In principle, a liquid sample subjected to test is added to a stationary solid phase with a ligand-specific binding reagent that contains the antigen for a certain antibody or vice versa. Subsequently, a labelled substrate for the antigen is added to the plate and the reaction of non-bound antigens with the substrate induces a color change. The resulting signal is measured using spectrophotometry and allows quantification of the antigen in the studied sample [Voller et al., 1978].

ELISA has been established as a standard for diagnostic tests such as the determination of serum antibody concentrations in HIV patients. In Chapter 5, we study Amyloid-β 40 and 42 measurements obtained with ELISA.

1.1.2 Microarrays

As mentioned before, regulatory mechanisms influence the transcription of the genome. In order to gain an insight into this transcriptional level, genome-wide hybridization arrays were designed and are used today to compare genome-wide features among individuals and tissues. For example, an investigation of samples from disease patients in comparison with healthy controls may show differences in the expression profiles that hint information on the gene products responsible for the defect. A schematic of a gene expression microarray assay is shown in Figure 1.1. Extracted mRNAs or amplified cDNAs from tissues are labelled with fluorescent markers and hybridized to the array. A laser and a confocal scanner then excite and detect the fluorescent dyes, which yields a digital image from the microarray. Using image processing algorithms, the spot intensities are measured and translated into numerical readings. After estimation and subtraction of the background noise, the final signal is an integer proportional to the concentration of the target sequence for each spot. For two-dye experiments, a ratio of the expression levels of a sample in respect to the reference is determined [Trevino et al., 2007]. A microarray is capable to detect and measure expression levels of thousands of genes in a single experiment.

The microarray technology developed by Illumina in particular is based on oligonucleotides attached to beads that are randomly deposited onto a glass surface. Using the address sequence of each bead, bead positions are decoded to determine which bead combination is located in which well [Gunderson et al., 2004]. Thus, each array has a unique layout file that is used to decode the data when scanning. For gene expression in human samples, Illumina provides an expression BeadChip (HT12v4) that targets more than 47 000 probes with up to 12 different samples per chip.

The design of the beads and therefore the applications of the Illumina BeadChip technology are multifaceted and include arrays used for genotyping, copy-number determination, sequencing, and methylation analysis. The versatility of the platform enables to obtain variety of data cost-efficiently, and was used to generate most of the data studied in Chapter 5.

1.1.3 DNA Methylation

DNA Methylation, occuring at the CpG dinucleotide, is probably the most studied epigenetic modification so far. Extensive mapping experiments in different cancers pointed out a key role of DNA methylation in oncogenic development [Boerno et al., 2010]. When studies showed the potential of methylation-based biomarkers that enhance early diagnosis, prognosis, and classification of cancer, the aim was set to perform epigenome-wide association analysis at reasonable costs.

The Illumina Infinium HumanMethylation 450k BeadChip covers 99% percent of all RefSeq genes with more than 480 000 CpG sites at high coverage with an average of 17 probes per gene. The array includes various functional elements, gene bodies and miRNA promoters are covered as well [Touleimat and Tost, 2012], and generates data for an extensive analysis with a single experiment. As we show in

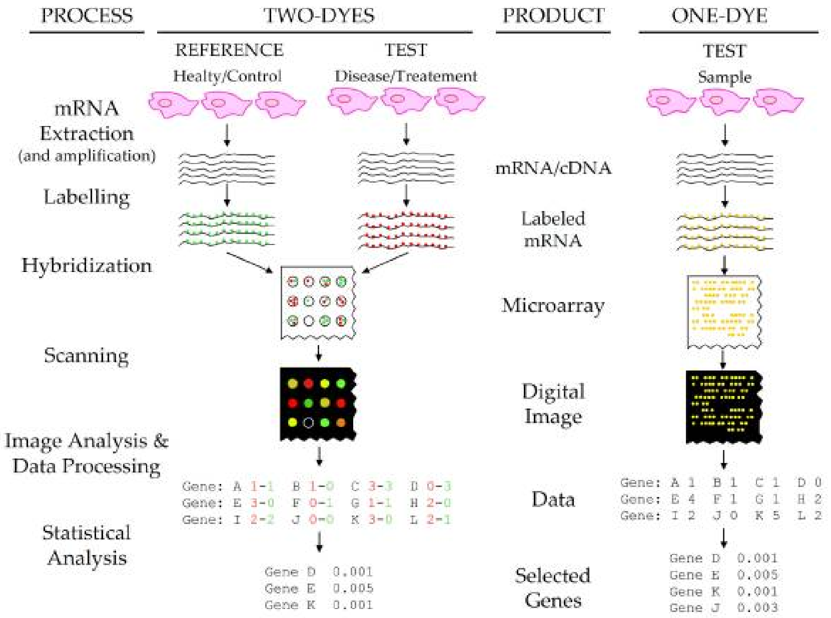


Figure 1.1: Overview of a microarray experiment for both a comparison experiment with two dyes and a single dye experiment. Left column shows the processing steps, third column the results of each step. Source: Trevino et al. [2007]

Chapter 5, we used this data to investigate methylation levels of gene and miRNA promoters together with expression and SNP data.

1.1.4 MicroRNA

Previously, we mentioned DNA methylation assuming a key role in the regulation of gene expression. Other than that, gene regulation may occur at the transcript level. Small non-coding RNAs, microRNAs (miRNA), bind to mRNA transcripts and prevent their translation to protein products. Moreover, targets may reciprocally influence level and function of miRNAs [Pasquinelli, 2012].

The mutual regulation of miRNAs and target genes is crucial to the understanding of gene-regulatory mechanisms.

The samples in our Alzheimer study were obtained using the Geniom RT Analyzer, which is another microarray-based platform. In order to quantify miRNA from the tissue samples, microRNAs are hybridized to microfluidic primers, labelled. After primer extension, a picture is obtained that is processed subsequently to translate the intensities into a numeric reading.

1.2 Single nucleotide polymorphisms (SNPs)

An investigation of single nucleotide polymorphisms (SNPs) enables to assess genetic variations between individuals on a genomic scale. For one position out of every 1 000 nucleotides, the human genome shows a base pair exchange among individuals [Syvänen, 2001]. Depending on the location of those SNPs, this may affect an individual in different ways. For example, a SNP that occurs within a gene coding region can change the amino acid composition of the encoded protein and, thus, alternate the structure and function of the product. In fact, many inherited disorders are associated with SNPs, such as the Apoϵ allele, which was shown to affect atherosclerosis [Davignon et al., 1988] and plays a role in our studies on Alzheimer's disease as well.

Apart from the obvious approach to sequence a sample genome, SNPs can be determined using a polymerase chain reaction with allele-specific oligonucleotides or, on larger scales, using microarrays. In our case, the Illumina Human610-Quad beadchip microarray was used to determine the SNPs for the samples in the Alzheimer study.

1.3 Systems Biology

Different biological data represent a collection of systematic measurements. Many-faceted system biology approaches that incorporate such genome-scale experiments have been developed to perform predictive, hypothesis-driven science [Chuang et al., 2010].

1.3.1 Gene regulatory networks

One systems biology strategy in particular aims at the reconstruction of gene regulatory networks (GRNs) from experimental data such as microarray gene expression profiles. The incorporation of more information such as interaction data, genome sequences, or epigenetic information helps to prune and thus to create more concise networks [Hecker et al., 2009].

The construction of regulatory networks and their pruning using other systemic data motivated large parts of the work presented in this thesis. We developed a gene regulatory network plugin based on co-expression data for our software framework Mebitoo presented in Chapter 3.3 and published subsequently the breast cancer study described in 3.4. In Chapter 4 we introduce the web service TFmiR, that was designed to build a transcription factor-miRNA regulatory network based on differentially expressed genes and miRNAs.

1.4 Similarity Networks

A different approach on data integration was done with networks of individuals [Barabási, 2007], for example in Christakis and Fowler [2007] where the authors investigated obesity of individuals and their social network and concluded obesity to spread through social ties.

Likewise, Wang et al. [2014] presented a method to create patient-similarity networks integrating biological data, and merge those into a single network that incorporates the information content of each data source. Thus, this network reflects a condensed representation of the dataset and offers insight into possible complementary characteristics of the different input sources.

Similarity networks and network merging are explained in more detail in section 2.3, and we outline their novel application on Alzheimer's disease data in Chapter 5.

1.5 Computational Tools for Data Integration

Unsurprisingly, the need for software that enables to create workflows for efficient processing rose together with the amount of biological data available.

Cytoscape, a software environment to integrate biological interaction networks with expression data, was presented by Shannon et al. [2003a] and gained large popularity due to a platform-independent architecture with a graphical user interface. Since then, the plugin-based platform received with well over 5000 citations and grew a large developer base that contributed various plugins for data visualization, ontology analysis, data integration, clustering, and many more as described in Saito et al. [2012]. However, some of the plugins mentioned are obsolete today, since the platform is under continuous development and was recently rebuild to be future-proof in a major version update to 3.x which is incompatible to 2.x.

Scripting languages on the other hand offer more flexibility to work with biological data. Dialects like Python allow dynamic typecasting and are based on an easy-to-understand syntax in comparison to their regular programming language counterparts and allow for very quick prototyping of data analysis pipelines. The potential was recognized, and groups like Cock et al. [2009] developed libraries that wrapped standard tasks like data import and export for biological data and the execution of BLAST queries or ClustalW alignments on sequences into their package called BioPython.

Statistics in computational biology provide important methods to rule out significant pieces from the large puzzle of biological data. Designed specifically for the purpose of statistical computing, the R framework naturally provides much of the required functionality. Because R features a packing protocol to extend the framework, Gentleman et al. [2004] presented their Bioconductor package to close the gap between R and biological data, starting with array-based expression data. Since then, Bioconductor has been extended continuously to allow processing of data from many different biotechnology platforms.

We present our own approach to a software framework for biological data integration and workflow creation in Chapter 3. Both Cytoscape and R were used to develop the pipeline behind TFmiR (Chapter 4), and large parts of the analysis in the Alzheimer's disease study was carried out using R scripts based on Bioconductor, and various subpackages designed to handle the variation of data sources.

1.6 Complex Diseases

Research has shown that many diseases show a genetic component [Davison et al., 1994]. Some disorders like sickle cell disease, cystic fibrosis, or Huntington's disease are linked to mutations in single genes or loci. However, many other disorders are likely to arise due to a combination of genetic factors but are induced by certain lifestyles and environmental factors as well, many of which are yet to be determined. We now know that there are genetic predispositions for certain diseases (such as the Apolipoprotein E (Apoϵ) allele in Alzheimer's Disease), but a genetic tendency alone proved not be sufficient as definitive predictors for many of them [Craig, 2008].

In the next sections, we describe briefly the complex diseases studied in the scope of this thesis, in particular breast carcinoma and the Alzheimer's disease.

1.6.1 Breast Cancer

Breast cancer (BC) is the most prevalent carcinoma in females, with one of ten women affected by the age of 80 years and accounts for the second-highest number of deaths of female cancer patients, after lung cancer [Siegel et al., 2014]. Because BC is a genetically heterogenous type of cancer, treatment and prognosis depends on correct classification of the carcinoma at hand [Volinia and Croce, 2013]. Due its complexity, molecular mechanisms and regulatory patterns of the disease are not yet completely understood.

In order to address the complexity with appropriate models, Cava et al. [2014], for example, presented an effective discrimination of cancer types based on a support vector machine classifier combining copy number variations, SNP data, and the expression values of miRNAs, and mRNAs.

In section 3.4, we describe our approaches to study BC with regulatory network approaches where we ruled out possible key driver genes and potential drug targets.

Additionally, we present the application of our TFmiR service to build a TF-regulatory network on breast cancer data in section 4.5.

1.6.2 Alzheimer's Disease

With improved healthcare and life standards in general, the average human life expectancy increased largely, and thus aging and aging-related disease research is regarded increasingly important. Besides cancer that represents in fact a collection of diseases, neurodegenerative diseases pose the most prevalent risk for the elderly. For the most common disorder, the Alzheimer's Disease (AD), cases double every five years from the age of 65 onwards. Since its discovery in the early 20th century, Alzheimer has been known as a complex disorder that is hard to diagnose in early stages, researchers started to search for possible connections and interactions between different regulatory pathways that point to the mechanism behind the disease.

Medical indications discern between Early-Onset Alzheimers Disease (EOAD) and Late-Onset Alzheimer's Disease (LOAD). The early-onset form occurs prior to the age of 60-65 years and often even before the age of 55 and is known to be mainly caused by mutations in three genes and inherited in an autosomal-dominant fashion. This familial form is implied by mutations in the genes related to encoding the amyloid precursor protein (APP), as well as mutations both in presenilin-1 and presinilin-2 (PSEN1, PSEN2). In a normal metabolism, APP is processed by the β-secretase 1 (BACE1) and the γ-secretase and is transformed to β-amyloid 40 and 42 (Aβ40, Aβ42) which are then decomposed. However, both peptides show neurotoxic characteristics and plaque accumulations have been found in brain tissues of patients diagnosed with AD as well as with the Down Syndrome. Interestingly, those plaques have been found in patients that suffered from traumatic brain injury [Johnson et al., 2010], which indicates its connection to neurodegeneration not to be limited to AD.

On the other hand, for the LOAD forms that show prevalence in individuals of 60-65 years and above, there are several genes known to be involved in the development of sporadic AD. So far, the gene coding for apoliprotein ϵ (Apoϵ) and its genotypes are considered a major factor but not a sufficient marker for diagnostics or prognosis. Statistically, AD patients are more likely to carry the Apoϵ4 allele than the population in general, while Apoϵ2 may be protective [Minati et al., 2009].

Alzheimer's disease largely affects the episodic and semantic memory as well as it induces noncognitive behavioural changes [Mega et al., 1996]. The temporal lobe - one of the four major brain lobes of the cerebral cortex - has been shown to be largely associated with those traits and thus, with AD [Visser et al., 2002].

In the closing chapter 5, we investigate a comprehensive dataset spanning microRNA and gene expression as well as methylome and Amyloid-β levels obtained for 64 samples of temporal lobe tissue from post-mortem patients, of which 39 suffered from Late-Onset Alzheimer's Disease.

1.7 Outline

This thesis is divided into six chapters. Closing the introduction at this point, the author presents the theoretical part relevant for this thesis in chapter 2, where statistical methods and network theory are explained in more detail.

The subsequent three chapters outline major projects the author participated in.

In chapter 3, the Mebitoo software framework for data integration and workflow pipelines and the application of the GRN plugin on breast cancer data and their downstream evaluation are presented.

Subject of chapter 4 is TFmiR, a web service we developed to build regulatory networks based on gene and miRNA expression data. Moreover, we show our studies with TFmiR in respect to breast cancer.

The last project presented in the scope of this thesis is the still ongoing study on Alzheimer's disease, for which we carried out the analysis for a comprehensive dataset ourselves. Subsequently, we applied the former approaches and additionally pursued a different strategy based on patient-similarity networks (Chapter 5).

Finally, the last chapter 6 concludes this thesis with a summary and outlook on future work.

Chapter 2

Theory

2.1 Statistical Tests

Statistical hypothesis testing is a common method to draw inference about populations using deviations within data from expected values or, in two-sample testing, to compare different groups of samples of a population. For instance, in the scope of this thesis, these are individual markers - such as expression levels for genes and microRNAs or methylated regions, that are tested for significant differences between case and control groups of different diseases.

For each test, one tries to determine a probability for a the dataset given a certain hypothesis is true.

2.1.1 Fundamentals

Classical statistical tests share the same concept:

1. A null-hypothesis H_0 and an alternative hypothesis H_1 are formulated. Basically, it is hypothesized that a population is different or not different from a certain mean, while the alternative is the antithesis.

2. When calculating a test-statistic, it is determined how probable a property is for the samples in question. As it is in general unlikely to match the exact mean in a statistical test, results are judged by confidence levels. A significance level for the statistic reduces the decision whether or not the null hypothesis is rejected on the so called p-value.

3. If the test statistic fails to satisfy the significance level, it is rejected as being unlikely to hold given the samples.

For the scope of this thesis, a variety of statistical tests has been used which will be introduced in the following sections.

2.1.2 The students's t-test

The t-test finds application when evaluating the mean values of a set of samples to an expected value.

Given a sample $X = X_1, \ldots, X_n$ of probes independent to each other that are $N(\mu, \sigma^2)$-distributed with unknown mean μ and variance σ^2, the investigated hypothesis is defined as:

$$H_0 : \mu = \mu_0 \tag{2.1}$$

The test statistic then is defined as shown in equation 2.2:

$$T := \sqrt{n} \cdot \frac{\bar{X} - \mu_0}{S_n} \tag{2.2}$$

with the sample variance of the mean $S_n = \frac{\sigma}{\sqrt{n}}$.

This test can be applied for quality assurance, for example to ensure a cohort is within certain specifications: if one is interested if the mean durability time of a set of lightbulbs compared to the specification a manufacturer warrants.

Obviously, the test confidence grows with larger sample sizes.

In case the samples are not normally distributed, non-parametrical test methods are required. For any t-test, there is an alternative non-parametrical test method.

Two-sample t-test

The two-sample t-test is applicable for two samples $X = X_1, \ldots, X_m \sim N(\mu_X, \sigma_X^2)$ and $Y = Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma_Y^2)$ with homogeneous variance ($\sigma_X^2 = \sigma_Y^2$ and both samples are independent to each other.

Similary as with the one-sample t-test, the null hypothesis is defined alike in equation 2.3.

$$H_0 : \mu_X = \mu_Y \tag{2.3}$$

The test statistic in this case is defined in equation 2.4.

$$T := \frac{\bar{X}_m - \bar{Y}_n}{\sqrt{\frac{1}{m} + \frac{1}{n}} \cdot S_p} \tag{2.4}$$

with

$$S_p^2 := \frac{(m-1)S_{X,m}^2 + (n-1)S_{Y,n}^2}{m+n-2} \tag{2.5}$$

In order to accept the null hypothesis H_0, T should be near to 0, otherwise it is rejected.

If the variance is proven to be heterogeneous instead of homogeneous, e.g. applying a Levene-Test, the t-test is not suitable. An alternative would be the Welch-Test.

In the scope of this thesis, the two-sample t-test is applied for the differential analysis of gene and microRNA expression data as well as methylation analysis.

2.1.3 Hypergeometric Test

The hypergeometric test is based on the hypergeometric distribution and provides a means to compute the statistical significance of specific k probes from n draws from a population sized N with a total of K success probes. In other words, this test is applicable to determine whether the amount of successful draws is over- or underrepresented.

A random variable that is hypergeometrically distributed follows equation 2.6:

$$P(X = k) = \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}} \tag{2.6}$$

The p-value is calculated by $1 - \sum_K P(k)$.

2.1.4 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test allows to validate whether or not an observed cumulative frequency distribution of samples matches 1.) an expected distribution or 2.) the distribution of another random variable.

To apply this test, observed frequencies are arranged in ascending order, and each cumulative observed frequency F_i is calculated as the sum from f_1 up to and including f_i. From this, cumulative relative observed frequencies are determined by:

$$\text{rel } F_i = \frac{F_i}{n} \tag{2.7}$$

with the number of data in the sample $n = \sum f_i$.

This distribution then tested against either an cumulative relative expected frequency rel $\hat{F}_i$, which is calculated alike. The test statistic is calculated first by the partial calculation

$$D_i = |\text{rel } F_i - \text{rel } \hat{F}_i| \tag{2.8}$$

and

$$D_i' = |\text{rel } F_{i-1} - \text{rel } \hat{F}_{i-1}| \tag{2.9}$$

for each i.

Finally, D is the largest value of the largest D_i or D'_i:

$$D = \max(\max D_i, \max D'_i) \tag{2.10}$$

Depending on a desired significance level α and the amount of samples n, D is rejected when it exceeds a critical $D_{\alpha,n}$, for large n approximated by

$$D_{\alpha(2),n} = \sqrt{\frac{-\ln(\alpha/2)}{2n}} \tag{2.11}$$

as given by Smirnov [1948]. Other computations have been suggested and literature provides tables for those [Miller, 1956; Zar, 2007].

In a summary, the Kolmogorov-Smirnov test allows to search for a maximum deviation between the observed distribution F and the hypothetic distribution $\hat{F}$ and in case this deviation exceeds a certain threshold, the hypothesis that both curves follow the same distribution is rejected.

In the scope of this thesis, this test has been used to calculate the significance of similar gene frequency within co-regulated and co-targeted genes of motifs in TFmiR.

2.1.5 Multiple Testing Correction

When statistical tests are applied, the probability to find a certain result "by chance" - the p-value usually defines the threshold whether a hypothesis is accepted or rejected. While this works well for few tests, dealing with genomes and microarray experiments leads to several thousand separate hypothesis tests. Accordingly, when testing 20 000 genes with a p-value cut-off at 0.05, still about 1.000 genes may mistakenly be considered significant. The possible outcomes for a hypothesis test are shown in Table 2.1.

Thus, the probability of making an error α accumulates with the number m of hypothesis tests. The probability of not making an error in such a series of tests can be written as:

$$P(\text{No errors in } m \text{ tests}) = (1-\alpha)^m \tag{2.12}$$

with the probability to make at least one error in m tests $1-(1-\alpha)^m$. Controlling the Type I error rate is the aim of the p-value adjustment for multiple testing. In the scope of this thesis we applied the False Discovery Rate correction presented by Benjamini and Hochberg [1995] (BH-FDR).

With the amount of mistakenly accepted hypotheses V - false positives -, the False Discovery Rate is defined as the expected proportion of Type I errors among the rejected hypotheses R:

	Actual Situation	
Decision	H_0 True	H_0 False
Accept H_0	Correct decision $1-\alpha$	Incorrect Decision Type II Error β
Reject H_0	Incorrect Decision Type I Error α	Correct Decision $(1-\beta)$

Table 2.1: Possible outcomes of a hypothesis test with $\alpha = P(\text{Type I Error})$ and $\beta = P(\text{Type II Error})$. Minimization of Type I errors are the purpose of multiple testing correction methods.

$$\text{FDR} = E(\frac{V}{R}|R > 0) \cdot P(R > 0) \tag{2.13}$$

To control false discoveries, this rate is to be kept below a certain threshold q. For example, if the threshold was 0.10 with 1000 hypotheses rejected for 20 000 genes, less than 100 of those are expected to be false positives.

In general, to control the FDR for m tests at level q, the following steps are applied:

1. Order unadjusted p-values: $p_1 \leq p_2 \leq \cdots \leq p_m$
2. Identify the test with the highest rank j for which $p_j \leq \frac{j}{m} \cdot q$
3. Tests of rank $1, 2, \ldots, j$ are declared significant

2.2 Graphs

First, fundamentals of graphs and the underlying abstract model is presented.

2.2.1 Fundamentals

On the most abstract level, a graph can be understood as a binary two-dimensional matrix, where rows and columns indicate the source and target elements of a graph and the binary value indicates an existing relation, as shown in equation 2.14.

$$\begin{array}{cccc} & v_1 & v_2 & v_3 \\ v_1 & 0 & 0 & 1 \\ v_2 & 1 & 1 & 0 \\ v_3 & 1 & 1 & 0 \end{array} \tag{2.14}$$

More intuitively, a graph consists of a set of elements called nodes $V = \{v_1, v_2, \dots\}$, which are connected via links that indicate connections between nodes, e.g. with $e_1 = (v_1, v_2)$. The network from 2.14 is translated to a graphical representation in Figure 2.1. Note that, depending on the type of the graph - directed or undirected -, the graphical representation may neglect the bidirectional relations inherently whereas the matrix representation is bijective.

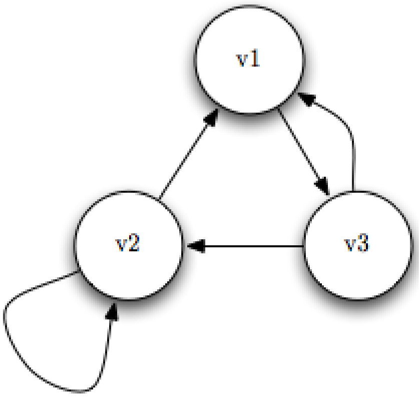


Figure 2.1: Example from 2.14 translated to a graphical representation

The number of links pointing from one node v to all other nodes is defined as the *degree* of a node.

2.2.2 Network Measures

For analyzing networks in general terms, a variety of centrality measures have been included in the scope of the work presented in Chapter 4.

Degree distribution

The degree distribution defines the probability for each node in a network to have a certain degree k. The trend of the curve in the resulting plot gives information about the characteristics of a network. In an evenly distributed random network the curve follows approximately a *Poisson distribution*, while the degree distribution of a scale-free network that is characterized by a few hubs with many links and a large amount of nodes with few links follows a negative power law [Erdős and Rényi, 1959; Newman, 2003]. In a biological context, the degree of a node in terms of a protein-protein interaction network can hint at which proteins are key drivers for metabolic processes in a cell.

Average path length

The average path length l_{G} is defined as the average number of steps along the shortest paths (d) between all pairs of all nodes (v_i, v_j) in a network G with size n:

$$l_{\mathrm{G}} = \frac{1}{n \cdot (n-1)} \cdot \sum_{i \neq j} d(v_i, vj) \tag{2.15}$$

Intuitively, for a biological interaction network to be efficient in terms of few intermediate steps, a short average path length indicates efficiency.

Density

The network density D denotes the ratio of the number of actual edges E existing in a network G to the number of possible edges, as shown in equation 2.16.

$$D = \frac{2E}{N(N-1)} \tag{2.16}$$

Network density has been shown to be an indicator of how robust a model network is to failures. But it has been shown for biological networks that even networks with low average density were robust to node losses. For that reason, Hayes et al. [2013] distinguished between local and global density and showed for real protein-protein interaction (PPI) networks, that the high local edge density in real world networks contributes to network stability as well.

Diameter

The length of the longest of all shortest paths in a network denotes the network diameter. This value is understood as a characteristic indicator for the linear size of a network.

Transitivity

The network transitivity - or clustering coefficient - is a measure for the clustering in a network.

Usually, one distinguishes between the global and the local clustering coefficient.

The global clustering coefficient is determined by the ratio of the actual number of 3 fully connected nodes to the number of connected triplets of nodes in a network, such as shown in equation 2.17

$$C = \frac{\text{no of closed triplets}}{\text{number of connected triplets}} \tag{2.17}$$

The local clustering coefficient measures the clustering coefficient for a single node i. In words, it is the ratio of all neighbours $N_i = \{v_j : e_{ij} \in E \wedge e_{ji} \in E\}$ of i that are interconnected to each other to all possible interconnected nodes, given by the size k_i of N_i, 2.18.

$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i \cdot (k_i - 1)} \tag{2.18}$$

Measuring the local clustering coefficient of a node in a PPI network, for example, may indicate the importance of the protein for the metabolism in question.

Closeness

The reciprocal sum of all shortest paths d from a node v to all other nodes in a network gives a measure for the centrality of a node, also called its closeness C, see Equation 2.19.

$$C(v) = \frac{1}{\sum_w d(v, w)} \tag{2.19}$$

This value, relative to the closeness of the other nodes in a network, gives a measure of how relevant a node is in a network in respect to the other nodes.

Note, since this value is an average, it may indicate very proximate distances to only a few nodes (with some very far away) or more similar distances to all nodes. A protein with high closeness could be heavily involved in regulation.

Betweenness

Another measure for the centrality of a node is the betweenness. For any node n and a pair of nodes v_1 and v_2, the total number of shortest paths linking v_1 and v_2 passing node n is calculated and related to the overall number of shortest paths between v_1 and v_2. A high betweenness indicates the importance of a node to maintain connections in a network.

Eigenvector

Other than the degree centrality of a node, the eigenvector centrality assigns a *quality factor* to each of the links connected to that node to weight the importance of the respective link. This factor is obtained by calculating the eigenvalues for the adjacency matrix and assigning those as weights. Intuitively the higher the eigenvector value of a certain node scores implicates a higher weight for the neighbouring nodes.

In a biological sense, a high eigenvector centrality protein is likely to be interacting with other important proteins, together resembling a critical part of an metabolic pathway.

In the context of drug discovery and a differential analysis of disease and healthy population, eigenvector centralities may help to reveal potential targets.

2.2.3 Minimum dominating sets

Originally developed to minimize resource allocation for wireless networks, Rai et al. [2009] presented an algorithm to find a minimum connected dominating set that represents the most efficient way to connect through the hubs in a network.

A connected dominating set (CDS) C of a graph G is the set S that induces a connected graph, and the minimum connected dominating set (MCDS) is the set with the minimal number of links necessary.

Because the problem is known to be NP-hard, it requires heuristics to solve a MCDS search in reasonable time [He et al., 2011].

For an adjacency matrix of a graph we define a vector X with X_i for each node $i = 1$, if the node has been recognized as a key node and 0 otherwise. Then, solving the optimization problem shown in equation 2.20 yields the dominating set of nodes.

$$\begin{gathered} \min \sum_{i=1}^{n} X_i \\ \text{subject to } \forall i \sum_{i}^{n} \text{adj } (i,j) X(j) \geq 1 \end{gathered} \tag{2.20}$$

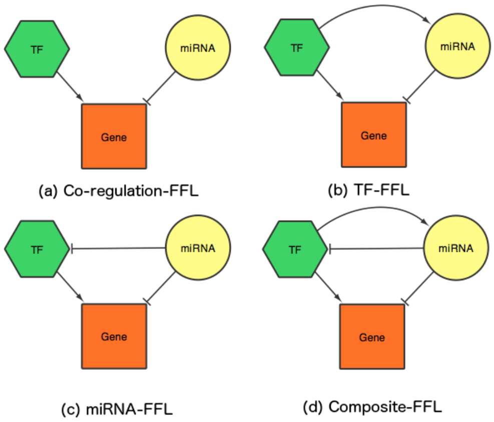
In collaboration with Maryam Nazarieh, the author adapted this algorithm to identify key driver nodes for regulatory networks, and we provide this functionality our webservice described in Chapter 4.

2.2.4 Network Motifs

Moreover, when it comes to understand network characteristics, an important question is whether the network can be decomposed into building blocks. This was shown in the transcription regulation of the bacterium *E. Coli* by Huerta et al. [1998] and Thieffry et al. [1998].

2.2.5 Motifs in TF-miRNA interaction networks

Key functional modules in a regulatory network are represented by Feed Forward Loops (FFLs). These have been shown to be important patterns in transcriptional regulation networks that are responsible in normal cell functions and diseases alike [Milo et al., 2002]. In Figure 2.2, the relevant motifs are shown.



(a) Co-regulation-FFL

(b) TF-FFL

(c) miRNA-FFL

(d) Composite-FFL

Figure 2.2: Schematic of the four motifs investigated in TFmiR, as shown in Hamed et al. [2015a]

Co-regulation feed forward loop (CR-FFL)

The co-regulation FFL resembles the regulation of a target gene by a transcription factor as well as the repression of the same gene by a miRNA.

miRNA feed forward loop (miRNA-FFL)

If a miRNA represses both the target gene and the TF which regulates the target gene as well, the miRNA-FFL pattern applies.

Transcription factor feed forward loop (TF-FFL)

This describes the regulation of expression of both miRNA and a target gene as well as the miRNA repression of the target gene.

Composite feed forward loop (C-FFL)

The composite FFL describes the most dense interaction between the three components. The transcription factor regulates both a miRNA and a target gene and the corresponding miRNA represses both the TF and the target gene.

2.2.6 Statistical evaluation

First, the hypergeometric test shown in Section 2.1.3 is applied to identify significant transcription factor and microRNA pairs that both regulate the same target:

$$p - \text{value} = 1 - \sum_{i=0}^{x} \frac{\binom{k}{i}\binom{M-k}{N-i}}{\binom{M}{N}} \tag{2.21}$$

with k the number of targets of the miRNA in question, N the number of genes regulated by a certain TF and x the number of common targets of both, and M the number of all genes regulated by miRNAs and TFs in the databases we queried. After applying the BH FDR correction (see 2.1.5), only remaining pairs with adjusted p-value < 0.05 were retained as significant.

2.2.7 Motif search algorithm

Since the motifs are similar and share an incremental structure, instead of searching for each motif individually, the algorithm could be applied based on the existing links in the network in such manner that all motifs are discovered during a single iteration over the whole set of edges in the network, which made the processing highly efficient.

Since all motifs share outgoing edges for a transcription factor, all edges that originate from a TF are identified. Then, for those edges targeting a gene, the gene is subsequently tested for miRNA regulations (edges incoming from miRNAs). For those microRNAs, the interactions to the original TF are checked and subsequently the motif can be classified as one of the four motifs defined above. Algorithm 1 shows a pseudocode listing of the motif search.

The motif search algorithm was implemented on network level using Java and Cytoscape.

Algorithm 1 Motif search

```
procedure MotifSearch
    MotifsList motifs
    for e: edges in network n do
        interactionType $\leftarrow$ InteractionType.determine(e)
        if interactionType.getSourceType() == 'TF' then
            Node tf $\leftarrow$ e.getSource()
            if interactionType.getTargetType() == 'GENE' then
                Node gene $\leftarrow$ e.getTarget()
                for Edge other : getAdjacentEdgeList(gene, INCOMING) do
                    InteractionType otherType = InteractionType.determine(other)
                    if otherType == 'miRNA' then
                        Node miRNA $\leftarrow$ other.getSource()
                        if n.containsEdge(miRNA, tf)
                                $\wedge$ n.containsEdge(tf, miRNA)
                                $\wedge$ n.getNeighbors(tf, OUTGOING).contains(miRNA)
                                $\wedge$ n.getNeighbors(tf, INCOMING).contains(miRNA) then
                            MotifType $\leftarrow$ COMPOSITE-FFL
                        else if n.containsEdge(miRNA, tf)
                                $\wedge$ n.getNeighbors(tf, INCOMING).contains(miRNA) then
                            MotifType $\leftarrow$ miRNA-FFL
                        else if n.containsEdge(tf, miRNA)
                                $\wedge$ n.getNeighbors(tf, OUTGOING).contains(miRNA) then
                            MotifType $\leftarrow$ TF-FFL
                        else
                            MotifType $\leftarrow$ CO-REGULATION
                        motifs.add(createMotif(motifType, tf, miRNA, gene))
    return motifs
```

2.2.8 Motif significance

In order to validate the significance of the motifs found in a certain network, we implemented a comparison to their occurrence in random networks based on the original network layout.

Network randomization

In order to retain the stronger attachment of key driver nodes, we decided to apply a degree-preserving randomization algorithm. For a network with L edges, two edges $e_1 = (v_1, v2)$ and $e_2 = (v_3, v_4)$ are randomly chosen for $2 \times L$ from all edges E of the network and rewired such that start and end nodes are swapped, i.e. $e_3 = (v_1, v_4)$ and $e_4 = (v_3, v2)$ if $\{e_3, e_4\} \notin E$.

Random network comparison

We calculate the p-value for a motif as follows:

$$p - \text{value} = \frac{N_h}{N_r} \tag{2.22}$$

which denotes the ratio of a certain motif to be acquired more or equal times in the tested network N_h to the number of created random networks, with $N_r = 100$ in our specific case.

Moreover, we calculate the Z-score for each motif type to investigate by how many standard deviations the observed motif was above or below the mean of random ones, defined as:

$$Z - \text{score} = \frac{N_o - N_m}{\sigma} \tag{2.23}$$

with number of motifs observed in the real network N_o and N_m, σ the mean and standard deviation of motif occurrence in the 100 random networks created.

2.3 Similarity Network Fusion

In order to investigate the individual datasets provided by Riemenschneider et al (see Chapter 5), we applied Similarity Network Fusion (SNF), a method proposed by Wang et al. [2014] which allows to integrate several different datasets retaining information about the individual samples. This approach was originally developed for computer vision and image processing by Wang et al. [2012].

2.3.1 SNF algorithm

The idea is to construct a graph $G = (V, E)$ that resembles a patient similarity network for each dataset. The vertices $V = v_1, \ldots, v_n$ correspond to the patients, while the edges E are weighted by how similar the patients are. For continous variables, with the Euclidian distance $\rho(v_i, v_j)$ between two patients i and j, the weight for this edge is determined by a scaled exponential similarity kernel defined as:

$$W(i,j) = \exp - \frac{\rho^2(v_i, v_j)}{\mu \epsilon_{i,j}} \tag{2.24}$$

with an empirically set hyperparameter μ and $\epsilon_{i,j}$ as a means to eliminate the scaling problem, defined as:

$$\epsilon_{i,j} = \frac{\rho(v_i, \bar{N_i}) + \rho(v_j, \bar{N_j}) + \rho(v_i, v_j)}{3} \tag{2.25}$$

with $\rho(v_x, \bar{N_x})$ being the average value of distances between v_x and each of its neighbors.

For discrete variables, the authors suggested using the chi-squared distance measure which in our work has been used to incorporate Apoϵ genotypes later on.

As the intention is to integrate different measured datasets, a fused matrix from the individual patient similarity networks had to be computed.

For this, a normalized weight matrix $P = D^{-1}W$ is calculated with the matrix $D(i,i) = \sum_j W(i,j)$ such that $\sum_j P(i,j) = 1$. This is called a full kernel on the vertex set V. To account for the normalization for self-similarities, the normalization is adapted for when $i = j$:

$$P(i,j) = \begin{cases} \frac{W(i,j)}{2\sum_{k \neq i} W(i,k)}, & j \neq i \\ \frac{1}{2}, & j = i \end{cases} \tag{2.26}$$

Now, to measure the local affinity of a node v_i to all its neighbors N_i (including v_i) in G, a k-nearest neighbors method applied as:

$$S(i,j) = \begin{cases} \frac{W(i,j)}{\sum_{k \in N_i} W(i,k)}, & j \in N_i \\ 0, & \text{otherwise} \end{cases} \tag{2.27}$$

That way, non-neighboring nodes are set to zero similarity values. While the matrix P contains full information about the similarities, S considers only the similarity to the k most similar patients for each patient.

The network fusion starts with P as initial states and incorporates S as a kernel to model the local structure of the graphs.

For m different datasets, similarity matrices $W_1, \ldots, W_m$ are computed using equation 2.24, likewise P_n and S_n from equations 2.26 and 2.27.

The idea behind SNF is to iteratively update the similarity matrix corresponding to each datatype. Assume two data sets for which two status matrices P^1 and P^2 have been calculated (iteration $t = 0$). The update process for both matrices is defined as:

$$P^1_{t+1} = S^1 \times P^2_t \times (S^1)^T \tag{2.28}$$

$$P^2_{t+1} = S^2 \times P^1_t \times (S^2)^T \tag{2.29}$$

with P^1_{t+1} and P^2_{t+1} the status matrices for the first and second datatype after t iterations, respectively. In that manner, status matrices are updated each time step in an interchanging fashion.

The final matrix after t steps is defined as:

$$P^c = \frac{P^1_t + P^2_t}{2} \tag{2.30}$$

In order to reduce noise, the method can be modified to include only common neighborhoods N_i of a node v_i:

$$P^1_{t+1}(i,j) = \sum_{k \in N_i} \sum_{l \in N_j} S^1(i,k) \times S^1(j,l) \times P^2_t(k,l) \tag{2.31}$$

As a result, if two nodes v_i and v_j share common neighbors in both similarity matrices, they are likely to be part of the same cluster. Also, if two nodes are not similar according to one dataset, the similarity within other datasets propagates into the other matrices during the fusion. After each iteration, the normalization from Equation 2.26 is applied after each iteration step to ensure that a node always is more similar to itself than to others and to guaranteee that the final network is full rank, i.e. a regular matrix with non-zero eigenvalues. This is important to apply classification and clustering to the final network.

The extension to more than two datasets is defined as an generalized version of Equations 2.28 and 2.29:

$$p^v = S^v \times \left(\frac{\sum_{k \neq v} P^k}{m-1} \right) \times (S^v)^T, v = 1, \ldots, m \tag{2.32}$$

The resulting fused matrix P^c can be used for clustering and classification of the nodes.

2.3.2 Network Clustering

In order to identify possible subtypes in a fused network graph with n samples and m measurements, we try to determine clusters of samples. A label vector is defined such that for each sample x_i, a vector is defined such that for each possible subtype k, $y_i(k)$, the value 1 and 0 is assigned depending on whether or not the sample belongs to the respective subtype. Thus, the partition matrix $Y = (y_1^T, y_2^T, \ldots y_n^T)$ describes the clustering scheme.

With SNF, spectral clustering is used to identify the network clusters by minimizing the RatioCut [Ng et al., 2001; Wei and Cheng, 1989] by solving the optimization problem defined as:

$$\min_{Q \in R^{n \times C}} \text{Trace } (Q^T L^+ Q) \text{s.t.} Q^T Q = I \tag{2.33}$$

with the scaled partition matrix $Q = Y(Y^T Y)^{-\frac{1}{2}}$, the normalized Laplacian matrix $L^+ = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ with similarity matrix W and the network degree matrix D (with 0 diagonal elements). This method has been shown to incorporate the global network structure [von Luxburg, 2007].

2.3.3 Normalized Mutual Information

The normalized mutual information (NMI) is a measure to evaluate a clustering against the number of clusters. The measure is defined as the ratio of the mutual information to the entropy H of clusters and classes, respectively.

The mutual information with clusters W and classes C is defined as:

$$I(W,C) = \sum_k \sum_j P(w_k \cap c_j) \log \frac{P(w_k \cap c_j)}{P(w_k) \cdot P(c_j)} \tag{2.34}$$

with $P(w_k)$ the probability for an element to be in cluster w_k, $P(c_j)$ the probability being class c_j and $P(w_k \cap c_j)$ the intersection of both.

The entropy H, a measure for the uncertainty of the respective outcomes for a probability function, is calculated as follows:

$$H(W) = -\sum_k P(w_k) \log P(w_k) \tag{2.35}$$

and analogously for the classes C.

Finally, the normalized mutual information can be written as:

$$\mathrm{NMI}(W,C) = \frac{I(W,C)}{[H(W) + H(C)]/2} \tag{2.36}$$

As both $H(W)$ and $H(C)$ each are an upper bound to $I(W,C)$, the dominator is divided by 2 to obtain values between 0 and 1 for the NMI score.

As the entropy increases with the number of clusters, the NMI score tends to value fewer clusters higher than a large number of clusters and, thus, allows to compare clusterings with different numbers of clusters.

We used NMI to compare the clusterings obtained with the similarity network fusion method using different combinations of data input sources.

Chapter 3

Mebitoo - an Extensible Software Framework for Bioinformatics Analysis Workflow Automatization

This chapter introduces the software framework Mebitoo. This software is a collaborative project for which the author contributed the mainframe software and plugin architecture, while additional plugins have been developed and used in various Bachelor projects supervised by the author. This work was published in Spaniol et al. [2015]. The case study on breast cancer describes a collaboration with Johannes Trumm and Mohamed Hamed that aimed at integrating regulatory network analysis.

3.1 Introduction

Initially, with Mebitoo we developed a software with the intention the perform analysis on membrane protein sequences [Spaniol, 2009]. However, applications in computational biology require to target a wider range of research fields beyond sequence analysis. When aiming at a more extensive view at biological data to gain a better understanding of complex processes, it has become popular to integrate data from various sources and perform a comprehensive analysis. For instance, in the field of complex disorders such as Alzheimers disease, the pathogenic factors have remained largely unexplained so far and it has been suggested that integration of various data such as gene expression, DNA methylation, single nucleotide polymorphisms (SNPs) and protein level measurements may lead to a better understanding of the disease [Rhinn et al., 2013; Zhang et al., 2013].

3.1.1 Related work

In the last years, various tools to pipeline data analysis have been presented. Web-based software like the Galaxy tool [Goecks et al., 2010] allow for interactive analysis of sequencing data. Users upload sequences to a remote public server or set up a local workspace environment. The software allows stepwise processing of generated data so that the workflow is highly adaptable to different types of analysis intended. Programmers can also extend the software using their own plugins to customize workflows with Python scripts. Another well-known tool is Geneious [Kearse et al., 2012], which is a Java Desktop application that allows for a variety of sequence manipulation analysis and may also be extended by own workflows or plugins.

While Galaxy performs on webserver basis that has the possibility to scale for large computational efforts, Mebitoo as desktop application does not require system administration knowledge or access to a server infrastructure and still qualifies for working with sensitive (e.g. clinical) data. On the other hand, Geneious emerged to become the standard software for many applications and excels as tool for the daily work with biological data and features a plugin development API that exceeds the concise plugin architecture of Mebitoo in complexity.

3.1.2 Motivation

With Mebitoo we developed a desktop software framework that enables for the implementation of a variety of analysis tools with the intention to minimize the overhead of plugin development. That way, the software is amenable to extensions by developers who can focus on implementing new methods while the framework provides the interface for data management and interaction with other modules. To manage module interaction, users are enabled to define tasks that automatically execute a variety of customized workflows. Mebitoo is set apart from other software by the ability to process arbitrary data supported by a database management system and XML (see Section 3.2.2), a task manager to enable workflows within a graphical user interface (Section 3.2.5) and the extendability with customized method plugins (Section 3.2.4).

3.2 Design

Mebitoo is a software application suite written in Java that is based on the Netbeans Rich-Client platform (RCP) project backed by Oracle.

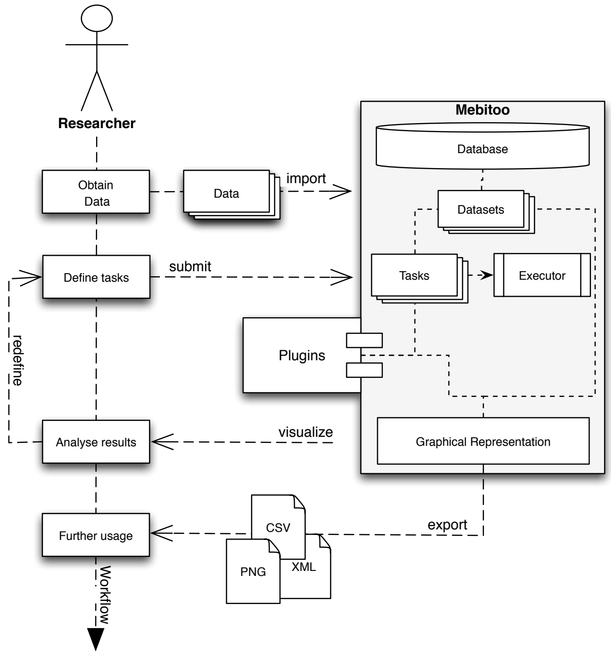


Figure 3.1: Mebitoo framework overview. The core (grey) provides the basic functionality for data import and task definition and execution as well as the plugin interface. Externally developed plugins can be mounted by the framework and executed by a task manager.

3.2.1 Core Module

The core module provides the framework for data management and processing. It (1) implements data import and storage, (2) defines the plugin system that provides an interface to communicate with extensions, and (3) provides a task management that is used to define and queue tasks that process the datasets. A framework overview is shown in Figure 3.1.

The data storage layout realized in Mebitoo uses the HSQL database engine and allows to store either collections of sequences as datasets or additional arbitrarily structured data as XML-documents.

The plugin structure is based on the module architecture provided by the Netbeans Rich-Client platform. It supports a concise abstract interface and provides plugin templates which can be easily adapted by developers to execute customized methods.

Tasks are defined as relations between plugins and datasets, that can be objected to parallel processing. Each task can be queued and finally submitted as a single processing job.

3.2.2 Database Design

Concerning the database itself, we provide a concise concept with few tables that are not allowed to be altered by plugin developers.

In order not to limit a plugin developer on the database concept presented here, if an application requires to apply custom data structures as applied when we worked with regulatory networks, the database design can be extended by plugins themselves. To maintain consistency for the core module, Mebitoo runs a database check at every startup. When working with separate projects, Mebitoo features an option to swap between different databases.

Entity Relationship Model

The diagram in Figure 3.2 shows a simplified entity relationship model, which omits the primary and foreign key attributes. Those have to be added for each entity and relation to obtain a complete model when the database is actually implemented.

Each entity and relation is modeled in the database, except for the plugins. Those are loaded from the file system dynamically. Instead of modeling a plugin representation within the database that requires to synchronize between database and plugins, we use the plugin identifiers (e.g. `org.mebitoo.plugin.aligner`) as symbolic foreign keys and use them to directly retrieve plugins from the module system service provider.

Any dataset may contain an arbitrary amount of sequences, but no sequence can belong to more than one dataset. A task can be linked to any dataset and plugin, but only once to every entity. This restriction is a key predicate for the task execution model. A plugin data entity has to be unique for each dataset and plugin – regardless of the plugin version – but any dataset can have as many data entries as there are plugins and a plugin may also produce XML files for any dataset.

Schema

The final database design applies to the schema shown in Table 3.1.

3.2.3 Dataset Structure

In its abstract sense, a dataset in Mebitoo is a collection of data that belongs to one certain entity. For example, a dataset could consist of a single sequence, one sample of sequences from a cohort, or an entire cohort itself. Whenever a task is processed, the dataset as a whole is passed to the processing. Thus, it depends on the modules how to interpret the plugin data associated with a certain dataset.

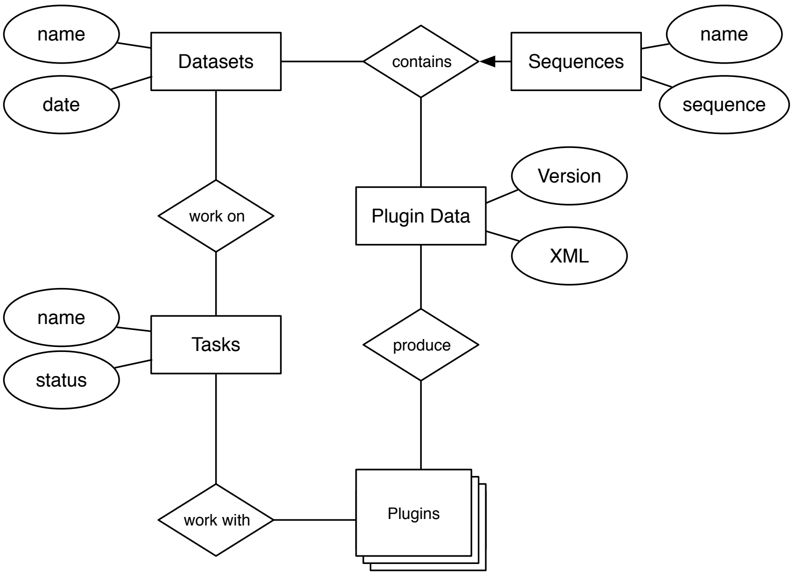


Figure 3.2: A compact Entity Relationship model of the Mebitoo database.

Table 3.1: Database Schema

Datasets	(*id*: `int`, *name*: `string`, *date*: `bigint`)
Sequences	(*id*: `int`, *dataset_id*: `int`, *name*: `string`, *sequence*: `string`)
Tasks	(*id*: `int`, *name*: `string`, *date*: `bigint`, *status*:`int`)
Plugin_Data	(*id*: `int`, *dataset_id*: `int`, *plugin_id*: `int`, *plugin_version*: `string`, *date*: `bigint`, *xml*:`CLOB`)
Task_Plugins	(*id*: `int`, *task_id*: `int`; *plugin_id*: `int`)
Task_Datasets	(*id*: `int`, *task_id*: `int`; *plugin_id*: `int`)

We defined the plugin data as XML documents that are stored as OTHER objects in the HSQL database. Basically an XML document is created for each dataset and plugin tuple. Thus, when a plugin and dataset are selected, the dataset container forwards the request to the database interface and builds a JDOM from the XML file on-the-fly.

Other than dataset, sequence, and task information, the plugin data is not read from the disk until required but the plugin data table is realized as a cached table. This affects the performance negatively because reading data from the hard disk

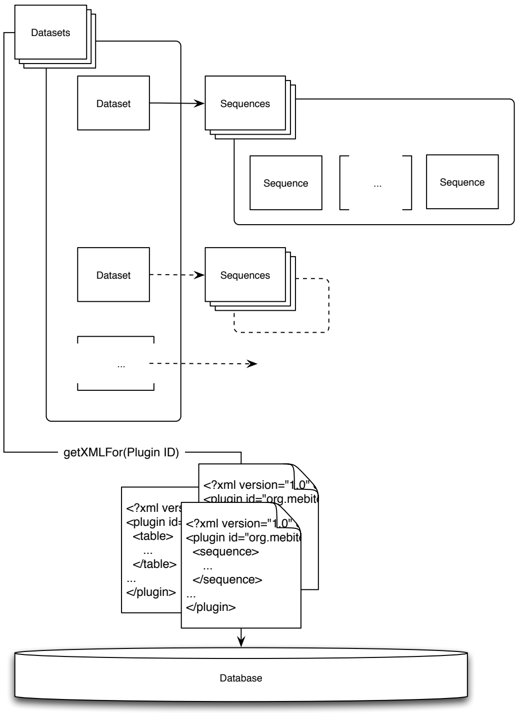


Figure 3.3: By default, a dataset consists of a set of string entries or arbitrary data stored in any XML-consistent format. During runtime, all datasets and the containing sequences are loaded into memory. On the other hand, the associated XML-files that contain the data of each plugin are not loaded from disk until required, e.g. during plugin processing. This is called a "lazy" initialization. In that way, we avoid the possible large memory requirements when loading the entire XML structure.

takes up significantly more time than accessing the RAM. We decided to accept this in trade-off for the possibility to store large XML files, although for small datasets it would be better to have in-memory tables to favor better runtimes.

In summary, the storage strategy of Mebitoo resembles a two fold concept: the dataset index and task tables, obligatory for the actual *use* of the software, are available in memory any time. Figure 3.3 illustrates the concept of a dataset. The plugin data is stored in XML files, which are completely open to the needs of a plugin developer. Instead of writing XML documents to the hard drive, what would entail the implementation of input/output (I/O) handling, we cache the files using the database engine and only load them into resident memory on demand.

3.2.4 Plugin Interface

The plugin system in Mebitoo adapts the concept from the plugin system that is implemented in the Netbeans RCP. We introduced an additional interface and provide an empty plugin template which can be easily customized to integrate methods into Mebitoo.

Thus, each plugin that should be part of the task processing system is required to implement a interface that is understood by core module. Basically, a class that complies with the required interface implements two functions that make the data available for processing and displaying objects.

Plugin Main Class

A plugin class should implement this interface:

```
@ServiceProvider(service=Plugin.class, path="org.mebitoo.plugin")
public class Template extends AbstractPlugin {
  DatasetProcessor getDatasetProcessor();
  DatasetView getDatasetView();
}
```

The easiest way would be to have static instances of both objects within the plugin and return them.

In fact, the implementation of both the dataset processor and the view is the more important part:

Dataset Processor

A dataset processor class is responsible for the computations performed on one dataset. The interface is very simple:

```
public class TemplateDatasetProcessor
    implements DatasetProcessor {

  public Result process(Dataset dtst) {
    ...
  }
}
```

Dataset View

Mebitoo automatically creates a window for each plugin. The dataset view is the interface for your plugin to publish the main panel to Mebitoo and defines the logic how to update a panel:

```
public class TemplateDatasetView implements DatasetView {
  // called to hook the panel in mebitoo
  public JPanel getRepresentation();
  // update on dataset focus change
  public void update(Dataset dtst);
  // empty the view, e.g. happens if no dataset selected
  public void reset();
}
```

3.2.5 Tasks

A key design requirement of Mebitoo is to easily queue and launch bulk dataset processing with little effort. This has been accomplished by introducing tasks that implement an interface. This interface can be understood by Java executors, which are part of the concurrency package.

A user can set up a task via a GUI-Dialog in which he specifies the plugins and the datasets.

Each of these datasets is processed by each plugin specified when the task is executed. The newly created task is queued and can be executed at any time the user invokes the processing by clicking on the "Start" button.

The task execution procedure can be understood as a three-fold loop, see Figure 3.4.

It is possible to queue many tasks, which are processed in serial order. This decision has been made to ensure that only one dataset is processed at a time, in order to avoid any anti-dependencies. Although those could be resolved easily, by synchronizing the access to the database, and if the same plugin accessed the same dataset, merging the XML trees, it lacks the necessity for such functionality. If a user wants all datasets to be processed at the same time, it is possible to queue all datasets into the same task.

By default the processing of datasets within a task happens in parallel, based on the number of CPU cores available. To assert correctness we define all datasets independent of each other and a plugin is not allowed to access foreign datasets during processing. This behavior restricts data processing to the scope of a dataset, a feature that has been sacrificed in order to allow execution parallelism.

The innermost loop, the plugin processing, is executed sequentially again, so we inherently solve any dependency issues when plugins depend on each other's data.

The greatest benefit of this design can be achieved by adding as many datasets and plugins as possible to a single task, which would result in a maximum execution parallelism.

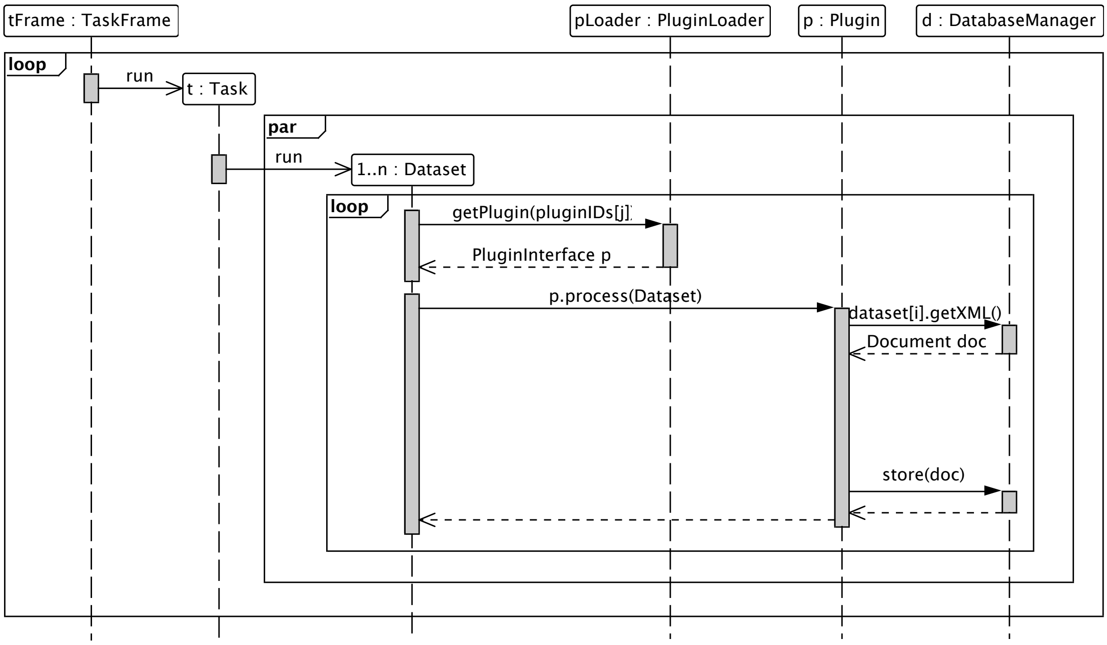


Figure 3.4: Sequence diagram for task processing.

3.3 Gene Regulatory Network Plugin

The GRN query plugin takes as input an undirected gene co-expression network (see Figure 3.5a) and then attempts to query several publicly available databases for regulatory information associated with the input co-expression network. Namely, it integrates data from the regulatory databases Transcriptional Regulatory Element Database (TRED) [Jiang et al., 2007], Molecular Signatures Database (MSigDB) [Liberzon et al., 2011] and JASPAR database [Sandelin et al., 2004] as data sources for identifying transcription factors, known regulatory interactions and associated binding motifs. At first, all those genes in an input network are marked that are listed in at least one of the databases to code for a transcription factor, as shown in Figure 3.5b, which shows an exemplary network.

Then, for each of these TF-genes, the tool searches whether the databases contain a known regulation of a target gene. In each such case, a directed edge is added between the transcription factor and the target gene.

Finally, the plugin uses the Motif Statistics and Discovery (MoSDi) software [Marschall and Rahmann, 2009] to run a motif search for all known binding motifs of the transcription factors represented in the current network against the promoter regions of all genes in the network. If a match is found, a new directed edge is added from the transcription factor to the gene. Moreover, the user has the option to expand the network by (a) adding further transcription factors that are annotated

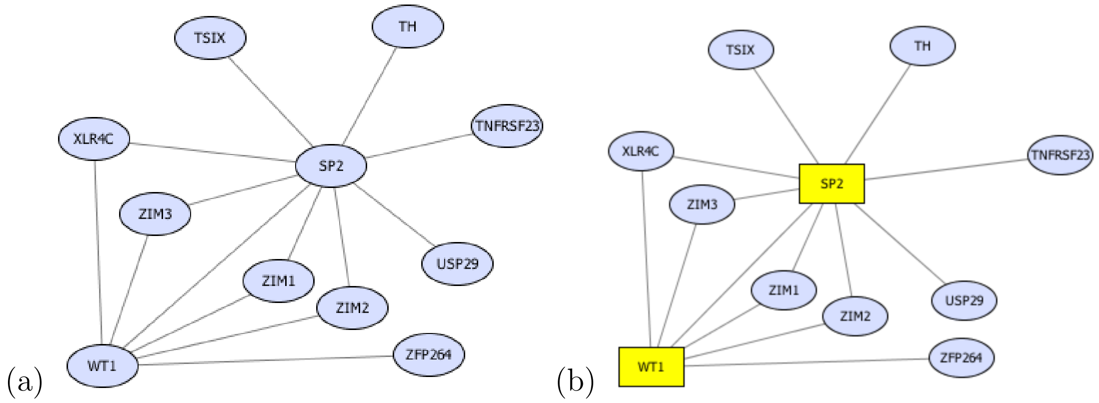


Figure 3.5: (a) Network visualization for the input co-expression network. (b) Transcription factors involved in the input network are identified and marked in yellow while the remaining genes are colored blue.

as known regulators of the input gene network, and/or by (b) including additional target genes that are annotated to be regulated by the TFs in the input network, and/or by (c) searching for regulatory interactions between the additional target genes and the additional TFs. Finally, the user can export the resulting network, as shown in Figure 3.6 in various formats as an image or as a network file (.sif) to be imported, visualized, and analyzed by other network analysis tools, such as Cytoscape [Smoot et al., 2011] or Visant [Hu et al., 2013].

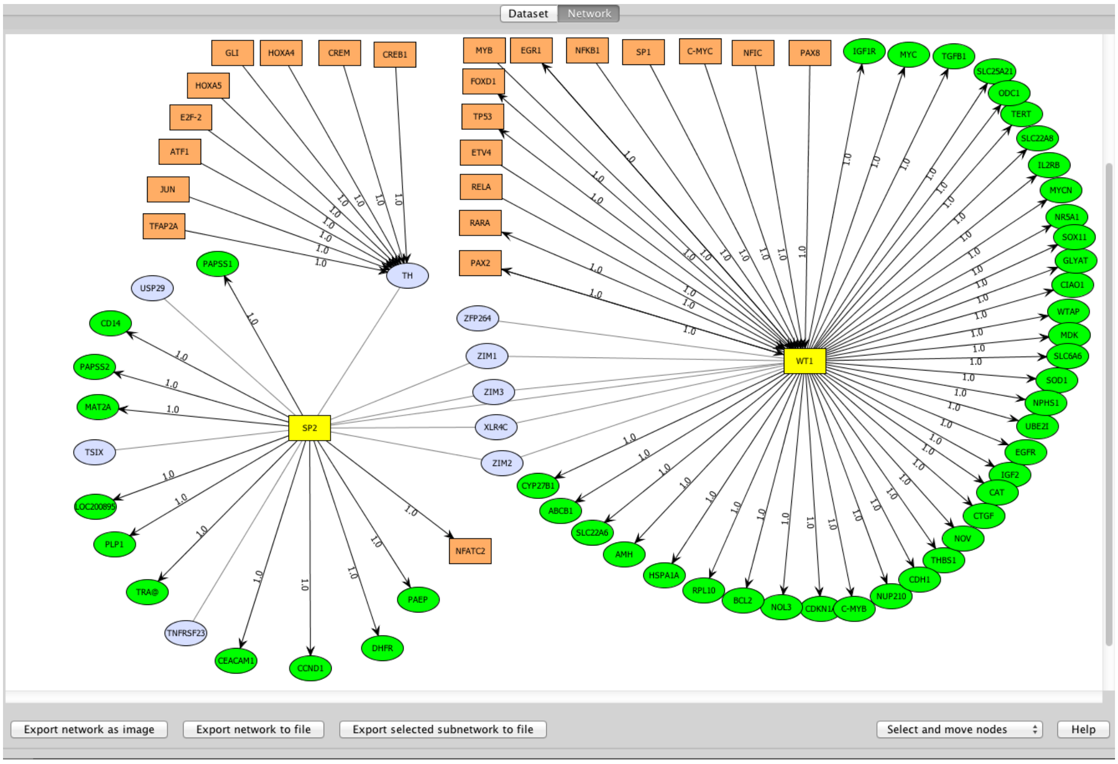


Figure 3.6: GRN query plugin applied to a gene co-expression sub network around the known imprinted genes (IGN). Transcription factors involved in the input network are identified and marked in yellow while the remaining genes are colored blue. The tool expands the IGN network by adding additional transcription factors (marked in orange) that are annotated as known regulators of the input genes and also by adding additional target genes (marked in green) that are annotated to be regulated by the TFs in the input network.

3.4 Case study - Integrative network-based approach identifies key genetic elements in breast invasive carcinoma

The following study, to which the author contributed in the key driver identification, is a collaborative work published in Hamed et al. [2015b]. As this study motivated the development of TFmiR, presented in Chapter 4, the general concept of the methology is depicted in the following sections.

3.4.1 Motivation

Using the integrative network approach to associate regulatory networks with the development of breast carcinoma we incorporated data from gene expression, DNA methylation, miRNA expression as well as somatic mutation datasets.

3.4.2 Approach

The datasets used for the integrative approach were collected from The Cancer Genome Atlas (TCGA) [Cancer Genome Atlas Research Network, 2008] . We investigated 151 cases and 20 controls for which all four data types we wanted to incorporate were available at TCGA.

The expression and methylation profiles were analyzed using 1) Significance Analysis of Microarray (SAM) , 2) a moderated student's *t*-test, and 3) the area under the curve receiver operator characteristics (AUC ROC) . All markers that were identified as significantly differentially expressed by at least two of the methods were accepted into the list of differentially expressed markers for genes, miRNAs, and methylated regions alike.

Differentially co-expressed gene clusters between cases and controls were identified by us using WGCNA[Langfelder and Horvath, 2008] and DiffCoEx [Tesson et al., 2010].

Based on the obtained data, we constructed a gene regulatory network as explained in section 3.3.

Additionally using the coexpression modules, we created a causal probabilistic Bayesian network and used the edges from the gene regulatory network as seed point to deduce node directionality. In order to evaluate the network topology, the Sparse candidate and likelihood-equivalence Bayesian Dirichlet methods were used to score the final network. At each iteration, edges could be added, removed or reversed as well as have the parent node swapped. This algorithm was applied three times and all edges that were inferred at least two times were considered for the next step.

The final network was built from the directed edges obtained by the three networks above. Subsequently, methylation and expression profiles were used to prune the network with respect to the data available. As increased methylation levels in gene promoter regions were shown to be responsible for reduced expression of those genes, anti-correlating regulatory interaction with respect to their expression and methylation profiles were removed from the network.

In order to put differentially expressed microRNAs and differentially expressed genes into context, we used miRTrail [Laczny et al., 2012] to identify the microRNA

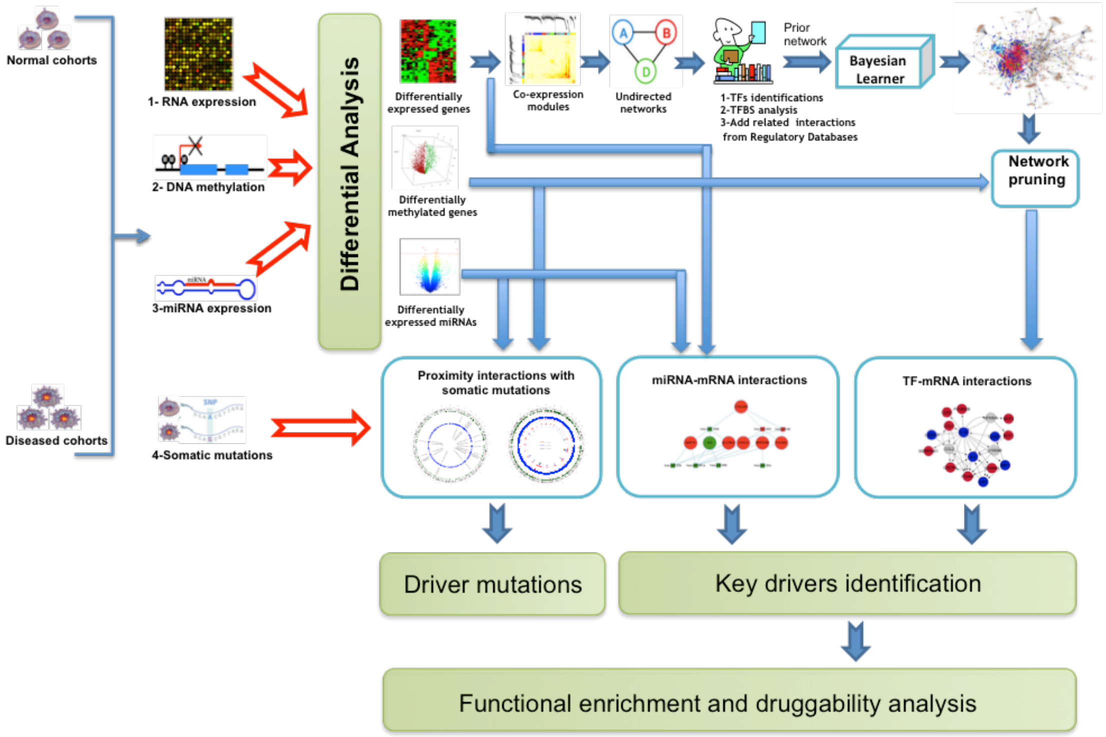


Figure 3.7: Overview integrated systems approach for key driver identification. Source: Hamed et al. [2015b]

Targets and calculated the intersection with the set of obtained differentially expressed mRNAs.

The other way around, we used TransmiR [Wang et al., 2009] to identify the genes (TF) that may be responsible for the regulation of the differentially expressed miRNAs.

The results were validated using the hypergeometric test (see Section 2.1.3) and merged into a final network. An overview of the method is shown in Figure 3.7

3.4.3 Results

In the end, we identified a minimal set of nodes that regulate the entire network using the GNU Linear Programming Kit (GLPK) [Makhorin, 2004] and OpenOpt [Kroshko, 2007] (covered in Section 2.2.3), which we assume to be possible key drivers for the regulatory network we generated.

In an extended enrichment analysis, we used overrepresented genes of in our study set and identified pathways from KEGG and GO functional annotations using the Database for Annotation Visualization and Integrated Discovery (DAVID)

[Dennis et al., 2003], and evaluated them applying a hypergeometric test. For microRNA enrichment, TAM [Lu et al., 2010] was used.

For the dataset of 131 cases and 20 controls, we obtained 1317 differentially expressed genes, 2623 differentially methylated gene promoter regions and 121 differentially expressed miRNAs.

The results of the study showed strong association between the regulatory elements of the heterogeneous data sources in terms of the interchangeable regulatory influence and genomic proximity. By analyzing three different types of regulatory interactions: TF-mRNA, miRNA-mRNA, and proximity analysis of somatic variants, we were able to identify various key driver elements (106 genes, 68 miRNAs, and 9 mutations) that could possibly drive the cancer developmental process and thus contribute to a better understanding of cancer development and new therapeutic strategies.

The derivation of a regulatory network from expression data and the downstream analysis with respect to various diseases has been the motivation for developing TFmiR, the webservice we present in Chapter 4. The approach presented in this study is applicable to other cancer types as well as other diseases, thus we applied the approach shown here also in our study on Alzheimer's disease (Chapter 5).

3.5 Discussion

We presented Mebitoo as a desktop software framework for sequential data processing pipelines. The core module provides functionality for data storage using a database engine and defines an interface for developers to implement their own methods via plugin extensions. The task system allows for large-scale automated execution of time-consuming processes. The usage of the Netbeans platform ensures future compatibility with possible upcoming hardware architectures and Java Runtime Environments. The RCP provides a large variety of functionality that developers may deploy in their own modules and enables the usage of a large variety of third-party libraries when desired.

We integrated a database engine to store datasets and use cached tables to enable the storage of large XML files. Pre-defined functionality enables to import sequence datasets either manually, as FASTA files either solitarily or grouped in folders using the BioJava library.

The software is extendable by mounting plugins that communicate with the core module. Plugins are derived from an interface class that provides functionality to assist inexperienced programmers with the development of new modules. The methods realized by the interface cover the data exchange between plugins and the database using JDOM.

In order to require little interaction with the end-user and to have automated workflows, we defined a task concept that allows for automated processing of datasets by plugins. Moreover, we modeled a thread-safe concurrency scheme for the task execution to enhance the processing performance for multiple datasets.

We found the decision to use Java and supplementary libraries reasonable, as the language supports fast application development. Moreover, the functionality of third-party modules enhanced the prototyping process because we were not required to implement all desired features on our own. We expect the database to handle at least two gigabytes of data because HSQL has been reported to easily deal with that size. According to the developers, HSQL theoretically supports a maximum capacity of 16 terabytes while the *character large objects* (CLOBs), which in case of Mebitoo contain the XML-information of the respective plugins, may hold multiple gigabytes.

Our approach to model the interaction between plugins and the main application is practical but plugin instances are not immutable by design. An immutable concept appears to be superior regarding the concurrent processing of data because of their inherent thread-safe nature [Bloch, 2001], but would interfere with plugin development. This would also enable the application of multiple dataset viewer frames, since that way a currently chosen dataset to be represented in the GUI is decoupled from the state of plugin instances.

In different projects, that were either conducted by the author in the course of this thesis (but are not described here due to space limitations) as well as in the Bachelor theses of Thorsten Klingen, Stefan Helfrich and Mustafa Kahraman that were supervised by the author, we connected the Mebitoo GUI to various command-line tools, such as the alignment tool BLAT, the motif search algorithm MEME, and recombination analysis with Recco in order to provide a pipeline working with those tools. We took different approaches using the Java Native Interface (JNI) to spawn C++ instances as part of the Java main application (MEME) and the Java Remote Message Invocation (RMI) to enable a client/server architecture to outsource complex computations with BLAT. In summary, the dependency on different platforms and various libraries turned those efforts into tedious endeavours and the projects showed that a limitation to command-line calls is the most feasible method to do this. However, tracking progress within the task manager for such methods is difficult and impairs the responsiveness of the application.

3.6 Outlook

Further development of Mebitoo is omnifarious by design. Although the plugin interface is non-varying, the restriction is one way. This means that Mebitoo is not supposed to adapt to new plugins but those are expected to implement a basic interface that is versatile enough to enable specialized development, taking advantage of the Java swing framework to do so.

The other way around, because a plugin is allowed to use all libraries the main program can provide, adding functionality to the core program can easily be accomplished. This flexibility permits further development as needed.

Possible enhancements could be to incorporate and include more probabilistic graphical approaches and other biological data types into Mebitoo to capture further insights on the regulatory regime of biological processes and human disease networks. For example, as shown in the case study on breast cancer, by incorporating the Bayesian learning approach, GO terms, and KEGG pathway knowledge, the network structure can be refined and made more informative with respect to a specific disease or a cellular process.

Chapter 4

TFmiR: A web server for constructing and analyzing disease-specific transcription factor and miRNA co-regulatory networks

The following project is a collaborative work to which the author contributed the web frontend software and the backend interface to handle the execution of the R scripts written by Mohamed Hamed as well as a Cytoscape plugin to perform TF-miRNA-Gene motif search on regulatory networks. This work was published in Hamed et al. [2015a].

4.1 Motivation

In the last chapter we presented a desktop software application to integrate analysis pipelines. However, if the analysis reaches a certain complexity in terms of computational requirements, it is feasible to gain additional power using a server architecture. Additionally, some of the downstream analysis presented before relies on R scripts and various 3rd party libraries which interfered with the fundamental concept of Mebitoo not to be dependent on additional software packages that have to be maintained outside the platform. Both considerations motivated the design of a web service that incorporates the software we used to build and analyze co-regulatory networks shown in the previous chapter.

Based on the input of differentially expressed mRNAs and/or microRNAs we extended our tool to provide a comprehensive downstream analysis to assess network motifs, network key drivers and investigate functional enrichments, optionally in context with specific diseases.

4.2 Methods

TFmiR is a software framework based on various programming languages and web technologies. The server-side computation of our methods was written in R. This is based on a generalized version of the scripts written for the work presented in the breast cancer study, see section 3.4. We refer to this part of the software as *backend*. The graphical representation of TFmiR consists of a *frontend* that is based on PHP, JavaScript, Java and various libraries thereof.

4.2.1 R statistical computing

The backend to a large extent consists of scripts based upon the R programming language. R has been chosen because it provides a large set of bioinformatics utility libraries, especially the Bioconductor software [Gentleman et al., 2004] and third party contributors thereof.

In the following sections, we briefly depict the software and tools incorporated into our framework.

4.2.2 igraph

The *igraph* package provides various tools to create, manipulate, and to visualize networks. It provides interfaces to high-level languages like R and Python and is capable to handle large graphs efficiently [Csardi and Nepusz, 2006].

Because of its versatility, the package was used in TFmiR to conduct backend computations on the generated networks as well as to generate network plots.

4.2.3 Bioconductor

Bioconductor itself is an extensible software project specifically developed for computational biology and bioinformatic applications, especially for analysis of high-throughput sequencing data including DNA and RNA sequencing, methylomes and downstream analysis such as annotations and graph analysis. It has been widely accepted in the scientific field so that it comprises nearly 1 000 packages so far written by a large community and enables rapid workflow implementation [Huber et al., 2015].

GoSemSim

The package GoSemSim by Yu et al. [2010] is used to compute semantic similarities between genes. Specifically, we compute gene functional similarities for our genes as well as a random set of genes from ENTREZ in order to rule out significant functional similarities the network motifs we identified using the Kolmogorov-Smirnov test, see Chapter 2.1.4.

4.2.4 HTML

The structure of any website is annotated using the Hypertext Markup Language (HTML). Based on the markup, the content of a website is rendered by a browser and displayed to the user.

4.2.5 PHP

The server-side frontend processing was implemented using the PHP Hypertext Preprocessor (PHP), which is nowadays - with over 80% of all websites based on PHP - the most used programming language to implement web services. The advantage over using plain HTML is that it enables a dynamic creation of the content displayed to the user. This meets the major requirement of TFmiR to allow interaction depending on user input data.

4.2.6 Javascript

In order to offer more responsiveness, interactive websites take advantage of the JavaScript language. Other than PHP, JavaScript is executed on the client-side, i.e. within the respective web browser. While this raises compatibility issues between several implementations of the scripting engines used in different browsers, such as between Microsoft Internet Explorer, Firefox, and Chrome, Javascript allows for asynchronous queries. Those allow for the processing of data in the background while the frontend interface is still usable and enables better interactivity in order to create a look and feel that narrows the gap between regular desktop applications and interactive websites.

4.2.7 JQuery

As Javascript evolved to be a key factor for interactive websites, entire libraries have been designed to simplify the client-side scripting of HTML pages. Among many others, the JQuery library emerged to be outstandingly the most popular framework today. The library allows for manipulation of DOM elements in a document and supports asynchronous queries and accounts for different browser engines, which makes its usage independent from the browser that is used client-side. The capability to spawn asynchronous processes, document manipulation and the compatibility to the Cytoscape Javascript library (see next Section) are used in various features implemented in TFmiR.

4.2.8 Cytoscape.js

A visual representation of the networks built was implemented using the Cytoscape Javascript library, which is compatible with the previously mentioned JQuery library. As this shares an interface to the well-known network visualization software Cytoscape [Shannon et al., 2003b], we used this to realize various layouts.

4.2.9 Apache

Underneath the entire framework runs a Apache webserver 2.0, an open-source project that was chosen since it proved to be a reliable server technology with reasonable performance and, thus, is worldwide accepted with ≈ 60% of all websites running Apache today.

4.2.10 Javascript Object Notation

The Javascript Object Notation (JSON) is a compact file format designed for data exchange between applications. It supports various basic datatypes such as numbers (whether integer or real), Strings, Arrays, and Objects. A brief example for a motif object JSON is shown in Listing 4.1.

Listing 4.1: "JSON Example"

```
{
        "type"      :       "COMPOSITE-FFL",
        "tf"        :       "ESR1",
        "mirna"     :       "hsa-mir-221",
        "gene"      :       "TP53"
}
```

In contrast to XML, JSON is less versatile but reduces the markup overhead which is favorable for transmission over networks.

We wrote a script in TFmiR that generates JSONs from the network interaction table in order to communicate with Cytoscape.js and Cytoscape.

4.2.11 Representational State Transfer

REST is a programming paradigm for distributed systems. In short, it states that a certain URI always has to contain the same content (but not identical). As an example, a news website would not be REST conform due to its changing content, other than a website that always displays the current weather in the same style and format. A REST Interface offers the functionality to access content using `GET`, `POST`, `PUT`, and `DELETE` operations and requires to be stateless, which means a server can not handle user sessions. Data transfer with rest usually occurs using HTML, JSON, or XML. REST has been used to realize the communication between TFmiR and the Cytoscape motif search plugin.

4.2.12 Cytoscape

Cytoscape evolved to be the standard application when dealing with networks and their visualization in computational biology. Based on Java, it offers a plugin architecture that enables to include a variety of tools by written by third party developers.

After an initial implementation of the motif search in R, it became apparent that R is too expensive in terms of execution time. Thus, we decided to increase the performance by implementing a plugin for Cytoscape, from which we expected the algorithm to perform significantly faster.

CyREST

CyREST is a plugin for Cytoscape developed by Keiichiro Ono [2015] that offers REST services for Cytoscape. We adapted the plugin and modified the source to support a REST query for the motif search in our regulatory networks in Cytoscape.

4.2.13 Databases

Within TFmiR we incorporated a variety of different interaction databases to model a regulatory network from the input data, an overview is given in Table 4.1.

In order to increase performance, the databases were mirrored into RData binary packages on our server.

4.3 Input data

As input for TFmiR for differentially expressed miRNAs and mRNAs alike, a user provides a file that contains the official symbol or miRNA identifier and a value

Table 4.1: Overview of the databases used in TFmiR. (P) means predicted interactions and (E) means experimentally validated interactions.

Interaction	Databases (P/E)	Published	Genes	miRNAs	Regulatory links	Version /frozen date
TF → Gene	TRANSFAC (E)	Matys et al. [2002]	1279	–	2943	V11.4
	ORegAnno (E)	Griffith et al. [2008]	1132	–	1083	Nov 2010
	TRED (P) [3]	Jiang et al. [2006]	3038	–	6462	2007
TF → miRNA	TransmiR (E)	Wang et al. [2009]	158	175	567	V1.2, Jan 2013
	PMID20584335 (E)	Qiu et al. [2009]	58	56	102	Apr 2009
	ChipBase (P)	Yang et al. [2012]	119	1380	33087	V1.1, Nov 2012
miRNA → gene	miRTarBase (E)	Hsu et al. [2010]	2244	551	5640	V4.5, Nov 2013
	TarBase (E)	Sethupathy et al. [2006]	422	79	492	V7.0
	miRecords (E)	Xiao et al. [2008]	543	157	780	Mar 2009
	starBase (P)	Yang et al. [2011]	5720	249	56051	V2.0, Sept 2013
miRNA → miRNA	PmmR(P)	Sengupta and Bandyopadhyay [2011]	–	312	3846	Mar 2011

$\{1, -1\}$ which indicates whether the corresponding markers has been classified as up- or downregulated within the sample dataset, an example shown in Listing 4.2.

Listing 4.2: Example TFmiR input for differentially expressed mRNAs

```
AATK     −1
ABCB8     1
ABCG4     1
ABHD10    1
ABLIM1   −1
ABT1      1
...
```

4.4 Frontend and Interface Architecture

The user frontend is divided into three separate parts: the front page for data input, the results page that displays the overall results and finally, detailed views for motifs, the respective co-regulated/co-targeted subnetworks, and functional similarity analysis.

In order to save computational time, the processing in TFmiR is split into three stages: first, the overall networks are calculated. Then at convience, TFmiR offers the option to perform a motif search on the created networks. Finally, a user may select a motif of interest and calculate functional similarities for the participants of the respective motif.

4.4.1 Front page

The first page presented to the user is the front page that displays the input forms for the files which contain the sets of deregulated microRNAs and mRNAs. Additionally, a user may define a p-value threshold for the analysis and, if available, specify the disease associated with the submitted dataset. Finally, there is an option to set an evidence level which offers the possibility to consider only experimental or predicted databases, or both. Finally, a user can initiate the first-stage processing.

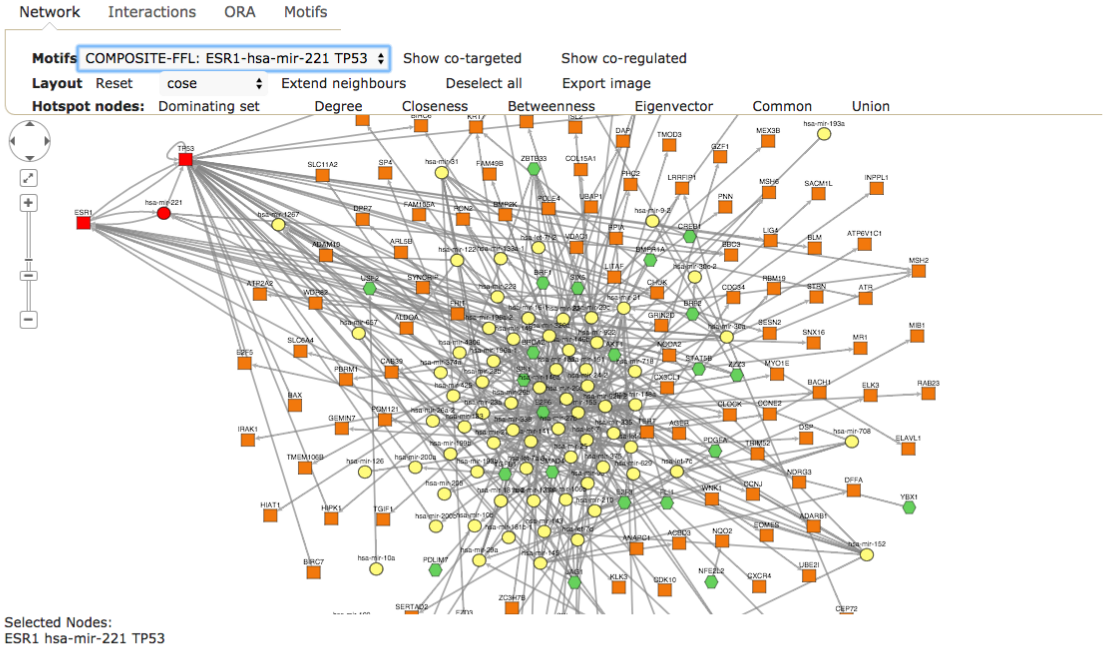
The first processing step is handled by a Javascript query that submits the datasets together with the options to the R script backend, where the respective interactions are calculated and finally merged into a full interaction network and - if specified - a disease specific network. When finished, a Javascript routine updates the front page and displays an overview of graphical buttons for the generated networks as well as for the individual interactions between miRNA, genes, and transcription factors, respectively. Those buttons display a brief summary, such as graph size, overlap for the respective interactions, and significance and lead to the more detailed result page.

4.4.2 Result page

The result page is split into four different parts: the network view, the interaction table, the overrepresentation analysis for both genes and miRNAs, and finally a motif view.

Network View

The first panel shows an interactive view that displays the network as a whole. An screen excerpt of the web service is shown in Figure 4.1. We offer a toolbar to modify the network layout and to select arbitrary nodes and a list of the selected nodes is shown at the bottom. Any network layout can be exported as an image. Additionally, the motif search and the minimum dominating set determination can be invoked on user demand.



Motifs COMPOSITE-FFL: ESR1-hsa-mir-221 TP53

Layout cose

Selected Nodes:
ESR1 hsa-mir-221 TP53

Figure 4.1: Shown are parts of the network view panel of TFmiR. The upper toolbar offers various options to layout the network and. Nodes can be selected based on various network measures. Additionally, a user may invoke a motif search, or highlight the respective motifs when the analysis is done. For any selected motif, co-targeted and co-regulated subnetworks can be investigated in an additional window. Here, the composite-FFL motif pattern from our breast cancer study is highlighted.

Interaction Table

The second panel contains an interaction table that gives detailed insight on every interaction in the network. The table shows each interaction between a regulator and its target, the type of the interaction, evidence (predicted or experimentally determined) and the source from which the interaction was obtained. Additionally, we provide a popup link which opens a window showing a list of the diseases related to the regulator and the target, respectively.

ORA Tables

The overrepresentation analysis is available for both, genes and miRNAs.

For miRNAs, TFmiR provides a list of diseases associated with the miRNAs submitted and calcutates a percentage value of how many of the related miRNAs are present in the current network.

The overrepresentation analysis is not done by TFmiR but we incorporate established web services for that purpose. We provide links to GO Term Biological Processes, KEGG Pathways, the Omim Database and David. When a user follows the links, TFmiR submits a list of the genes present in the network to the respective webservice and the browser opens the corresponding website in a new windows.

Motif Table

Once the motif search is finished, the results are stored on the server. To investigate in detail, users may switch from the network overview to the motif table panel. Here, each motif and the participating nodes are shown as well as the confidence scores for the motif in question. A link is provided to investigate the respective co-regulated und co-targeted subnetworks.

Co-regulation and Co-targeted subnetworks

We implemented two approaches to investigate the genes that interact with a motif, as shown in Figure 4.2. Co-targeted subnetworks contain all targets that were identified to be regulated by both the TF and the miRNA in the motif. Co-regulated subnetworks on the other hand incorporate all targets that are regulated by at least one of the former, TF *or* miRNA. An example for a co-regulated subnetwork is shown in our case study in section 4.5, Figure 4.3.

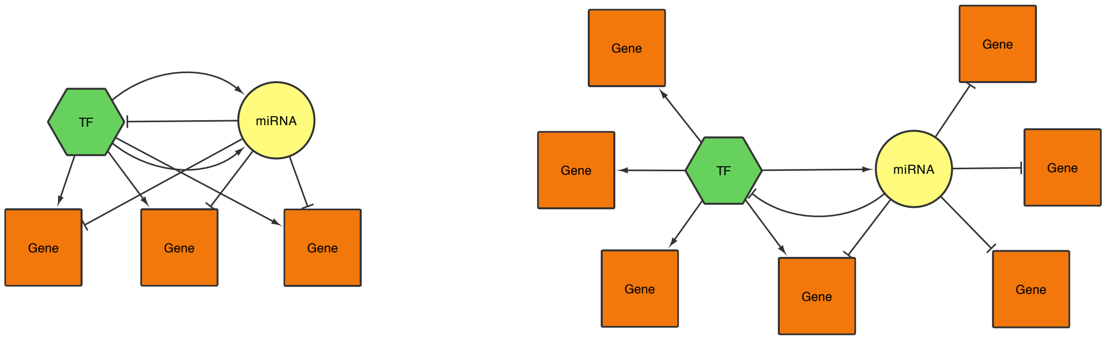


Figure 4.2: Shown are graphical representations of how co-targeted (l) and co-regulated (r) subnetworks are defined.

Functional similarity

For the co-targeted and co-regulated subnetworks, a user may compute a functional similarity score. TFmiR selects a random set of genes of the same size as the gene set in the motif from all GO annotated Entrez genes and computes their similarity scores. After 1 000 permutations, a kolmogorov-smirnov test is used to check if the gene pairs similarity scores are significantly higher than those determined in the randomly selected gene pairs. TFmiR displays a plot of the cumulative scores and the *p*-value obtained from the statistic test, an example is shown in Figure 4.4.

4.4.3 Data management

As TFmiR generates a variety of output during its computations, we thought of a concept that would allow to easily assess the data even outside of the web application. The results of each processing step are stored into a user's session folder incrementally. For example, if interested in the functional similarity plot in a certain motif context, the costly computation is performed once and subsequently stored in the respective session folder.

We implemented a routine that allows to compress a users full session data into a ZIP file and provide a button on the result page download the entire dataset that has been created up to that point.

Sessions

We tried to implement a tradeoff between security and usability of our tool. An intuitive way to track a user without the requirement implement a login process can be realized using a session management system. Each client is assigned its own session identifier which is used to store the uploaded and in the analysis process generated data physically into a directory on the webserver. In order to ensure reasonable capacities, the session information is deleted automatically after two weeks when it has not been accessed.

4.4.4 Cytoscape Plugin

In order to speed up the motif search in our regulatory network, we implemented a Cytoscape plugin according to our algorithm shown in Section 2.2.7 that our webservice is able to connect with using the CyREST plugin for Cytoscape.

In order to do so, and since the vanilla REST plugin only supports basic network operations such and storage functionality, we modified the existing REST plugin to have an additional function that offered the possibility to invoke our motif search plugin within a resident Cytoscape instance running on our server. We defined a path `motifsearch` within the plugin so that it can accept a HTTP POST query that submits a JSON file. The is built from the network files created in our backend and contains the information about each interaction type and the respective nodes involved.

When a user invokes the motif search routine, the network displayed by Cytoscape.js is converted to a JSON and submitted to the Cytoscape REST interface plugin, where the algorithm is automatically started. After the motif search routine finished, the motifs are serialized to JSON and returned as a result to TFmiR.

Cytoscape “headless”

As Cytoscape incorporates a scripting engine and is usually mandatory to be executed within a GUI, we initially thought of invoking our plugin by starting Cytoscape with a script as parameter. However, this is not possible for custom plugins yet. Thus, the easier way was to modify the REST plugin which worked well with the JSON interface Cytoscape.js provides.

In order to adapt Cytoscape to act as a server without a GUI in a “‘headless mode”, we used the `X Window Virtual Framebuffer` (Xvfb) software as a virtual framebuffer that captured the GUI and resolved our issues with Cytoscape.

4.5 Case Study - Breast Cancer

The following case study was taken from the Hamed et al. [2015a] article.

TFmiR was applied to several data sets related to complex diseases such as cancer, Alzheimer and diabetes. In a recent study on breast cancer [Hamed et al., 2015b], we identified 1262 deregulated genes and 121 deregulated miRNAs using gene and miRNA expression data from the TCGA portal (`https://tcga-data.nci.nih.gov/tcga/`). These two sets of deregulated genes and miRNAs are the default sample input files provided by the TFmiR web server. Next, TFmiR was used to reveal the co-regulation network between the deregulated genes/TFs and deregulated miRNAs and to better understand the pathogenic mechanisms associated with breast tumorigenesis. As user input parameters we set the p-value cut off to 0.05, disease was set to breast neoplasms, and the evidence level was set to both experimentally validated and predicted interactions.
For this data set TFmiR constructed a total of 427 regulatory interactions comprising 263 nodes of deregulated miRNAs and deregulated TFs/genes. The breast cancer-specific network involved 345 interactions and 212 nodes of deregulated miRNAs and genes with node and edge coverage ratios (CR) of 80.6%, and 80.8%, respectively. The provided ORA analysis of the disease network nodes revealed their implications in many cancer types as well as cancer-related KEGG pathways. Moreover, ORA analysis of the network miRNAs showed their involvement in cancerogenesis of multiple organs such as lung neoplasms, ovarian cancer, and

adenocarcinoma (Table 4.2). Additionally, TFmiR identified 22 key network players (10 genes and 12 miRNAs) based on the union set of four centrality measures described above (Table 4.3). Interestingly, some of the identified key genes such as BRCA2, ESR1, AKT1, and TP53 were previously implicated and significantly mutated in breast cancer samples [Koboldt et al., 2012]. More importantly, the protein products of the genes ESR1, TP53, TGFB1, AKT1, and BRCA2 are binding targets for anti-breast cancer drugs [Hamed et al., 2015b] (Table 4.4).
Next, we examined the TF-miRNA co-regulatory motifs that were significantly enriched in the entire interaction network. We identified 53 FFL motifs (3 composite-FFLs, 2 TF-FFLs, 6 miRNA-FFLs, and 42 coreg-FFLs). An interesting motif involving the TF SPI1, the miRNA hsa-mir-155, and the target gene FLI1 reveals how FFL motifs may help to better understanding the pathogenicity of breast cancer (Figure 4.3). Recent studies reported that the oncogene SPI1 is involved in tumor progression and metastasis [Guo et al., 2005; Rimmelé et al., 2010]. However, the co-regulation of the oncogene FLI1 [Sakurai et al., 2007] by both SPI1 and the oncomiR hsa-mir-155 was not reported before. As the co-regulated genes of SPI1 and hsa-mir-155 have significantly more similar cellular functions than randomly selected genes (Figure 4.4), this FFL motif provides novel insights on SPI1-miRNA networks alteration in breast cancer and suggests a cooperative functional role between SPI1 and potential miRNA partners.

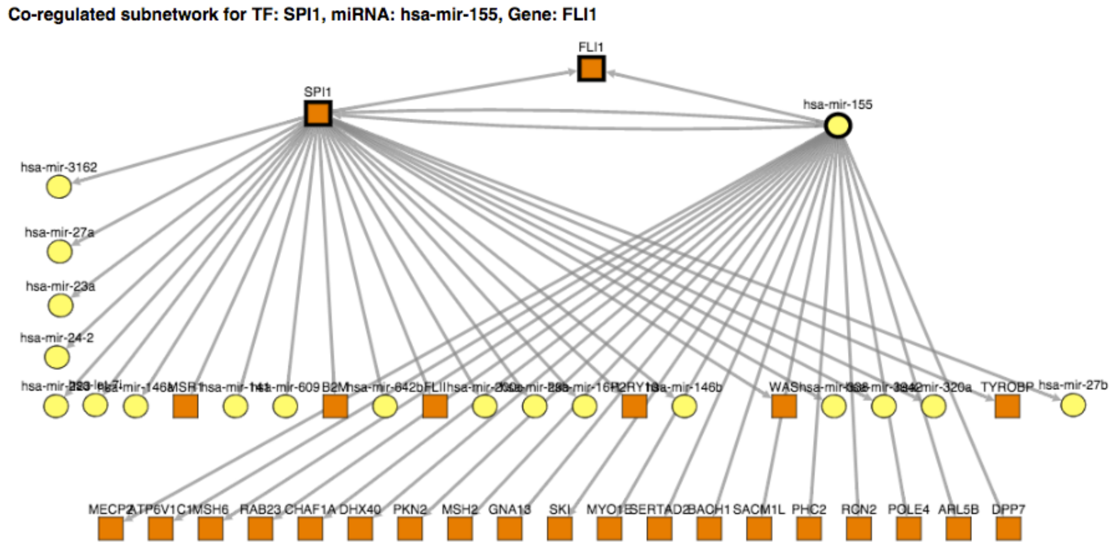


Figure 4.3: A composite FFL motif involves the TF SPI1, the miRNA has-mir-155, and the target gene FLI1. The co-regulated nodes are also visualized and are further tested whether they compose a cooperative functional module in breast cancerogenesis (see Figure 4.3)

Table 4.2: The twelve most significant functions and diseases enriched in the miRNA nodes of the breast cancer disease network

Category	Term	miRNAs Count	p-value
Function	Epithelial-mesenchymal transition	17	0.022
Function	Glucose metabolism	4	0.048
Disease	Breast neoplasms	67	1.43×10^{-25}
Disease	Lung neoplasms	50	4.33×10^{-17}
Disease	Neoplasms	44	3.15×10^{-15}
Disease	Ovarian neoplasms	43	1.3×10^{-14}
Disease	Adenocarcinoma	27	2.59×10^{-13}
Disease	Pancreatic neoplasms	39	7.3×10^{-13}
Disease	Prostatic neoplasms	41	3.49×10^{-12}
Disease	Melanoma	45	1.25×10^{-11}
Disease	Colonic neoplasms	32	4.67×10^{-11}
Disease	Colorectal neoplasms	45	5.69×10^{-11}

Table 4.3: Key driver genes and miRNAs in the breast cancer network

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

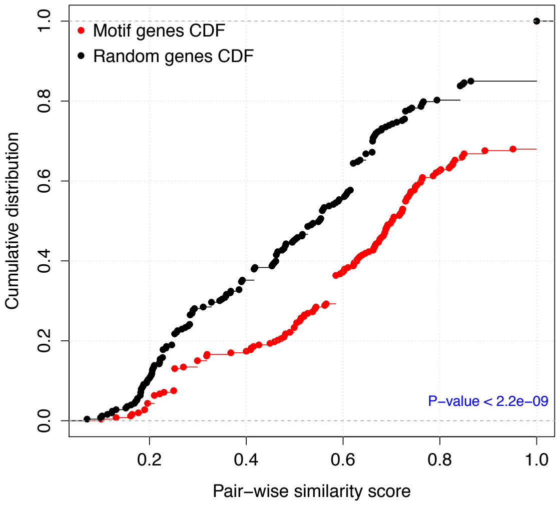


Figure 4.4: Cumulative distribution of GO functional semantic scores of gene pairs of co-regulated genes in the examined motif (red) versus randomly selected genes (black). The p-value was calculated using the Kolmogorov- Smirnov test.

Table 4.4: The identified key gene nodes in the breast cancer network whose protein products are targeted by anti-cancer drugs. (1) means that at least one drug that targets this gene product is reported in this database, and (0) means no drugs are reported for the respective gene in this database. Not included are substances that are known to be cancerogenous or mutagenic.

Target gene	Drug and antineoplastic agents	CTD	PharmGKB	CancerResource
AKT1	U 0126; Tyrphostin AG 1478; Ursodeoxycholic Acid; Valproic Acid; Tyrphostin AG 1024; Trametinib; Tretinoin	1	0	1
BRCA2	Tretinoin; Trichostatin A; Estradiol; Transplatin; Troglitazone; Tunicamycin; Fulvestrant	1	0	1
ESR1	Exemestane; Tamoxifen	0	1	1
TGFB1	Doxorubicin; Fluorouracil; Thalidomide; Entinostat; Hyaluronidase	0	0	1
TP53	4-biphenylmine; Alliin; Apigenin; Atropine; Bicalutamide; Butylidenephthalide	0	0	1

4.6 Discussion & Perspective

Our webservice approach to create and investigate comprehensively a co-regulatory network for diseases was a technical challenge in terms of developing an interface with reasonable responsiveness and performance. As most operations are expensive in execution time, we decided to split the initial network generation from downstream motif search, and this again from functional similarity determination of their subnetworks.

To enable downstream analysis outside TFmiR, for example with Cytoscape, the entire data that is generated by TFmiR at any time during the analysis can be downloaded with a single click. This includes the functional similarity images generated up to that point. The efforts to implement the Cytoscape/CyREST/TFmiR interface were reasonable, the transfer of the motif search to Cytoscape could increase the performance for our example data set by a factor of 10.

In comparison to other webservices with similar functionality that allow to investigate single gene ↔ miRNA interactions, TFmiR offers extended functionality. The tool allows for an investigation of the molecular interactions between sets of deregulated genes and miRNas within and without the pathogenic pathways of a disease with custom evidence levels. The downstream network analysis with co-regulatory motif detection, visualization, ORA analysis, and evaluation of the interaction significance is novel in its functionality.

TFmiR was able to confirm regulators known from literature and revealed new aspects of TF/gene/miRNA interactions in breast cancer pathogenesis. The identified novel hub nodes may be investigated in respect to their druggability.

In future, TFmiR could be extended to include time series expression data. This would enable to investigate how regulatory mechanisms may evolve.

Chapter 5

Alzheimer's Disease: Integrative Differential Analysis of Temporal Cortex Brain Samples

The last project presented in the scope of this thesis is a collaborational project with the group of Prof. Matthias Riemenschneider, head of the neurobiological laboratory and the psychiatric department at the Universitätklinikum des Saarlandes with the aim to study the Alzheimer's disease integratively across diverse datasets. The main wet lab work has been done by Sabrina Pichler and Dr. Gilles Gasparoni, while the author performed the raw data processing and the data integration. Additionally, Mohamed Hamed and Alexander Zapp helped with the key driver identification (section 5.3) and proximity analysis (section 5.4), while Lukas Tost studied putative connections between miRNA expression and the methylome (section 5.5) in his Bachelor thesis that was supervised by the author. The work is in preparation for publication.

5.1 Motivation

In this study, we try to account for the evident complexity of AD by integrating different data collected for individual samples into a single model. The group of Prof. Matthias Riemenschneider obtained and prepared brain samples from 64 individuals from the Munich Brain Center (MUC) and generated an extensive dataset including gene and microRNA expression levels, methylome, single nucleotide polymorphisms and Amyloid-β protein level measurements from temporal lobes.

In the following, we discuss the dataset and describe subsequently our approaches to integrate the different types of data.

5.2 Dataset

The brain samples - further referred to as MUC samples - were collected from 64 individuals with 39 cases and 25 controls, see Table 5.1. The age distribution, which is shown in Figure 5.1, suggested a cohort of late-onset cases.

	Male	Female	Total
Case	15	24	39
Control	15	10	25
	30	34	64

Table 5.1: Number of males and females in the example dataset

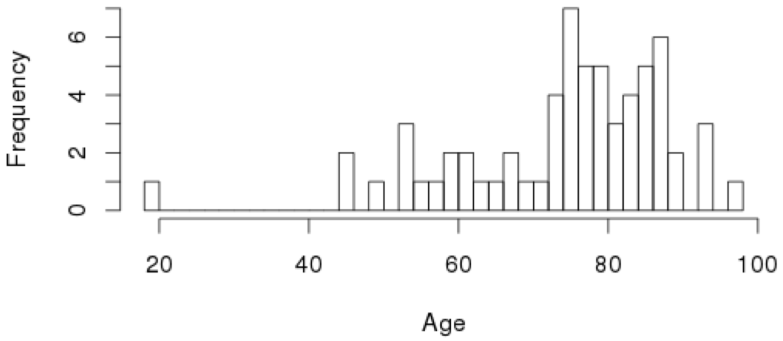


Figure 5.1: Age distribution of the patients in the sample set. The outlier being a control, and EOAD onset expected to be prior to 55-65, we consider our dataset to represent the late onset form of Alzheimer's disease.

In the following sections, the individual datasets suchs as protein levels, miRNA and mRNA expression profiling, and methylome are presented in more detail.

5.2.1 Amyloid-β protein levels

As mentioned before, the formation of neuritic Amyloid-β is considered to be a key pathological feature of AD. Protein levels for both peptides (Amyloid-β40 and 42) were obtained using an quantitative Enzyme Linked Immunosorbent Assay (ELISA).

Evaluation of the assay, log transformation and determination of the extinction values, was done at the laboratory in Homburg. In Section 2.3, We incorporated the protein level measurements as additional input for the SNF method described in section 2.3.

We log transformed the protein levels to validate the presumed case and control fold change. Amyloid-β 40 and 42 showed differential levels between both groups with p-values at 1.052×10^{-8} and 1.984×10^{-8}, respectively. Figure 5.2 shows the distribution of both protein level measurements.

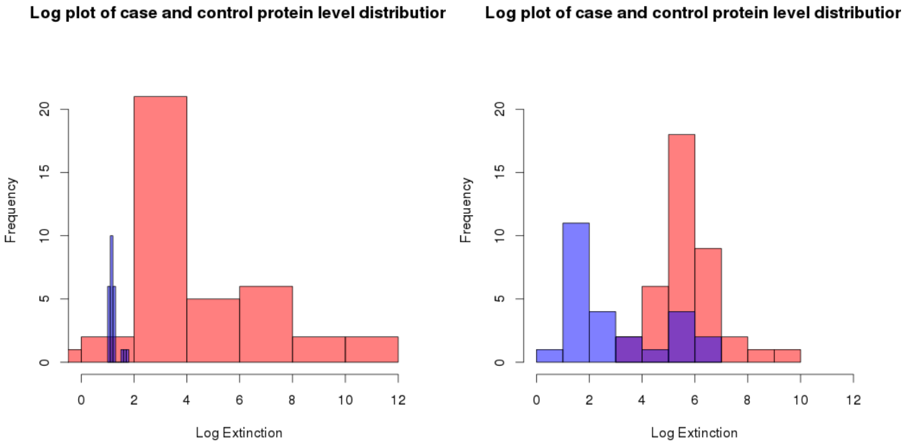


Figure 5.2: Shown are the respective Amyloid-β 40 (l) and 42 (r) level distributions for cases (red) and controls (blue)

5.2.2 miRNA Expression

We analyzed the miRNA-expression profiles in the temporal cortex from the MUC brain samples for both control group and disease group. Quantile normalization and background correction was done in Homburg. The full dataset contained information about the expression of 1281 miRNAs and was filtered several times to identify differentially relevant markers.

Data preparation

The dataset needed to be filtered for noise, since a large fraction of the miRNA markers showed relative expression values < 50, which was established in Homburg as reasonable threshold when working with the Febit Geniom 16 array. Those values were marked as `NA`.

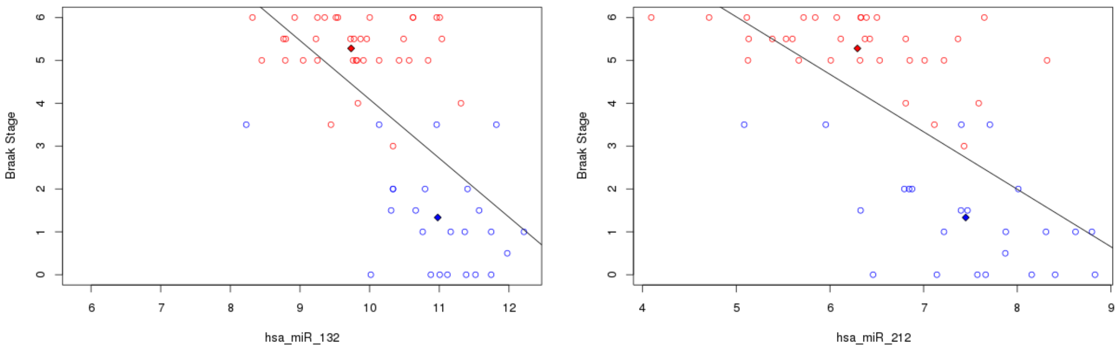


Figure 5.3: miRNA expression plotted against the respective Braak Stage of the samples. Red circles indicate individuals diagnosed with AD, blue circles are control. Filled rhombs indicate median values for respective groups. The most significant candidates, miRNA 132 (left panel) and miRNA 212 (right panel), both show higher expression levels in control than in AD group.

We define a call ratio C as the fraction of samples p for a marker M that satisfy the threshold over all samples for a marker:

$$C_M = \frac{p}{p+n} \tag{5.1}$$

with the markers n that exceed this threshold.

The dataset was split into control and AD group. For both groups, every marker was checked for a call ratio above 90%. If at least one marker in of both groups exceeded that ratio, the marker was accepted.

For both groups, the median for each marker was calculated. If the median did not meet the threshold (50.0) at least one group, the marker was rejected. In the same step, only markers which had a significant higher expression (twofold) in one group (AD or control) were accepted. The filtering left 560 acceptable markers and the remaining dataset was $\log_2$ transformed.

Sample redundancy For some samples, two profiles were measured. Similar to the marker call-ratio, we calculated a sample call-ratio to decide which one should be rejected.

The profile that scored lower was dropped from the analysis, see table 5.2. Call rates were similar except for `AD_F_TP_16`.

Differential Analysis

To find differentially expressed miRNAs, a Student's t-test was applied to the dataset. After correction for multiple testing, two significant miRNAs remained, namely has_miR_132 and hsa_miR_212 as shown in table 5.4.

Table 5.2: Call ratios for the doubled samples

Sample name	Call ratio	Call ratio REP
AD_F_TP_05	0.38	0.39
AD_F_TP_09	0.41	0.40
AD_F_TP_16	0.34	0.39
ctrl_F_TP_05	0.408	0.406

Table 5.3: Student's t-test results with p-Value < 0.01 *before multiple testing correction, see Table 5.4*

miRNA	p-Value
hsa_miR_132	8.2E-07
hsa_miR_212	6.5E-06
hsa_miR_129*	0.00028
hsa_miR_296_5p	0.00032
hsa_miR_129_3p	0.00093
hsa_miR_590_5p	0.0018
hsa_miR_4284	0.0023
hsa_miR_323_3p	0.0027
hsa_miR_148b	0.003
hsa_miR_1207_5p	0.0031
hsa_miR_1274b	0.0035
hsa_miR_543	0.0037
hsa_miR_4298	0.0059
hsa_miR_495	0.0061
hsa_miR_126	0.0063
hsa_miR_129_5p	0.0063
hsa_miR_1289	0.007
hsa_miR_4270	0.0079
hsa_miR_1972	0.0085
hsa_let_7g	0.0087
hsa_miR_874	0.0091
hsa_miR_744*	0.0093
hsa_life_18	0.0094
hsa_miR_296_3p	0.0098

Table 5.4: Significant miRNAs after FDR correction with p-value < 0.05

miRNA	Corrected p-value
hsa_miR_132	0.0024
hsa_miR_212	0.008

For those, the average expression in control groups is higher than in cases, see Figure 5.3.

The heatmap for miRNAs with p-value < 0.01 is shown in Figure 5.4.

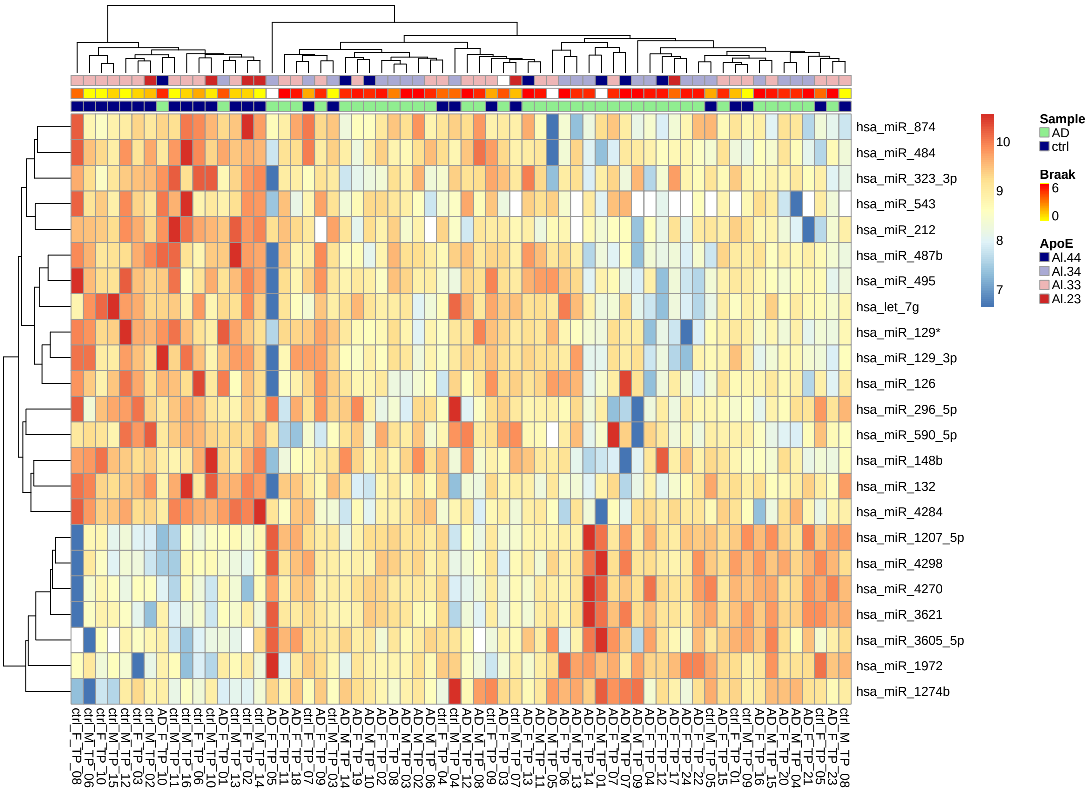


Figure 5.4: Heatmap for the 23 miRNAs with p-value < 0.01 for AD and control group. Note that the cluster of samples to the left indicates all but two Apoϵ4 non-carriers, the exceptions being actual AD cases.

5.2.3 Gene Expression

The gene expression levels were determined using the Illumina HT12 v4 Bead Array. Bioinformatic analysis was carried out with R using the *limma* package [Ritchie et al., 2015].

First, using the negative array control probes, the Illumina data was background corrected and normalized, and $\log_2$ transformed afterwards. The comparison between the raw data and the processing is shown in Figure 5.5.

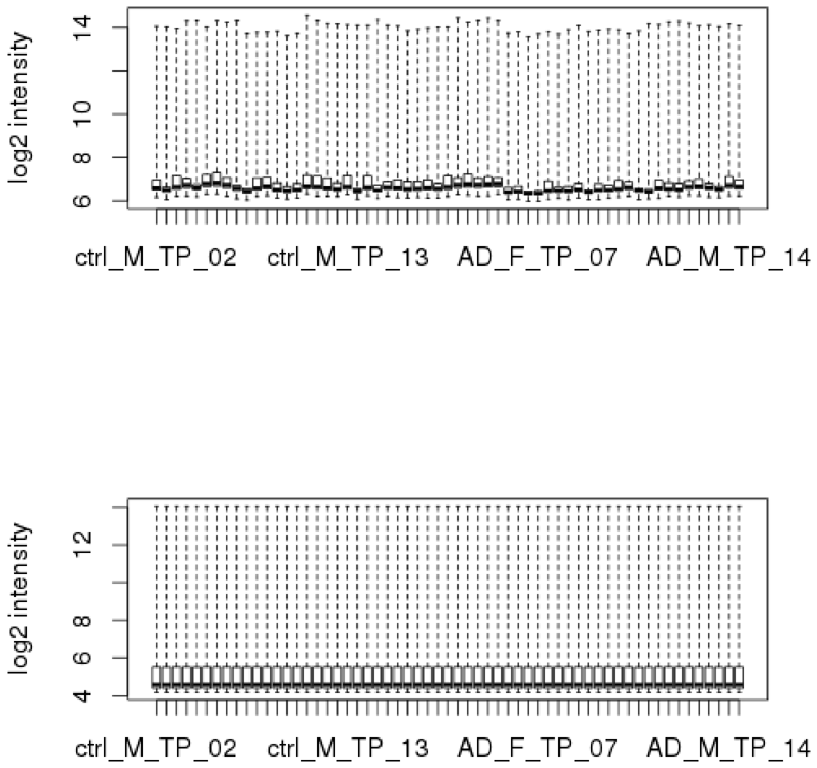


Figure 5.5: $\log_2$ *intensity plot before and after normalization and background correction*

Results

The test gave 1346 underexpressed genes in contrast to 814 overexpressed genes in Alzheimer's disease, see Table 5.5.

Table 5.5: Number of up- and downregulated genes in test set

	AD - ctrl
Downregulated (-1)	1346
Non-decisive (0)	27389
Upregulated (1)	814

We could identify 148 differentially expressed genes between case and control, using an adjusted p-value threshold of 0.01. For a less strict threshold of 0.05, 1918 genes were considered to be significant. Table 5.6 shows the 30 top genes from the analysis (see Appendix A.1 for the full table)

Table 5.6: Table shows an excerpt of the top 30 genes from 148 differentially expressed genes with p-Value < 0.01 using limma. (5 differentially expressed sites could not be annotated)

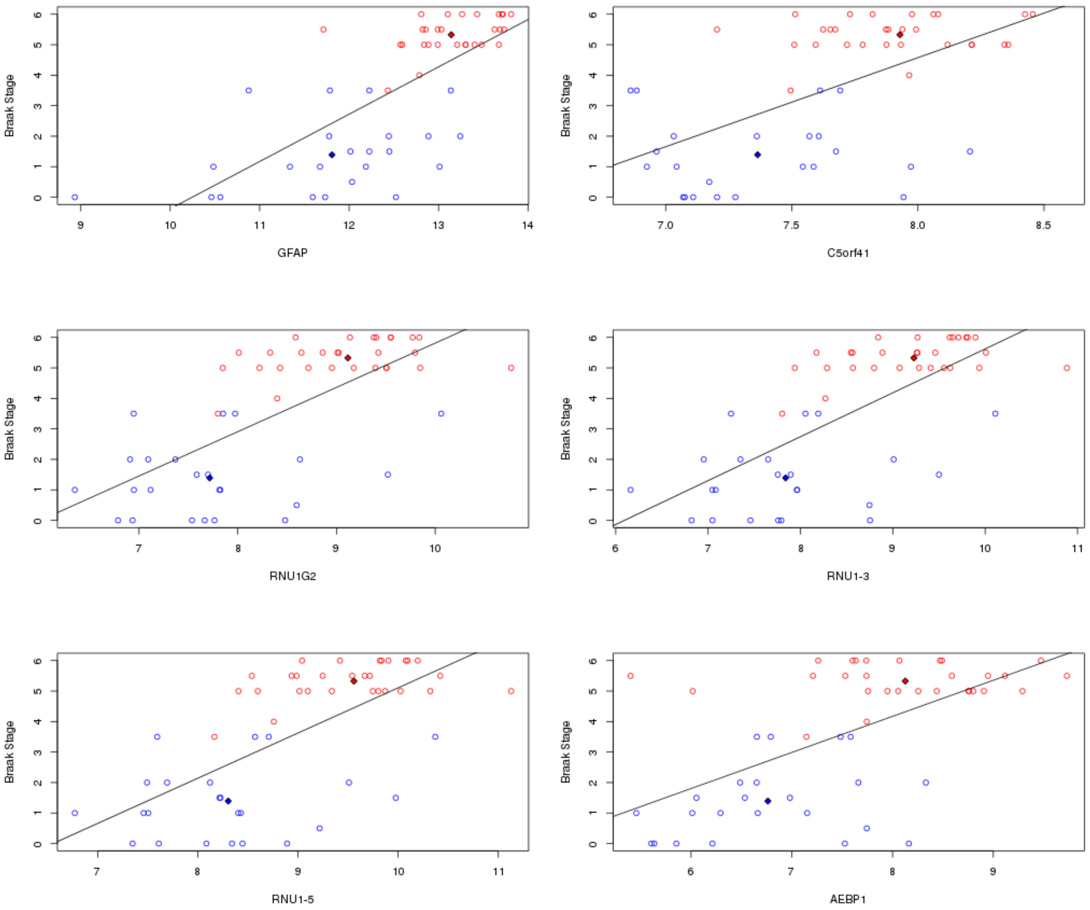
No	Gene Symbol	log FC	t	p-Value	adj. p-Value	B
1	GFAP	1.4	7.1	1.6e-09	4.8e-05	11
2	C5orf41	0.58	6.7	7.6e-09	9e-05	9.8
3	RNU1G2	1.4	6.7	9.1e-09	9e-05	9.7
4	RNU1-3	1.3	6.3	3.5e-08	0.00026	8.4
5	RNU1-5	1.2	6.1	7.8e-08	0.00044	7.7
6	AEBP1	1.5	6.1	8.8e-08	0.00044	7.6
7	HBP1	0.59	5.9	1.7e-07	0.00064	7
8	C1orf110	1.2	5.9	1.8e-07	0.00064	6.9
9	RHOQ	0.77	5.9	2e-07	0.00064	6.9
10	PPARBP	0.51	5.8	2.8e-07	0.00076	6.5
11	MYBPC1	1.1	5.8	2.8e-07	0.00076	6.5
12	EIF3E	0.38	5.7	3.4e-07	0.00079	6.4
13	NFKB1	0.49	5.7	3.5e-07	0.00079	6.4
14	TNPO1	0.56	5.7	4.2e-07	0.00089	6.2
15	C5orf41	0.57	5.5	8.6e-07	0.0017	5.5

...

Table 5.6 – *Continued*

	Gene Symbol	log FC	t	p-Value	adj. p-Value	B
16	PKN2	0.62	5.4	1.3e-06	0.0024	5.1
17	AK1	0.53	5.4	1.4e-06	0.0024	5.1
18	MAP1LC3A	-0.51	-5.3	1.5e-06	0.0024	5
19	ITPKB	1.2	5.3	1.6e-06	0.0024	5
20	SRGAP1	0.78	5.2	2.2e-06	0.003	4.7
21	KCNF1	-0.97	-5.2	2.9e-06	0.0038	4.4
22	SST	-1.5	-5.2	3e-06	0.0038	4.4
23	PPEF1	-0.99	-5.1	3.1e-06	0.0038	4.4
24	DDR2	0.96	5.1	3.8e-06	0.0045	4.2
25	LOC284988	0.74	5.1	4e-06	0.0045	4.1
26	HBQ1	-0.85	-5.1	4.2e-06	0.0046	4.1
27	FOXJ1	1.3	5	4.7e-06	0.0048	4
28	FAM107A	0.64	5	4.9e-06	0.0048	4
29	LOC388481	-0.49	-5	5e-06	0.0048	3.9
30	KCNA4	-0.28	-5	5.1e-06	0.0048	3.9

For the top 12 genes, Figure 5.6 shows the expression values for cases and controls and the corresponding Braak stages that indicate the progression of the neurodegenerative process.



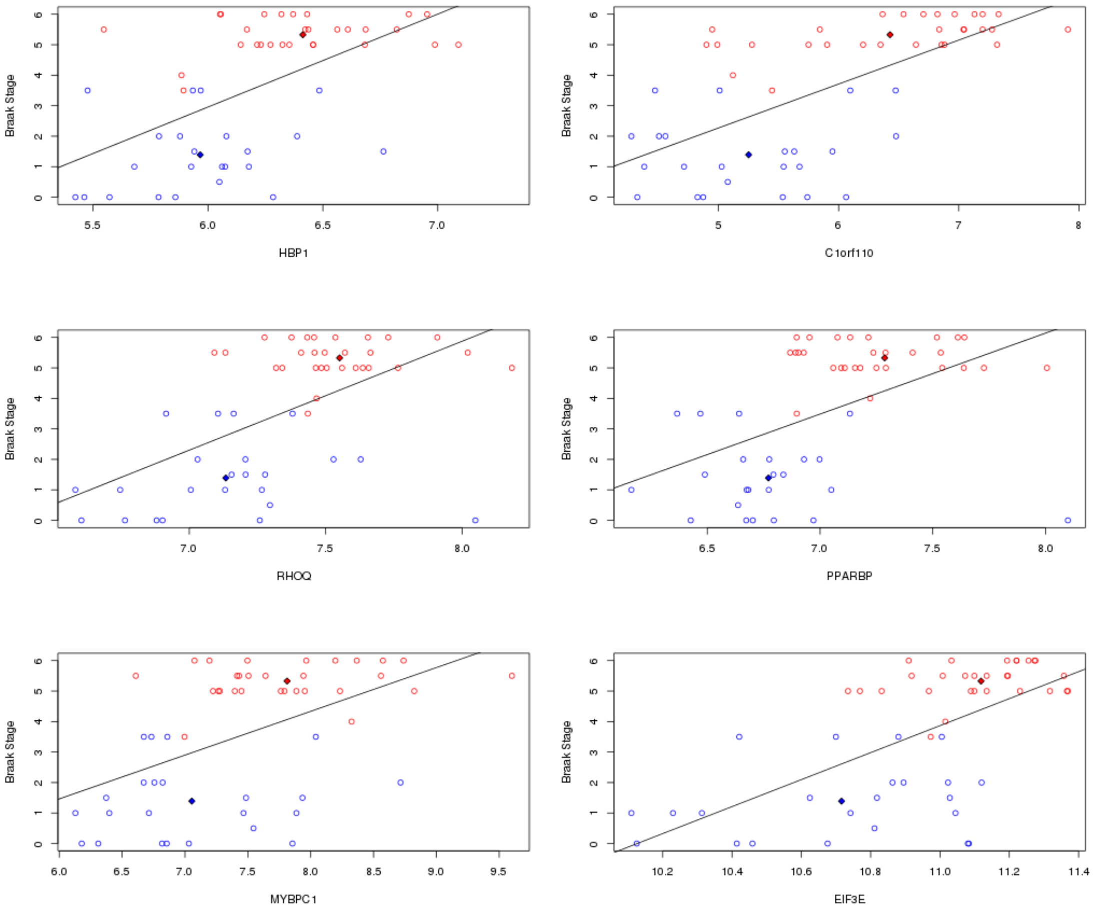
(a) Genes 1-6

Figure 5.6: Figures 5.6a and 5.6b show the expression values and their corresponding Braak stage of the top 12 genes as indicator for AD progression.

The most prevalent genes for early-onset Alzheimer's disease, such as *APP*, *PSEN1* and *PSEN2* show either non-significant p-values within the expression data or dropped out of analysis due to FDR correction (*PSEN2*). Since the mean age of the targets studied is 74.5 and Alzheimer's disease shows a mean progression of 10 years (ranging from 1 to 25 in some cases), we consider the studied individuals as late-onset alzheimer's disease cases (LOAD) for which those genes may not be determinative.

The tau protein coding gene *MAPT* shows significantly higher expression in the case group, with an adjusted p-value of 0.0328.

A comparison of the identified 1918 genes (1291 downregulated, 845 upregulated) shows an overlap of 307 (24.8%) of 1271 in reference downregulated genes



(b) Genes 7-12

Figure 5.6: (cont.)

and 218 (12.9%) of 1691 in reference upregulated genes, see [Blalock et al., 2004], Figure 5.7 or *Molecular Signatures Database*, where the data is published.

Figure 5.8 shows the overlap between the different external datasets and the data retrieved with limma.

Also, there were significant overlaps with gene sets that have been associated to various disorders such as bipolar disorder.

A heatmap of the 148 top genes with known gene names is shown in Figure 5.9.

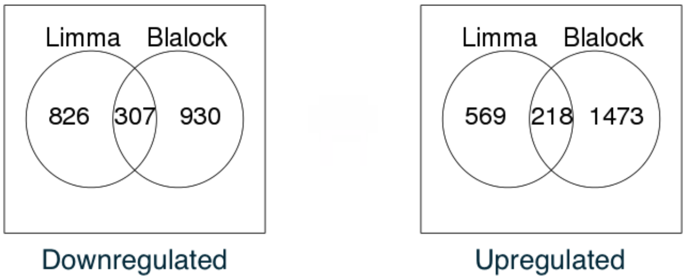


Figure 5.7: Overlaps of the up- and downregulated genes in our study with those of Blalock et al. [2004].

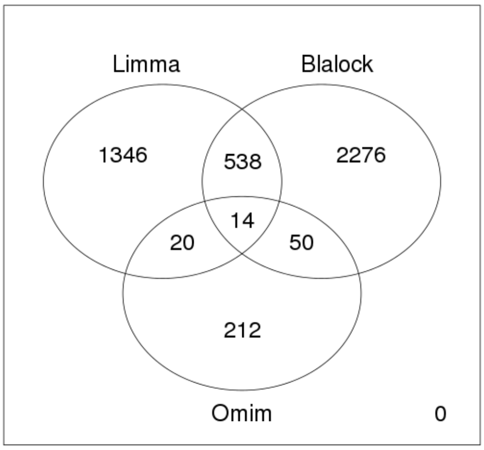


Figure 5.8: Diagram shows the overlap of the results retrieved from Limma analysis (with p-Value < 0.05*) and the genes associated with Alzheimer from Omim and a dataset published on MSigDB, released by Blalock et al. [2004], who worked with a threshold of* 0.05 *as well.*

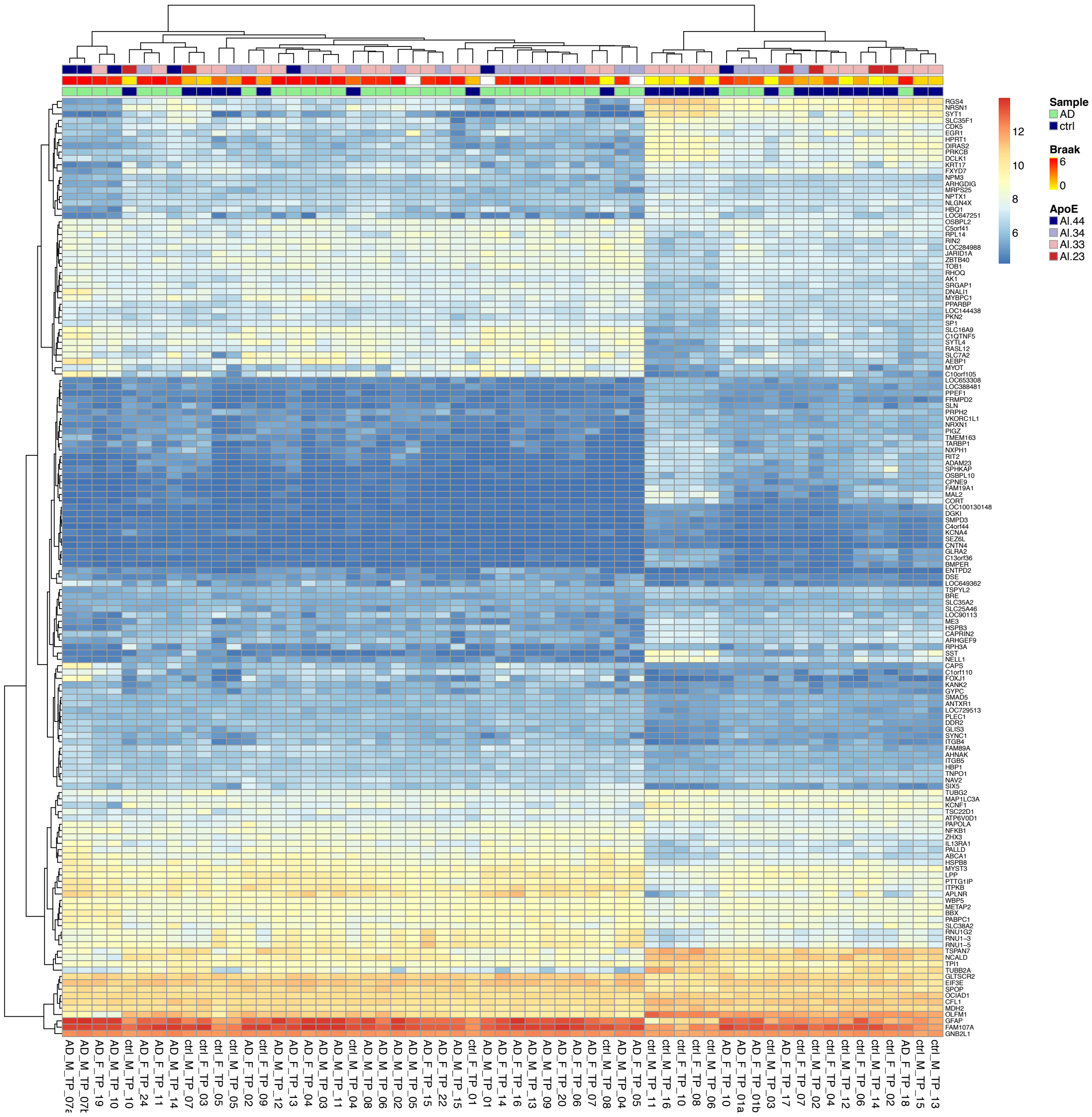


Figure 5.9: Heatmap for the top 148 differentially expressed genes. A separation between case and control group can be seen. Even more interesting, "misclassified" individuals (due to expression profiles) show the respective Apoϵ genotypes that are considered to have a protective or promoting function for Alzheimer's disease, e.g. ctrl_M_TP_10 and ctrl_M_TP_07 with an 23 genotype, or the AD patients AD_F_TP_10 and AD_F_TP_01 with expression profiles similar to control groups but with at least one Apoϵ 4 allele. On the other hand, information for clear separation is still incomplete, since patient AD_F_TP_17 shows an expression similar to controls as well as the Apoϵ 23-genotype but still were diagnosed with Alzheimer's.

5.2.4 Methylation

The epigenetic methylation data were generated using an Illumina 450k BeadChip Array. In the following, the data has been analyzed using the *methylumi* R package [Davis et al., 2012].

First, we preprocessed the data following the protocol as described in the manual provided with methylumi and subsequently identified differentially methylated regions (DMRs) using the *methyAnalysis* R-package [Du and Bourgon, 2013].

The comparisons of the density between both channels for both methylated and unmethylated probes after (1) color balance adjustment, (2) background correction and (3) quantile normalization are listed in Appendix A.2. In Figure 5.10 the comparison before and after preprocessing is shown.

Then, we identified the DMRs using a smoothing window of 250 base pairs. The thresholds were set at a p-value of 0.05 and a required fold difference of 0.1. The fold difference threshold was selected to account for the fact that within tissues, methylation changes are more subtle in same type tissues than in more heterogeneous samples, such as in cancer and controls for example. Finally, the DMRs were annotated and filtered for promoter regions 2kb base pairs upstream.

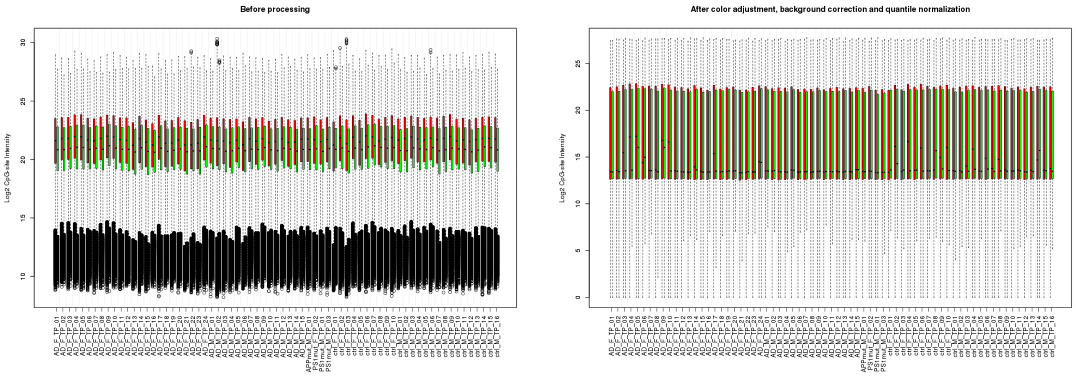


Figure 5.10: Box plot of color bias before and after processing of the methylation data

Results

We obtained a set of 5543 differentially methylated regions between cases and controls, whereof 1021 lie within promoter regions (2kb upstream). The top 50 differentially methylated regions – sorted by the adjusted p-value – are listed in Table 5.7.

Table 5.7: *Top 50 differentially methylated regions, sorted by p-Value*

Rank	Chr	Start	End	Width	min adj.*p*-val	Mean AD	Mean Ctrl	EntrezID	Symbol	TSS Dist.
1	chr14	64010427	64010844	418	0.0038	-4.7	-4.6	5529	PPP2R5E	-348
2	chr20	16710288	16710841	554	0.0039	-4.3	-4.1	6629	SNRPB2	0
3	chr10	85954312	85954622	311	0.0039	-5.6	-5.4	92211	CDHR1	0
4	chr17	79268988	79269242	255	0.0046	-3.8	-3.7	124565	SLC38A10	0
5	chr1	21112556	21113199	644	0.0046	-4.6	-4.5	50809	HP1BP3	0
6	chr17	79670206	79670933	728	0.0052	-5.6	-5.5	1468	SLC25A10	0
7	chr11	809697	810321	625	0.0052	-5.3	-5.1	6181	RPLP2	0
8	chr20	23470274	23470461	188	0.0052	3.5	3.8	10047	CST8	-1305
9	chr16	85060822	85061228	407	0.007	-3.8	-3.6	9764	KIAA0513	-147
10	chr2	10861364	10861710	347	0.007	-4.8	-4.5	245973	ATP6V1C2	-65
11	chr15	42840397	42841312	916	0.0079	-5	-4.9	255252	LRRC57	0
12	chr16	11343164	11343701	538	0.0079	5.5	5.3	116028	RMI2	0
13	chr1	11332969	11333577	609	0.0082	-4.5	-4.3	29914	UBIAD1	0
14	chr4	205522	205522	1	0.0086	0.99	0.68	642280	ZNF876P	-867
15	chr10	81838583	81839181	599	0.0088	-4.6	-4.5	219347	TMEM254-AS1	0
16	chr18	77711560	77712060	501	0.0088	-5.1	-4.9	80148	PQLC1	0
17	chr2	9697298	9697298	1	0.0088	2.3	1.8	6868	ADAM17	-1381
18	chr8	121137288	121137734	447	0.0094	-4.1	-3.9	7373	COL14A1	0
19	chr9	120177435	120177700	266	0.01	-3.4	-3.7	23245	ASTN2	-118
20	chr1	9149026	9149026	1	0.01	0.8	0.36	6518	SLC2A5	-516
21	chr2	17720407	17720407	1	0.011	-5.7	-5	7447	VSNL1	-1400
22	chr13	20420028	20434104	14077	0.011	-0.2	-0.089	9205	ZMYM5	0
23	chr21	27542749	27543410	662	0.011	-5.3	-5.2	351	APP	0
24	chr5	148786302	148786377	76	0.011	1.1	0.89	728264	MIR143HG	-63
25	chr5	134094363	134094672	310	0.011	-5.3	-5.3	9879	DDX46	0
26	chr1	113257872	113258477	606	0.011	-4.9	-4.7	333926	PPM1J	0
27	chr16	30569726	30569978	253	0.011	-5.1	-4.9	92595	ZNF764	-84
28	chr10	99609481	99611477	1997	0.011	-5.2	-4.9	401647	GOLGA7B	0
29	chr1	147400503	147400867	365	0.011	-5.4	-5.2	51463	GPR89B	0
30	chr6	53659348	53659618	271	0.011	-3.7	-3.5	55227	LRRC1	0
31	chr12	120427178	120427182	5	0.011	-3.7	-3.6	92558	CCDC64	-466
32	chr22	33453994	33454444	451	0.011	-3.9	-3.7	8224	SYN3	0
33	chr17	39240343	39240343	1	0.011	1.7	1.2	100132476	KRTAP4-7	-116
34	chr4	39699140	39699959	820	0.011	-4.7	-4.5	3093	UBE2K	0
35	chr2	73341598	73341598	1	0.011	0.74	0.48	26056	RAB11FIP5	-1452
36	chr15	90808551	90809150	600	0.011	-5.1	-5	10519	CIB1	0
37	chr7	129592730	129593010	281	0.012	-3.8	-3.6	7328	UBE2H	0
38	chr15	83316640	83316838	199	0.013	-5.6	-5.3	64506	CPEB1	0
39	chr15	64454827	64455548	722	0.013	-4.5	-4.3	5479	PPIB	0
40	chr17	79633496	79633857	362	0.013	-5.7	-5.6	339229	OXLD1	0
41	chr7	33102336	33102794	459	0.013	-5.2	-5	51251	NT5C3A	0
42	chr7	66057363	66057450	88	0.013	-4.8	-4.5	493754	LOC493754	0
43	chr4	126236316	126236816	501	0.013	-3.3	-3.1	79633	FAT4	-751
44	chr2	111435449	111435943	495	0.013	-4.1	-4	699	BUB1	0
45	chr11	71493933	71497688	3756	0.013	-4.3	-4.4	55199	FAM86C1	-869
46	chr21	35014635	35015322	688	0.013	-4.8	-4.5	6453	ITSN1	0
47	chr9	88357129	88357129	1	0.013	-3.3	-3	23287	AGTPBP1	-185
48	chr8	75233610	75233613	4	0.014	-6.3	-5.7	56704	JPH1	-48
49	chr6	170124933	170125241	309	0.014	-5.4	-5.3	55274	PHF10	-827
50	chr8	145133263	145133899	637	0.014	-5.2	-5.1	54512	EXOSC4	0

5.3 Application of the key driver identification pipeline

We used the set of differentially expressed miRNAs, mRNAs and differentially methylated regions to apply the computational pipeline described in Sections 3.4 and 4.5. To widen the search space, we extended the microRNAs to the set before multiple test correction. The identified key drivers are shown in Table 5.8.

Notably, one would expect mir-212 to be a key driver because it has been widely associated with the Alzheimer's Disease [Wong et al., 2013]. When investigating the resulting TF-miRNA network shown in Figure 5.11, this is a result from using the MCDS algorithm to the identify the key drivers. When looking for a minimum set, mir-132 is a more contributing node than mir-212 and thus favorably selected.

We could identify the cyclin-dependent kinase 5 gene (CDK5) as a key driver gene, which has as well been shown in other studies to be associated with pre-neurofibrillary tangles and neurofibrillary tangles in temporal lobes from AD brain samples [Hernandez et al., 2009; Shukla et al., 2012].

Table 5.8: Table shows the results for the key drivers identified from differentially expressed miRNAs and mRNAs. Among other miRNAs we identified mir-132 as a key driver, but not mir-212.

Interactions	Type	Count	Key Drivers	Total KD	Top GO	Top KEGG
TF-mRNA	Genes	148	TSPAN7, JARID1A, AHNAK, ITPKB, NELL1, PPARBP, KRT17, BBX, C10orf105, CDK5, HBP1, CPNE9, C13orf36, RASL12, SLC16A9	15	cell-matrix adhesion, multicellular organismal development,intracellular signaling cascade	Inositol phosphate metabolism, Axon guidance, Alzheimer's disease (CDK5)
miRNA-miRNA	Genes	59	GFAP, ITGB5, RIN2, TPI1	4		
	miRNAs	15	mir-148b, mir-323-3p, let-7g ,mir-126, mir-129*, mir-132, mir-129- 3p, mir-484, mir-296-5p, mir-590-5p,mir-543	11	Hormones Regulation, hESC regulation, inflammation	Autistic Disorder, Stomach Neoplasms

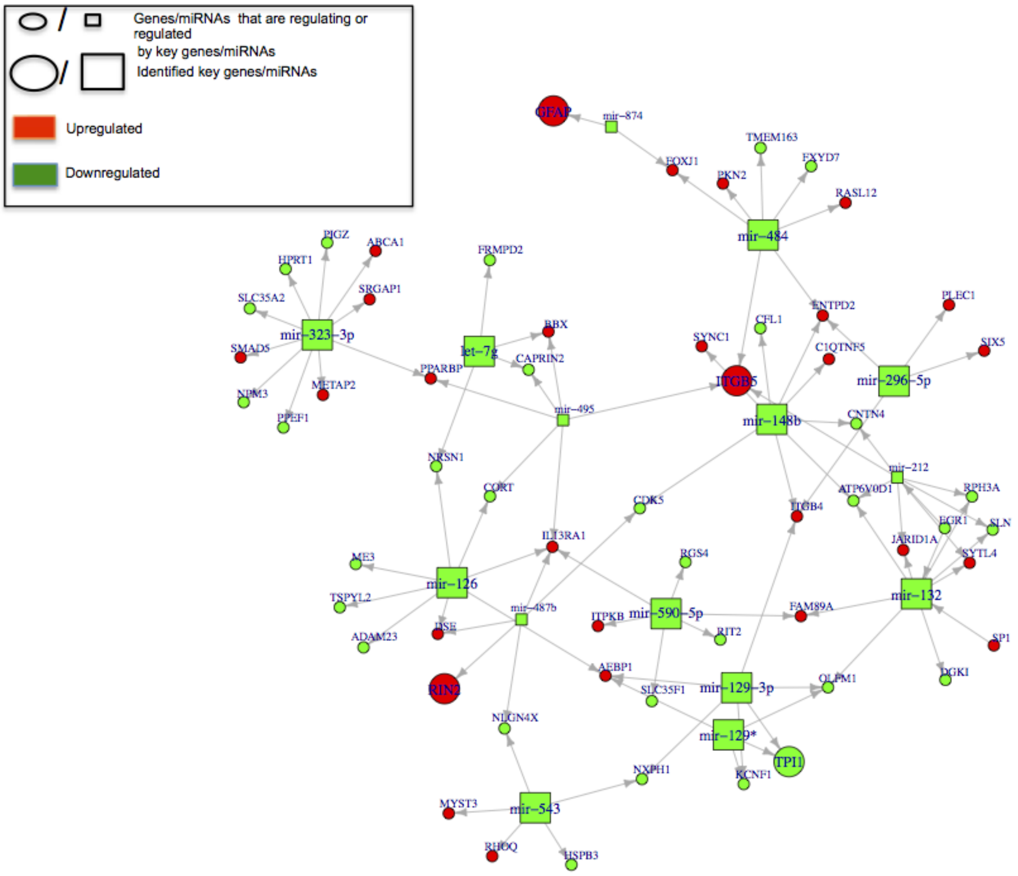


Figure 5.11: Transcription factor-mRNA network (created with igraph*)*

5.4 Proximity Analysis

Using the set of differentially expressed miRNAs and gene methylation, we examined SNPs that are adjacent to their respective coding regions.

5.4.1 microRNA proximity

Figure 5.12 shows the result from the proximity analysis of miRNAs and their corresponding SNPs.

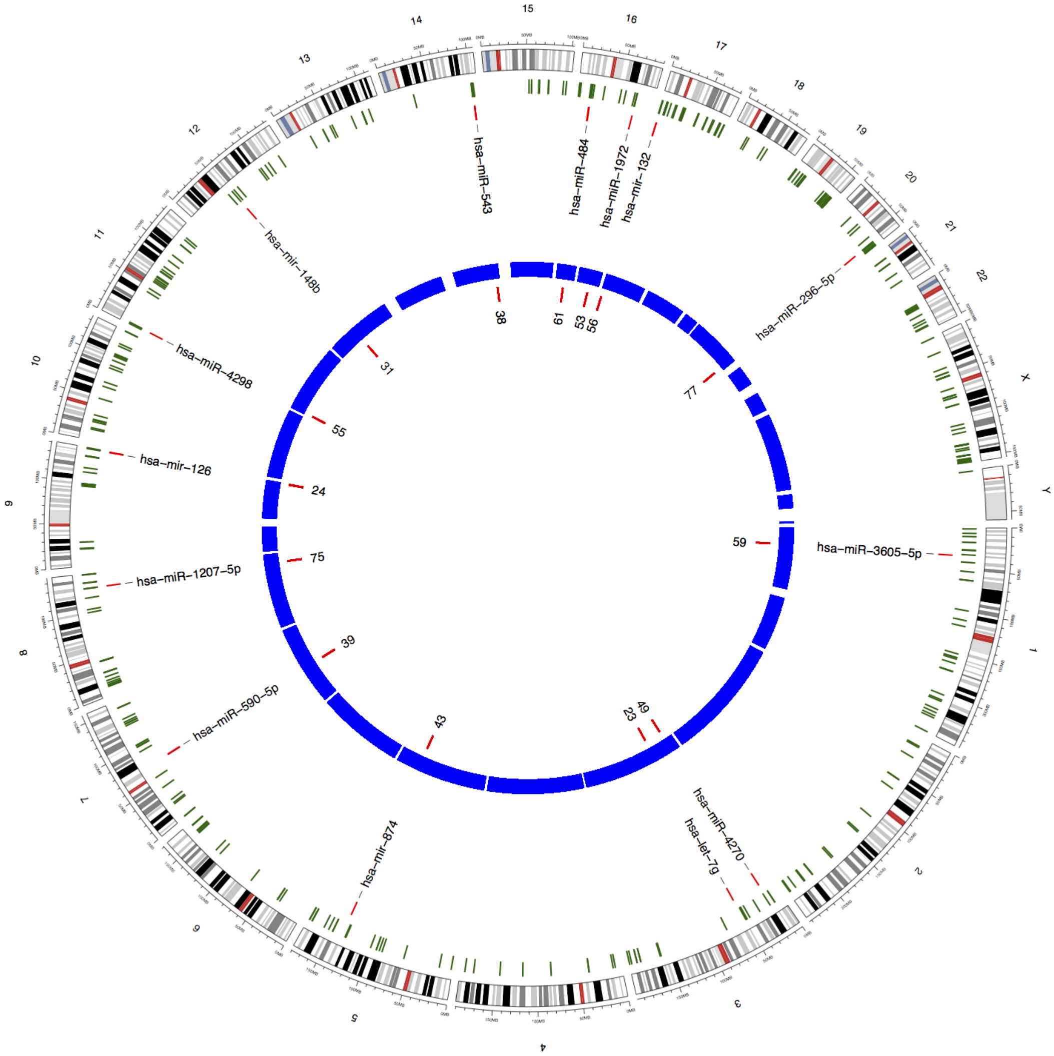


Figure 5.12: Proximity analysis of micro RNAs with known SNPs. Green are the known micro RNAs, blue known SNPs. Red are differentially expressed microRNAs as shown in section 5.2.2.

5.4.2 Methylation proximity

For the differentially methylated regions, we identified four genes and their respective SNPs in our dataset. We examined the results with DAVID and in the UCSC Genome Browser. A graphical overview of the findings is shown in Figure 5.13.

For TMEM254-AS1 Antisense RNA Gene, COL14A (involved with collagen binding) and CST8, as well as genes in their genomic context displayed no connection to AD or neurogenerative diseases is in general known so far.

UbiA prenyltransferase domain containing protein 1 (UBIAD) is associated with Schnyder corneal dystrophy, an autosomal dominantly inherited disease that affects the cornea and causes opacification. Other than increasing severity of the disease with aging patients, the connection to AD is questionable.

This analysis suggests that SNPs near methylated regions in our dataset are unlikely to influence the development of AD.

5.5 Epigenetic analysis of miRNA

In the scope of a Bachelor thesis, Lukas Tost worked on a project in which we aimed for an investigation of the effect of miRNA promoter methylation in the MUC samples and implemented an R package to do so. We combined the previously obtained methylation data, miRNA and mRNA expression profiles using various database sources. Based on the differentially methylated transcription start sites (TSSs), we built a network for the respective miRNAs and their targets, shown in Figure 5.14. We analyzed the respective targets using DAVID [Dennis et al., 2003], see Table 5.9.

The four genes EPHA4, CBLN4, ZNF148 and KAT2B were previously described in literature to be correlated with AD.

EPHA4 is a substrate of γ-secretase, and the γ-secretase-cleaved EPHA4 intracellular domain (EICD) is known to enhance the formation of dendritic spines via activation of the Rac signaling pathway [Matsui et al., 2012]. Rac1 levels are correlated with the amount of EICD in frontal lobes and negatively correlated with the amount of tau phosphorylation, which indicates involvment in the synaptic pathogenesis of AD.

Significant expression decrease of the gene coding for synaptic protein cerebellin 4 (CBLN4) was shown to be connected to amyloid-β induced cell death, as *in vitro* experiments with increased recombinant CBLN4 showed preserving effects in cultivated neuron cells Chacón et al. [2015].

ZBPF belongs to the family of zinc binding protein factors which were described by Augustin et al. [2011] to be part of a module of transcription factor families that AD-related genes share.

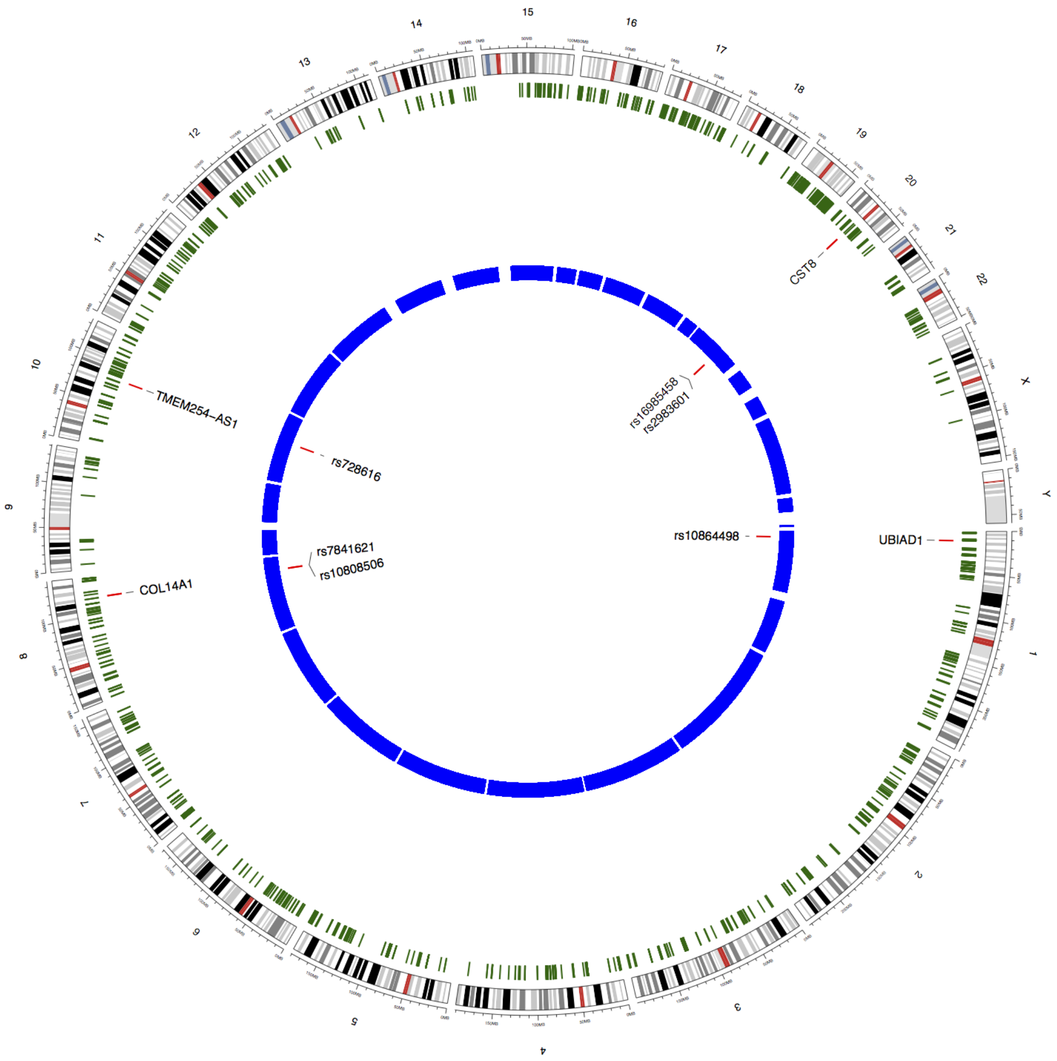


Figure 5.13: Proximity analysis of gene methylation with known SNPs

The gene encoding the lysine acetyltransferase 2B (KAT2B), also known as P300/CBP-associated factor (PCAF), was shown to regulate the expression of proteins involved in amyloid-β generation and degradation [Duclot et al., 2010], especially previously mentioned γ-secretase.

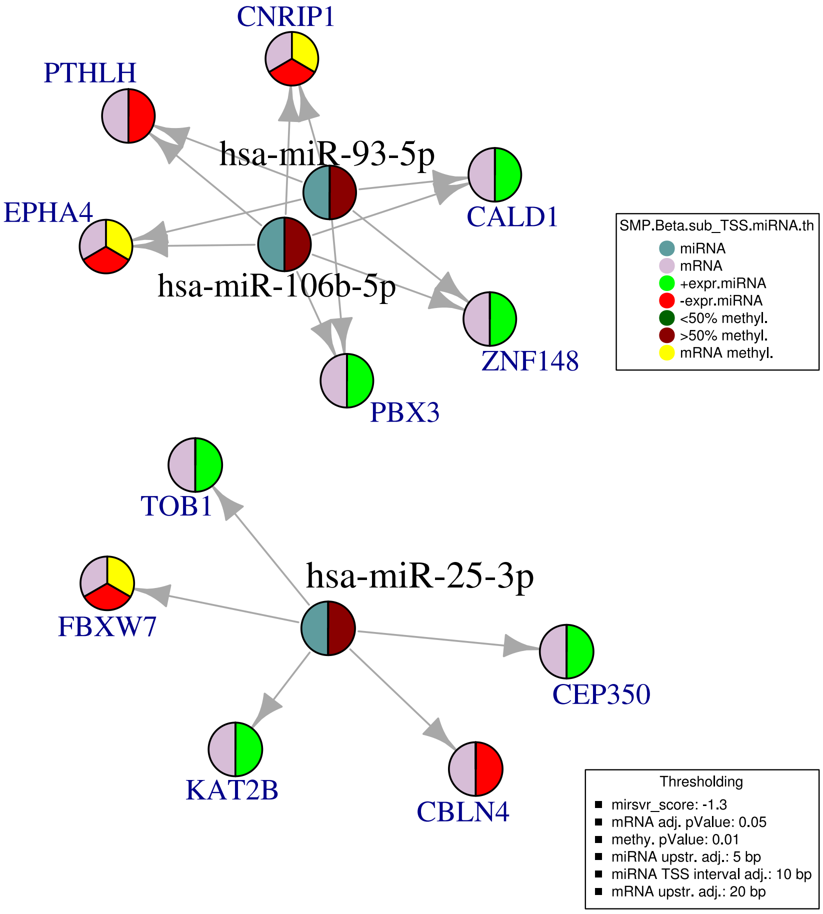


Figure 5.14: Shown is the network for differentially methylated miRNAs and their gene targets. The node filling indicates methylation and expression levels, respectively.

5.6 Integration with Similar Network Fusion

While the key driver analysis is a method to generally assess expression and methylome data, the MUC dataset enabled us to specifically compare the profiles for each sample. Thus, in order to assess the datasets sample-wise, we applied the similar network fusion (SNF) method presented in Chapter 2.3. Since the dataset is incomplete for some samples, the SNF study was limited to an intersection of 56 samples for which the whole dataset was available.

We used the lists of the differential analysis for each dataset as an input in order to increase the significance of the output network and reduce noise by non-decisive probes. That way, we tried to highlight possible decisive correlations between the

Table 5.9: Functional annotation for the putative miRNA regulated gene targets

Gene Symbol	Gene Name	Function
EPHA4	receptor A4	Axon guidance
CBLN4	cerebellin 4 precursor	Putative involvement in synaptic function in central nervous system
ZNF148	zinc finger protein 148	negative regulation of transcription from RNA polymerase II promoter
KAT2B	K(lysine) acetyltransferase 2B	chromatin structure and dynamics / transcription
CNRIP1	cannabinoid receptor interaction protein 1	modulates the constitutive CB1 receptor activity in the central nervous system
FBXW7	F-box and WD repeat domain containing 7	Ubiquitin mediated proteolysis
PTHLH	parathyroid hormone-like hormone	skeletal system development
CEP350	centrosomal protein 350kDa	Cell division and chromosome partitioning
PBX3	pre-B-cell leukemia homeobox 3	regulation of respiratory gaseous exchange by neurological system process
CALD1	caldesmon 1	muscle system process
TOB1	transducer of ERBB2, 1	Anti-proliferative protein

samples spanning over the different datatypes that enable a more precise prediction outcome whether or not to classify a sample to be case or control.

Thus, the final input for SNF contained the following datasets with continously measured data of 2 miRNAs, 148 mRNAs, and 119 differentially methylated sites, including non-promoter regions.

Moreover, from the protein level measurement (Amyloid-β 40/42) we extracted the Apoϵ genotype (presumed relevance shown in 5.9 and created a matrix with discrete entries 0, 1, and 2, each representing the count of the respective allele Apoϵ 2, 3, and 4:

Based on the clustering shown in Figure 5.9, we expect a stronger influence of the Apoϵ genotype, for that reason a different measure for this feature has

Table 5.10: Example of the Apoϵ genotype table

Sample	Apoϵ 2	Apoϵ 2	Apoϵ 3
AD_F_TP_01	0	1	1
AD_F_TP_06	0	1	1
AD_F_TP_07	0	2	0
AD_F_TP_10	0	0	2
⋮	⋮	⋮	⋮

been included. We derived the probabilities for each genotype from Raber et al. [2004] and created a vector containing the allele, diagnosis, and the corresponding probability, see Table 5.11

Table 5.11: Apoϵ probability table [Raber et al., 2004]

Genotype	Population[a] (%)	AD[b] (%)	#Population	#AD	Risk[c] (%)	If all US
ε2/2	1	0.1	0.5M	0.004M	0.08	0.4M
ε2/3	12	4	5.5M	0.18M	3.2	1.5M
ε3/3	60	35	27.6M	1.4M	5.1	2.3M
ε3/4	21	42	9.6M	1.7M	18	8.2M
ε4/4	2	16	0.9M	0.6M	67	30.7M

Please note that ε2/ε4 subjects are not included in table.

[a] Using estimate of 46 million in US over 60 y/o in 2000.

[b] Assuming 4 million individuals have AD.

[c] Data from [13,46,49].

5.6.1 Results

We could reach an Mutual Information score of 0.35 and 84% accuaracy using these four datasets.

The scores for the resulting network scores using different combinations of input networks are shown in Table 5.12. The clusters for each individual dataset and the final clustering of the fused network is shown in Figure 5.15.

Interestingly, when we included the Amyloid-β ratios as an additional marker into the network, the accuracy of the method suffered largely. Thus, for this study, we focused on expression, methylome and the Apoϵ genotype.

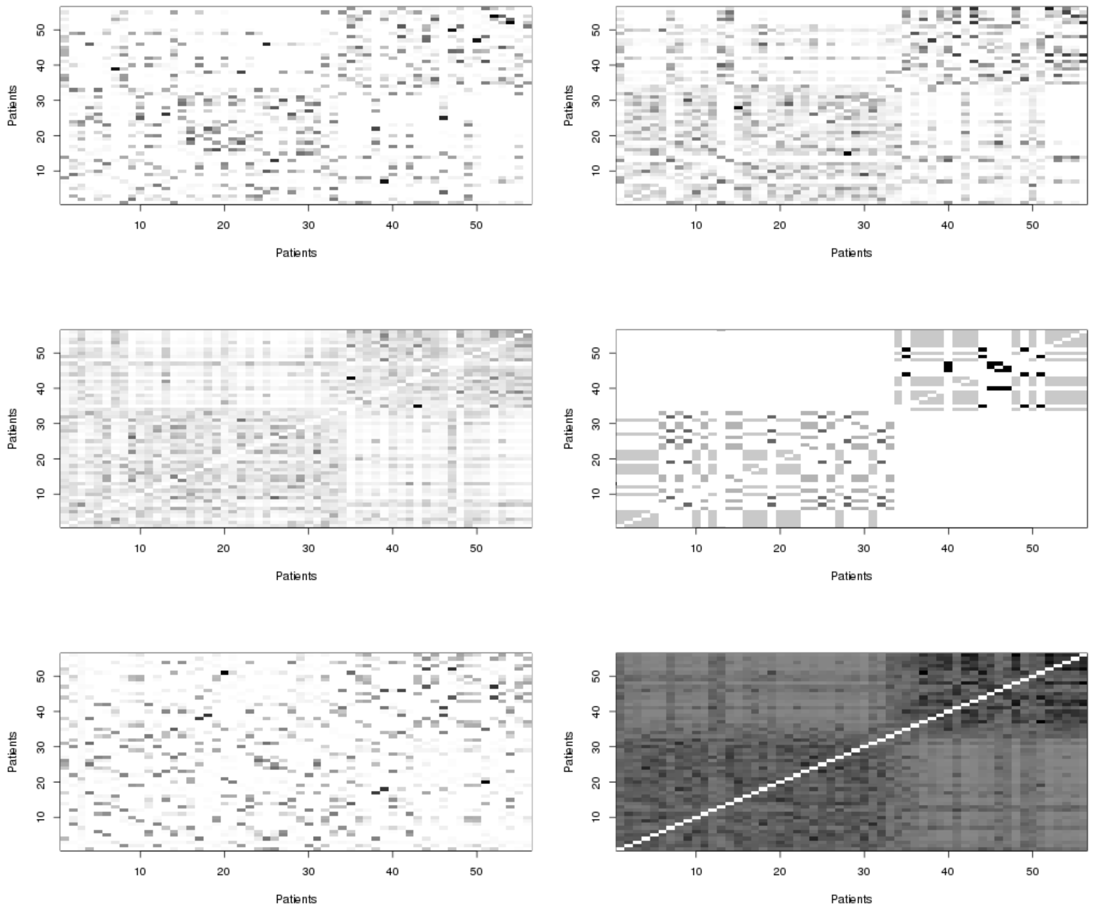


Figure 5.15: From top left to top right: clusters for each dataset, (1) miRNA, (2) mRNA, (3) methylation, (4) ApoE genotype, (5) Amyloid-β ratios and (6) the fused network

5.7 Discussion

In the following section, we assess the results from the (1) data analysis, (2) key driver identification, (3) microRNA promoter methylation analysis and (4) the similar network fusion and possible strategies to improve the approaches.

5.7.1 MUC Dataset

Subject of this study is a cohort of 64 samples from 39 Alzheimer cases and 25 controls. Although the size of the data suffices to show clear trends between both groups, presumably the study gains statistical power if the tested population could be increased. However, comprehensive datasets that include the entire set of input

Table 5.12: Scoring of different combinations for Similar Network Fusion

	Specifity (TPR)	Sensivity (TNR)	Accuracy	NMI
miRNA/mRNA/Methylation/ApoE	0.83	0.85	0.84	0.36
mRNA/Methylation	0.70	0.91	0.82	0.32
miRNA/mRNA/ApoE	0.83	0.82	0.82	0.32
miRNA/mRNA	0.78	0.85	0.82	0.31
miRNA/Methylation	0.70	0.91	0.82	0.32
miRNA/mRNA/Methylation	0.70	0.88	0.80	0.27
miRNA/Methylation/ApoE	0.87	0.73	0.79	0.27
mRNA/Methylation/ApoE	0.91	0.64	0.75	0.25
mRNA/ApoE	0.83	0.64	0.71	0.16
miRNA/ApoE	0.83	0.64	0.71	0.16
Methylation/ApoE	0.83	0.64	0.71	0.16

data that was generated for this project are not available publicly and, thus, hard to study at a large scale.

Moreover, the post-mortem timespan of the tissues before preservation differs largely, see Table 5.13. This leads to different degradation states within the samples and may distort the results of a differential analysis.

Table 5.13: Post-mortem preservation times for MUC samples

Time	< 12h	12-24h	24-48h	> 48h	NA
	6	17	16	5	20

Methylation Noise

Major concerns when investigating the methylation data are related to the nature of the brain tissue, methylation heterogeneity between different tissues, and the putative cells affected in AD. Epigenetic modifications are known differ largely for different tissues in an individual [Ma et al., 2014], while the samples contain as well neurons as neuroglia cells. Because AD is associated with neuronal degeneration, the measured methylome is expected to contain noise in respect to how many cells of one or the other type have actually been measured for each sample at hand. A possible remedy could be to identify specific markers for neuronal epigenetic modifications and glia to determine the assumable fraction of neurons and correct the signals.

5.7.2 Early stage samples

When investigating Alzheimer's disease, it has to be considered that the pathology develops in a comparatively large time frame. Other than cancer, late-onset AD shows early symptoms that *may* indicate a case, but are diagnosed years later. This uncertainty impedes the need for a dataset that is representative for samples in early stages of AD. A possible remedy, and not limited to research on AD, could be a project to collect biological data from individuals spanning several decades. In our particular interest, if we were able to track changes in the metabolism of patients that were later shown to suffer from AD, it would be highly interesting to search for significant differences between earlier "snapshots" of AD patients and healthy population. Such studies would allow to investigate the origins of the disease rather than late-stage effects, as factors such as amyloid-β plaques are observed in many diseases involved in neurodegeneration [Hardy and Selkoe, 2002].

5.7.3 microRNA Promoter Methylation

While we found genes related to AD to be targeted by miRNAs with differentially methylated promoters, the expression levels did not meet the assumption of direct methylation → miRNA expression → gene expression relationship. Beside the issues regarding the methylation data (see section 5.7.1), miRNA-mediated regulation is subjected to other possible factors that influence their targeting efficiency that were not accounted for, such as ribosome interference, RNA-binding proteins or unpredictable effects with overlapping target sites [Pasquinelli, 2012]. For this reason, more sophisticated approaches, e.g. incorporating protein levels, to estimate post-translational effects may enhance the understanding of the underlying mechanisms.

5.7.4 Similarity Network Fusion

The similarity network fusion was previously used to incorporate TCGA data. Cancer is considered a heterogeneous disease and descriptive markers for cases and controls are usually separated clearly in a differential analysis [Hamfjord et al., 2011; Hibbs et al., 2004; Jerónimo et al., 2004; Melnikov et al., 2009]. In comparison, we expected the investigation of late-onset Alzheimer's disease to pose a challenge because there are several markers known to be risk factors yet not conclusive predictors. We think a combination of each may enhance the discrimination between cases and controls, but probably some regulatory mechanism yet to be understood can not be modeled with the statistical model behind the patient-similarity network fusion.

Input for the method

When Wang et al. [2014] published their method, whole datasets from TCGA without differential analysis were used. Although the tumor samples could be stratified for different cancer types, our approach was different as we wanted to distinguish combined markers between affected and unaffected individuals with a high accuracy. The pipeline for the SNF analysis can be easily adapted for other/-more input data or pose different questions. Other approaches to implement the method could include (1) using the whole datasets without differential anaysis and (2) stratify unaffected and affected samples each in respect to different features, such as we did for the Apoϵ 4 allele where we compared carriers/non-carriers to study a possible relation between the Apoϵ genotype and the markers used in our patient-similarity profiles (which did not yield satisfactory results so far).

Chapter 6

Summary and Discussion

In this thesis, the author presented our approach to address data integration in systems biology.

First, we introduced Mebitoo, a software framework for data integration. In the beginning designed to handle with sequence data, we evolved the application to work with arbitrary data and introduced an extension to work with gene regulatory networks. We extended the regulatory network in a study on breast cancer where we used methylation data and miRNA expression profiles to prune and enrich the network, respectively.

However, the concept of Mebitoo to be a desktop application independent of execution environment limits the software in a few ways. While the database storage system is able to manage several gigabytes of size with ease, unprocessed systemic data exceed the feasible dimensions easily. For example, a methylome or genome wide association studies generate raw data that reach terabytes depending on the study size, up to several gigabytes for each sample.

On the other hand, strengths of Mebitoo are the concise plugin design that benefits from the rich client platform underneath and eases the development of plugins as shown in several Bachelor and Master projects. While less flexible than interpreter based environments, a graphical user interface enhances usability for users not familiar with command shells. A task manager allows to define and queue processes and enables to generate workflows, thus Mebitoo excels when applying a custom set of various methods to datasets of moderate size.

We presented a pipeline for generating regulatory transcription factor-miRNA networks and generalized the approach for public use with TFmiR. As we incorporate various databases and R packages, we chose a web server architecture and designed a contemporary web application. We extended the functionality with a search algorithm for regulatory motifs and provide downstream analysis with their interacting co-regulated and co-targeted subnetworks. In comparison to other services, TFmiR is distinguished by the ability to investigate comprehensively the interaction between all participating genes and miRNAs, the motif detection functionality, and the ORA analysis of the generated TF-miRNA regulatory networks. Continuing our studies on breast cancer, TFmiR was able to confirm regulators from literature and hint possible new aspects of TF-miRNA-gene interactions and to characterize co-regulatory motifs that form functional modules in breast oncogenesis.

In final study presented in this thesis, the author analyzed various data from a cohort of Alzheimer's disease patients and respective controls. We carried out a differential analysis on each and applied the previously shown pipelines to identify key drivers, putative drug targets, and modeled a regulatory network based on the dataset. As the data was specific for each individual, we applied a patient-similarity network based method and a network fusion algorithm to rule out distinctive features. The first results showed that the complementary information of the differential analysis enhanced the capability to discriminate between AD patients and control groups.

6.1 Outlook

The aim of systems biology to model living systems in its entirety remains a demanding task. Integrating experimental und theoretical techniques has been tackled by many approaches and yet, the complexity of regulatory systems still remains to be fully understood.

Our Mebitoo framework is suitable to implement various methods with small effort. The platform ensures efficiency with the advantage of compiled bytecode instead of scripting languages and easy usability due to a graphical user interface. This allows the embedding of pipelines that up to today are frequently done manually by biologists, such as a PCR primer identification on sequenced data, followed by a BLAT mapping to identify regions, annotate the corresponding genes, and subsequent investigation of the gene neighborhood.

We intend to continue development of TFmiR. In order to investigate how regulatory mechanisms may evolve, we think of a concept to elucidate multi-case expression data to model a time series. Moreover, an elucidation of the networks in respect to cellular processes such as stem cell differentiation in addition to diseases could yield insights in the underlying regulatory mechanism.

The most recent work on Alzheimer's disease leaves some open questions, such as the disputable impact of amyloid-β levels on the similar network fusion method. The author intends to incorporate larger datasets that are in preparation into the study to refine the marker selection, additionally integrating the SNP dataset that was used for the proximity analysis so far.

6.2 Closing Remarks

We learned that data integration in bioinformatics, while making large progresses, remains a relevant topic to unravel systemic relationships and explore possible regulatory mechanisms in complex organisms and their diseases. Further improvements of biotechnology and the methods to assess systemic data may reveal specific connections on a molecular level. Guided by this, researchers may strategically scrutinize the natural causality of biological processes.

Bibliography

Augustin, R, Lichtenthaler, S. F, Greeff, M, Hansen, J, Wurst, W, and Trümbach, D.
Bioinformatics Identification of Modules of Transcription Factor Binding Sites in Alzheimer's Disease-Related Genes by In Silico Promoter Analysis and Microarrays.
International Journal of Alzheimer's Disease, 2011(2):1–13, 2011.

Barabási, A.-L.
Network Medicine — From Obesity to the "Diseasome".
New England Journal of Medicine, 357(4):404–407, 2007.

Benjamini, Y and Hochberg, Y.
Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.
Journal of the Royal Statistical Society Series B-Methodological, 57(1):289–300, 1995.

Bird, A. P.
CpG-rich islands and the function of DNA methylation.
Nature, 321(6067):209–213, 1986.

Blalock, E. M, Geddes, J. W, Chen, K. C, Porter, N. M, Markesbery, W. R, and Landfield, P. W.
Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses.
Proceedings of the National Academy of Sciences of the United States of America, 101(7):2173–2178, 2004.

Bloch, J.
Effective Java : programming language guide.
Boston : Addison-Wesley, 2001.

Boerno, S. T, Grimm, C, Lehrach, H, and Schweiger, M.-R.

Next-generation sequencing technologies for DNA methylation analyses in cancer genomics.
Epigenomics, 2(2):199–207, 2010.

Cancer Genome Atlas Research Network.
Comprehensive genomic characterization defines human glioblastoma genes and core pathways.
Nature, 455(7216):1061–1068, 2008.

Cava, C, Bertoli, G, Ripamonti, M, Mauri, G, Zoppis, I, Rosa, P. A. D, Gilardi, M. C, and Castiglioni, I.
Integration of mRNA Expression Profile, Copy Number Alterations, and microRNA Expression Levels in Breast Cancer to Improve Grade Definition.
Plos One, 9(5):e97681, 2014.

Chacón, P. J, del Marco, Á, Arévalo, Á, Domínguez-Giménez, P, García-Segura, L. M, and Rodríguez-Tébar, A.
Cerebellin 4, a synaptic protein, enhances inhibitory activity and resistance of neurons to amyloid-β toxicity.
Neurobiology of aging, 36(2):1057–1071, 2015.

Christakis, N. A and Fowler, J. H.
The Spread of Obesity in a Large Social Network over 32 Years.
New England Journal of Medicine, 357(4):370–379, 2007.

Chuang, H.-Y, Hofree, M, and Ideker, T.
A Decade of Systems Biology.
dx.doi.org, 26(1):721–744, 2010.

Cock, P. J. A, Antao, T, Chang, J. T, Chapman, B. A, Cox, C. J, Dalke, A, Friedberg, I, Hamelryck, T, Kauff, F, Wilczynski, B, and de Hoon, M. J. L.
Biopython: freely available Python tools for computational molecular biology and bioinformatics.
Bioinformatics (Oxford, England), 25(11):1422–1423, 2009.

Craig, J.
Complex Diseases: Research and Applications .
Nature Education, 2008.

Csardi, G and Nepusz, T.
The igraph software package for complex network research.
InterJournal, Complex Systems, 1695(5):1–9, 2006.

Davignon, J, Gregg, R. E, and Sing, C. F.
Apolipoprotein E polymorphism and atherosclerosis.
Arteriosclerosis (Dallas, Tex.), 8(1):1–21, 1988.

Davis, S, Du, P, Bilke, S, Triche Jr, T, and Bootwalla, M.
methylumi: Handle Illumina methylation data.
R package version, 2(0), 2012.

Davison, C, Macintyre, S, and Smith, G. D.
The potential social impact of predictive genetic testing for susceptibility to common chronic diseases: a review and proposed research agenda.
Sociology of Health & Illness, 16(3):340–371, 1994.

Dayhoff, M. O.
Computer analysis of protein sequences.
In *Computers in Life Science Research*, pages 9–14. Springer US, Boston, MA, 1974.

Dayhoff, M. O, Barker, W. C, Schwartz, R. M, Orcutt, B. C, and Hunt, L. T.
Data base for protein sequences.
In *AFIPS '76: Proceedings of the June 7-10, 1976, national computer conference and exposition.* ACM Request Permissions, 1976.

Dayhoff, M. O.
Atlas of Protein Sequence and Structure.
National Biomedical Research Foundation, 1965.

Dayhoff, M. O.
Computer Analysis of Protein Evolution.
Scientific American, 221(1):86–95, 1969.

Dennis, G, Sherman, B. T, Hosack, D. A, Yang, J, Gao, W, Lane, H. C, and Lempicki, R. A.
DAVID: Database for Annotation, Visualization, and Integrated Discovery.
Genome Biology, 4(5):P3, 2003.

Du, P and Bourgon, R.
methyAnalysis: an R package for DNA methylation data analysis and visualization.
R package version, 2013.

Duclot, F, Meffre, J, Jacquet, C, Gongora, C, and Maurice, T.
Mice knock out for the histone acetyltransferase p300/CREB binding protein-associated factor develop a resistance to amyloid toxicity.

Neuroscience, 167(3):850–863, 2010.

Erdős, P and Rényi, A.
On random graphs.
Publicationes Mathematicae Debrecen, 6:290–297, 1959.

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

Goecks, J, Nekrutenko, A, Taylor, J, and lastName, a. f. a.
Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.
Genome Biology, 11(8):R86, 2010.

Griffith, O. L, Montgomery, S. B, Bernier, B, Chu, B, Kasaian, K, Aerts, S, Mahony, S, Sleumer, M. C, Bilenky, M, Haeussler, M, Griffith, M, Gallo, S. M, Giardine, B, Hooghe, B, Van Loo, P, Blanco, E, Ticoll, A, Lithwick, S, Portales-Casamar, E, Donaldson, I. J, Robertson, G, Wadelius, C, De Bleser, P, Vlieghe, D, Halfon, M. S, Wasserman, W, Hardison, R, Bergman, C. M, Jones, S. J. M, and Open Regulatory Annotation Consortium.
ORegAnno: an open-access community-driven resource for regulatory annotation.
Nucleic Acids Research, 36(Database issue):D107–13, 2008.

Gunderson, K. L, Kruglyak, S, Graige, M. S, Garcia, F, Kermani, B. G, Zhao, C, Che, D, Dickinson, T, Wickham, E, Bierle, J, Doucet, D, Milewski, M, Yang, R, Siegmund, C, Haas, J, Zhou, L, Oliphant, A, Fan, J.-B, Barnard, S, and Chee, M. S.
Decoding Randomly Ordered DNA Arrays.
Genome research, 14(5):870–877, 2004.

Guo, Z. S, Naik, A, O'Malley, M. E, Popovic, P, Demarco, R, Hu, Y, Yin, X, Yang, S, Zeh, H. J, Moss, B, Lotze, M. T, and Bartlett, D. L.
The enhanced tumor selectivity of an oncolytic vaccinia lacking the host range and antiapoptosis genes SPI-1 and SPI-2.
Cancer Research, 65(21):9991–9998, 2005.

Hamed, M, Spaniol, C, Nazarieh, M, and Helms, V.

TFmiR: a web server for constructing and analyzing disease-specific transcription factor and miRNA co-regulatory networks.
Nucleic Acids Research, 2015a.

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

Hamfjord, J, Stangeland, A. M, Hughes, T, Skrede, M. L, Tveit, K. M, Ikdahl, T, and Kure, E. H.
Differential expression of miRNAs in colorectal cancer: comparison of paired tumor tissue and adjacent normal mucosa using high-throughput sequencing.
Plos One, 7(4):e34150–e34150, 2011.

Hardy, J and Selkoe, D. J.
The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics.
Science (New York, NY), 297(5580):353–356, 2002.

Hayes, W, Sun, K, and Przulj, N.
Graphlet-based measures are suitable for biological network comparison.
Journal of Gerontology, 29(4):483–491, 2013.

He, J. S, Cai, Z, Ji, S, Beyah, R, and Pan, Y.
A Genetic Algorithm for Constructing a Reliable MCDS in Probabilistic Wireless Networks.
In *Wireless Algorithms, Systems, and Applications*, pages 96–107. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

Hecker, M, Lambeck, S, Toepfer, S, van Someren, E, and Guthke, R.
Gene regulatory network inference: Data integration in dynamic models—A review.
Biosystems, 96(1):86–103, 2009.

Hernandez, P, Lee, G, Sjoberg, M, and Maccioni, R. B.
Tau phosphorylation by cdk5 and Fyn in response to amyloid peptide Abeta (25-35): involvement of lipid rafts.
Journal of Alzheimer's disease : JAD, 16(1):149–156, 2009.

Hibbs, K, Skubitz, K. M, Pambuccian, S. E, Casey, R. C, Burleson, K. M, Oegema, T. R, Jr, Thiele, J. J, Grindle, S. M, Bliss, R. L, and Skubitz, A. P. N.
Differential Gene Expression in Ovarian Carcinoma : Identification of Potential Biomarkers.

The American Journal of Pathology, 165(2):397–414, 2004.

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

Hu, Z, Chang, Y.-C, Wang, Y, Huang, C.-L, Liu, Y, Tian, F, Granger, B, and Delisi, C.
VisANT 4.0: Integrative network platform to connect genes, drugs, diseases and therapies.
Nucleic Acids Research, 41(Web Server issue):W225–31, 2013.

Huber, W, Carey, V. J, Gentleman, R, Anders, S, Carlson, M, Carvalho, B. S, Bravo, H. C, Davis, S, Gatto, L, Girke, T, Gottardo, R, Hahne, F, Hansen, K. D, Irizarry, R. A, Lawrence, M, Love, M. I, MacDonald, J, Obenchain, V, Oleś, A. K, Pagès, H, Reyes, A, Shannon, P, Smyth, G. K, Tenenbaum, D, Waldron, L, and Morgan, M.
Orchestrating high-throughput genomic analysis with Bioconductor.
Nature Methods, 12(2):115–121, 2015.

Huerta, A. M, Salgado, H, Thieffry, D, and Collado-Vides, J.
RegulonDB: a database on transcriptional regulation in Escherichia coli.
Nucleic Acids Research, 26(1):55–59, 1998.

Jerónimo, C, Henrique, R, Hoque, M. O, Mambo, E, Ribeiro, F. R, Varzim, G, Oliveira, J, Teixeira, M. R, Lopes, C, and Sidransky, D.
A quantitative promoter methylation profile of prostate cancer.
Clinical Cancer Research, 10(24):8472–8478, 2004.

Jiang, C, Xuan, Z, Zhao, F, and Zhang, M. Q.
TRED: a transcriptional regulatory element database, new entries and other development.
Nucleic Acids Research, 35(Database issue):D137–D140, 2006.

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

Johnson, V. E, Stewart, W, and Smith, D. H.

Traumatic brain injury and amyloid-β pathology: a link to Alzheimer's disease? *Nature Reviews: Neuroscience*, 11(5):361–370, 2010.

Kearse, M, Moir, R, Wilson, A, Stones-Havas, S, Cheung, M, Sturrock, S, Buxton, S, Cooper, A, Markowitz, S, Duran, C, Thierer, T, Ashton, B, Meintjes, P, and Drummond, A.
Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data.
Bioinformatics, 28(12):1647–1649, 2012.

Koboldt, D. C, Fulton, R. S, McLellan, M. D, Schmidt, H, Kalicki-Veizer, J, McMichael, J. F, Fulton, L. L, Dooling, D. J, Ding, L, Mardis, E. R, Ally, A, Balasundaram, M, Butterfield, Y. S. N, Carlsen, R, Carter, C, Chu, A, Chuah, E, Chun, H.-J. E, Coope, R. J. N, Dhalla, N, Guin, R, Hirst, C, Hirst, M, Holt, R. A, Lee, D, Li, H. I, Mayo, M, Moore, R. A, Mungall, A. J, Pleasance, E, Gordon Robertson, A, Schein, J. E, Shafiei, A, Sipahimalani, P, Slobodan, J. R, Stoll, D, Tam, A, Thiessen, N, Varhol, R. J, Wye, N, Zeng, T, Zhao, Y, Birol, I, Jones, S. J. M, Cherniack, A. D, Saksena, G, Onofrio, R. C, Pho, N. H, Carter, S. L, Schumacher, S. E, Tabak, B, Hernandez, B, Gentry, J, Nguyen, H, Crenshaw, A, Ardlie, K, Beroukhim, R, Winckler, W, Getz, G, Gabriel, S. B, Chin, L, Park, P. J, Hoadley, K. A, Todd Auman, J, Fan, C, Turman, Y. J, Shi, Y, Li, L, Topal, M. D, He, X, Chao, H.-H, Prat, A, Silva, G. O, Iglesia, M. D, Zhao, W, Usary, J, Berg, J. S, Adams, M, Booker, J, Wu, J, Gulabani, A, Bodenheimer, T, Hoyle, A. P, Simons, J. V, Soloway, M. G, Mose, L. E, Jefferys, S. R, Balu, S, Parker, J. S, Neil Hayes, D, Malik, S, Mahurkar, S, Shen, H, Weisenberger, D. J, Triche, T. J, Lai, P. H, Bootwalla, M. S, Maglinte, D. T, Berman, B. P, van den Berg, D. J, Baylin, S. B, Creighton, C. J, Noble, M, Voet, D, Gehlenborg, N, Dicara, D, Zhang, J, Zhang, H, Wu, C.-J, Yingchun Liu, S, Lawrence, M. S, Zou, L, Sivachenko, A, Lin, P, Stojanov, P, Jing, R, Cho, J, Sinha, R, Park, R. W, Nazaire, M.-D, Robinson, J, Thorvaldsdottir, H, Mesirov, J, Reynolds, S, Kreisberg, R. B, Bernard, B, Bressler, R, Erkkila, T, Lin, J, Thorsson, V, Zhang, W, Ciriello, G, Weinhold, N, Schultz, N, Gao, J, Cerami, E, Gross, B, Jacobsen, A, Sinha, R, Arman Aksoy, B, Antipin, Y, Reva, B, Shen, R, Taylor, B. S, Ladanyi, M, Anur, P, Lu, Y, Liu, W, Verhaak, R. R. G, Mills, G. B, Akbani, R, Zhang, N, Broom, B. M, Casasent, T. D, Wakefield, C, Unruh, A. K, Baggerly, K, Coombes, K, Haussler, D, Benz, C. C, Stuart, J. M, Benz, S. C, Zhu, J, Szeto, C. C, Scott, G. K, Yau, C, Paull, E. O, Carlin, D, Wong, C, Sokolov, A, Thusberg, J, Mooney, S, Ng, S, Goldstein, T. C, Ellrott, K, Grifford, M, Wilks, C, Ma, S, Yan, C, Hu, Y, Gastier-Foster, J. M, Bowen, J, Ramirez, N. C, Black, A. D, Xpath Error Unknown Variable Tname, R. E, White, P, Zmuda, E. J, Frick, J, Lichtenberg, T. M, Brookens, R, George, M. M, Gerken, M. A, Harper,

H. A, Leraas, K. M, Wise, L. J, Tabler, T. R, McAllister, C, Barr, T, Tarvin, K, Saller, C, Sandusky, G, Iacocca, M. V, Brown, J, Rabeno, B, Czerwinski, C, Dolzhansky, O, Abramov, M, Voronina, O, Suchorska, W. M, Murawa, D, Kycler, W, Ibbs, M, Korski, K, Spychała, A, Murawa, P, Brzeziński, J. J, Perz, H, Łaźniak, R, Teresiak, M, Tatka, H, Leporowska, E, Bogusz-Czerniewicz, M, Malicki, J, Mackiewicz, A, van Le, X, Kohl, B, Viet Tien, N, Thorp, R, van Bang, N, Sussman, H, Duc Phu, B, Hajek, R, Phi Hung, N, Viet The Phuong, T, Quyet Thang, H, Penny, R, Mallery, D, and Curl...
Comprehensive molecular portraits of human breast tumours.
Nature, 490(7418):61–70, 2012.

Kroshko, D.
OpenOpt: Free scientific-engineering software for mathematical modeling and optimization.
URL http://www. openopt. org, 2007.

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

Langfelder, P and Horvath, S.
WGCNA: an R package for weighted correlation network analysis.
BMC Bioinformatics, 9(1):559, 2008.

Liberzon, A, Subramanian, A, and Pinchback, R.
Molecular signatures database (MSigDB) 3.0.
. . . , 2011.

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

Ma, B, Wilker, E. H, Willis-Owen, S. A. G, Byun, H.-M, Wong, K. C. C, Motta, V, Baccarelli, A. A, Schwartz, J, Cookson, W. O. C. M, Khabbaz, K, Mittleman, M. A, Moffatt, M. F, and Liang, L.
Predicting DNA methylation level across human tissues.
Nucleic Acids Research, 42(6):3515–3528, 2014.

Makhorin, A.

Glpk (gnu linear programming kit), version 4.42.
URL http://www. gnu. org/software/glpk, 2004.

Marschall, T and Rahmann, S.
Efficient exact motif discovery.
Bioinformatics, 25(12):i356–64, 2009.

Matsui, C, Inoue, E, Kakita, A, Arita, K, Deguchi-Tawarada, M, Togawa, A, Yamada, A, Takai, Y, and Takahashi, H.
Involvement of the γ-secretase-mediated EphA4 signaling pathway in synaptic pathogenesis of Alzheimer's disease.
Brain Pathology, 22(6):776–787, 2012.

Matys, V, Fricke, E, Geffers, R, Gossling, E, Haubrock, M, Hehl, R, Hornischer, K, Karas, D, Kel, A. E, Kel-Margoulis, O. V, Kloos, D. U, Land, S, Lewicki-Potapov, B, Michael, H, Munch, R, Reuter, I, Rotert, S, Saxel, H, Scheer, M, Thiele, S, and Wingender, E.
TRANSFAC(R): transcriptional regulation, from patterns to profiles.
Nucleic Acids Research, 31(1):374–378, 2002.

Mega, M. S, Cummings, J. L, Fiorello, T, and Gornbein, J.
The spectrum of behavioral changes in Alzheimer's disease.
Neurology, 46(1):130–135, 1996.

Melnikov, A, Scholtens, D, Godwin, A, and Levenson, V.
Differential Methylation Profile of Ovarian Cancer in Tissues and Plasma.
The Journal of Molecular Diagnostics, 11(1):60–65, 2009.

Miklos, G. L. G and Rubin, G. M.
The Role of the Genome Project in Determining Gene Function: Insights from Model Organisms.
Cell, 86(4):521–529, 1996.

Miller, L. H.
Table of Percentage Points of Kolmogorov Statistics.
Journal of the American Statistical Association, 51(273):111, 1956.

Milo, R, Shen-Orr, S, Itzkovitz, S, Kashtan, N, Chklovskii, D, and Alon, U.
Network motifs: simple building blocks of complex networks.
Science (New York, NY), 298(5594):824–827, 2002.

Minati, L, Edginton, T, Bruzzone, M. G, and Giaccone, G.
Reviews: Current Concepts in Alzheimer's Disease: A Multidisciplinary Review.

American Journal of Alzheimer's Disease and Other Dementias, 24(2):95–121, 2009.

Newman, M. E. J.
The Structure and Function of Complex Networks.
SIAM Review, 45(2):167–256, 2003.

Ng, A. Y, Jordan, M. I, and Weiss, Y.
On Spectral Clustering: Analysis and an algorithm.
ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, pages 849–856, 2001.

Ono, K.
idekerlab/cyREST, 2015.
URL https://github.com/idekerlab/cyREST.

Pasquinelli, A. E.
MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship.
Nature Reviews Genetics, 13(4):271–282, 2012.

Qiu, C, Wang, J, Yao, P, Wang, E, and Cui, Q.
microRNA evolution in a human transcription factor and microRNA regulatory network.
BMC systems biology, 4:90–90, 2009.

Raber, J, Huang, Y, and Ashford, J. W.
ApoE genotype accounts for the vast majority of AD risk and AD pathology.
Neurobiology of aging, 25(5):641–650, 2004.

Rai, M, Verma, S, and Tapaswi, S.
A Power Aware Minimum Connected Dominating Set for Wireless Sensor Networks.
Journal of networks, 4(6), 2009.

Rhinn, H, Fujita, R, Qiang, L, Cheng, R, Lee, J. H, and Abeliovich, A.
Integrative genomics identifies APOE $\varepsilon4$ effectors in Alzheimer's disease.
Nature, 500(7460):45–50, 2013.

Rimmelé, P, Komatsu, J, Hupé, P, Roulin, C, Barillot, E, Dutreix, M, Conseiller, E, Bensimon, A, Moreau-Gachelin, F, and Guillouf, C.
Spi-1/PU.1 oncogene accelerates DNA replication fork elongation and promotes genetic instability in the absence of DNA breakage.
Cancer Research, 70(17):6757–6766, 2010.

Ritchie, M. E, Phipson, B, Wu, D, Hu, Y, Law, C. W, Shi, W, and Smyth, G. K.
limma powers differential expression analyses for RNA-sequencing and microarray studies.
Nucleic Acids Research, 43(7):gkv007–e47, 2015.

Saito, R, Smoot, M. E, Ono, K, Ruscheinski, J, Wang, P.-L, Lotia, S, Pico, A. R, Bader, G. D, and Ideker, T.
A travel guide to Cytoscape plugins.
Nature Methods, 9(11):1069–1076, 2012.

Sakurai, T, Kondoh, N, Arai, M, Hamada, J.-i, Yamada, T, Kihara-Negishi, F, Izawa, T, Ohno, H, Yamamoto, M, and Oikawa, T.
Functional roles of Fli-1, a member of the Ets family of transcription factors, in human breast malignancy.
Cancer science, 98(11):1775–1784, 2007.

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

Sanger, F, Nicklen, S, and Coulson, A. R.
DNA sequencing with chain-terminating inhibitors.
Proceedings of the National Academy of Sciences of the United States of America, 74(12):5463–5467, 1977.

Sengupta, D and Bandyopadhyay, S.
Participation of microRNAs in human interactome: extraction of microRNA-microRNA regulations.
Molecular BioSystems, 7(6):1966–1973, 2011.

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

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

Shannon, P, Markiel, A, Ozier, O, Baliga, N. S, Wang, J. T, Ramage, D, Amin, N, Schwikowski, B, and Ideker, T.

Cytoscape: a software environment for integrated models of biomolecular interaction networks.
Genome research, 13(11):2498–2504, 2003b.

Shukla, V, Skuntz, S, and Pant, H. C.
Deregulated Cdk5 activity is involved in inducing Alzheimer's disease.
Archives of Medical Research, 43(8):655–662, 2012.

Siegel, R, Ma, J, Zou, Z, and Jemal, A.
Cancer statistics, 2014.
CA: a cancer journal for clinicians, 64(1):9–29, 2014.

Smirnov, N.
Table for Estimating the Goodness of Fit of Empirical Distributions.
The Annals of Mathematical Statistics, 19(2):279–281, 1948.

Smoot, M. E, Ono, K, Ruscheinski, J, Wang, P.-L, and Ideker, T.
Cytoscape 2.8: new features for data integration and network visualization.
Bioinformatics, 27(3):431–432, 2011.

Spaniol, C.
Mebitoo - A Membrane Bioinformatics Sequence Analysis Toolbox.
Master's thesis, Universität des Saarlandes, Saarbrücken, 2009.

Spaniol, C, Hamed, M, Trumm, J, and Helms, V.
Mebitoo - an Extensible Software Framework for Bioinformatics Analysis Workflow Automatization.
In *Proceedings of the ISCA 7th International Conference on Bioinformatics and Computational Biology, BICoB-2015, March 9-11, 20 15, Waikiki Beach Ressort, Honululu, Hawaii, USA*, 2015.

Syvänen, A.-C.
Accessing genetic variation: genotyping single nucleotide polymorphisms.
Nature Reviews Genetics, 2(12):930–942, 2001.

Tesson, B. M, Breitling, R, and Jansen, R. C.
DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules.
BMC Bioinformatics, 11(1):497, 2010.

Thieffry, D, Huerta, A. M, Pérez-Rueda, E, and Collado-Vides, J.
From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli.
Bioessays, 20(5):433–440, 1998.

Touleimat, N and Tost, J.
Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation.
dx.doi.org, 4(3):325–341, 2012.

Trevino, V, Falciani, F, and Barrera-Saldaña, H. A.
DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research.
Molecular Medicine, 13(9-10):527–541, 2007.

Visser, P. J, Verhey, F. R. J, Hofman, P. A. M, Scheltens, P, and Jolles, J.
Medial temporal lobe atrophy predicts Alzheimer's disease in patients with minor cognitive impairment.
Journal of Neurology, Neurosurgery & Psychiatry, 72(4):491–497, 2002.

Volinia, S and Croce, C. M.
Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer.
In *Proceedings of the National Academy of Sciences*, pages 7413–7417, 2013.

Voller, A, Bartlett, A, and Bidwell, D. E.
Enzyme immunoassays with special reference to ELISA techniques.
Journal of Clinical Pathology, 31(6):507–520, 1978.

von Luxburg, U.
A tutorial on spectral clustering.
Statistics and Computing, 17(4):395–416, 2007.

Wang, B, Jiang, J, 0028, W. W, Zhou, Z.-H, and Tu, Z.
Unsupervised metric fusion by cross diffusion.
CVPR, pages 2997–3004, 2012.

Wang, B, Mezlini, A. M, Demir, F, Fiume, M, Tu, Z, Brudno, M, Haibe-Kains, B, and Goldenberg, A.
Similarity network fusion for aggregating data types on a genomic scale.
Nature Methods, 11(3):333–337, 2014.

Wang, J, Lu, M, Qiu, C, and Cui, Q.
TransmiR: a transcription factor-microRNA regulation database.
Nucleic Acids Research, 38(Database issue):D119–D122, 2009.

Wei, Y.-C and Cheng, C.-K.
Towards efficient hierarchical designs by ratio cut partitioning.

IEEE, 1989.

Wong, H.-K. A, Veremeyko, T, Patel, N, Lemere, C. A, Walsh, D. M, Esau, C, Vanderburg, C, and Krichevsky, A. M.
De-repression of FOXO3a death axis by microRNA-132 and -212 causes neuronal apoptosis in Alzheimer's disease.
Human Molecular Genetics, 22(15):3077–3092, 2013.

Xiao, F, Zuo, Z, Cai, G, Kang, S, Gao, X, and Li, T.
miRecords: an integrated resource for microRNA-target interactions.
Nucleic Acids Research, 37(Database issue):D105–D110, 2008.

Yang, J.-H, Li, J.-H, Jiang, S, Zhou, H, and Qu, L.-H.
ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data.
Nucleic Acids Research, 41(Database issue):D177–D187, 2012.

Yang, J.-H, Li, J.-H, Shao, P, Zhou, H, Chen, Y.-Q, and Qu, L.-H.
starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data.
Nucleic Acids Research, 39(Database issue):D202–9, 2011.

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

Zar, J. H.
Biostatistical Analysis (5th Edition).
Prentice-Hall, Inc., 2007.

Zhang, B, Gaiteri, C, Bodea, L.-G, Wang, Z, McElwee, J, Podtelezhnikov, A. A, Zhang, C, Xie, T, Tran, L, Dobrin, R, Fluder, E, Clurman, B, Melquist, S, Narayanan, M, Suver, C, Shah, H, Mahajan, M, Gillis, T, Mysore, J, MacDonald, M. E, Lamb, J. R, Bennett, D. A, Molony, C, Stone, D. J, Gudnason, V, Myers, A. J, Schadt, E. E, Neumann, H, Zhu, J, and Emilsson, V.
Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease.
Cell, 153(3):707–720, 2013.

List of Abbreviations

AD	Alzheimer's Disease
APP	Amyloid Precursor Protein
AUC ROC	Area Under the Curve Receiver Operator Characteristics
BC	Breast Cancer
BH	Benjamini Hochberg
CDS	Connected Dominating Set
CLOB	Character Large Objects
CR	Coverage Ratio
DAVID	Database for Annotation Visualization and Integrated Discovery
DMR	Differentially Methylated Region
ELISA	Enzyme Linked Immunosorbent Assay
EOAD	Early-Onset Alzheimer's Disease
FDR	False Discovery Rate
FFL	Feed Forward Loop
GLPK	GNU Linear Programming Kit
HTML	Hypertext Markup Language
JSON	Javascript Object Notation
LOAD	Late-Onset Alzheimer's Disease
MCDS	Minimum Connected Dominating Set

miRNA	microRNA
mRNA	messenger RNA
	Normalized Mutual Information
PHP	PHP Hypertext Preprocessor
PPI	Protein-protein interaction
PSEN	Presenilin
REST	Representational State Transfer
RNA	Ribonucleic acid
SAM	Significance Analysis of Micoarray
SNF	Similarity Network Fusion
SNP	Single Nucleotide Polymorphism
TCGA	The Cancer Genome Atlas
TCGA	The Cancer Genome Atlas
TSS	Transcriptional Start Site

Appendix A

Alzheimer Study

A.1 Table of differentially expressed genes

Table A.1: Table shows 148 differentially expressed genes with p-Value < 0.01 using limma. (5 differentially expressed sites could not be annotated)

No	Gene Symbol	log FC	t	p-Value	adj. p-Value	B
1	GFAP	1.4	7.1	1.6e-09	4.8e-05	11
2	C5orf41	0.58	6.7	7.6e-09	9e-05	9.8
3	RNU1G2	1.4	6.7	9.1e-09	9e-05	9.7
4	RNU1-3	1.3	6.3	3.5e-08	0.00026	8.4
5	RNU1-5	1.2	6.1	7.8e-08	0.00044	7.7
6	AEBP1	1.5	6.1	8.8e-08	0.00044	7.6
7	HBP1	0.59	5.9	1.7e-07	0.00064	7
8	C1orf110	1.2	5.9	1.8e-07	0.00064	6.9
9	RHOQ	0.77	5.9	2e-07	0.00064	6.9
10	PPARBP	0.51	5.8	2.8e-07	0.00076	6.5
11	MYBPC1	1.1	5.8	2.8e-07	0.00076	6.5
12	EIF3E	0.38	5.7	3.4e-07	0.00079	6.4
13	NFKB1	0.49	5.7	3.5e-07	0.00079	6.4
14	TNPO1	0.56	5.7	4.2e-07	0.00089	6.2
15	C5orf41	0.57	5.5	8.6e-07	0.0017	5.5
16	PKN2	0.62	5.4	1.3e-06	0.0024	5.1
17	AK1	0.53	5.4	1.4e-06	0.0024	5.1

. . .

Table A.1 – *Continued*

	Gene Symbol	log FC	t	p-Value	adj. p-Value	B
18	MAP1LC3A	-0.51	-5.3	1.5e-06	0.0024	5
19	ITPKB	1.2	5.3	1.6e-06	0.0024	5
20	SRGAP1	0.78	5.2	2.2e-06	0.003	4.7
21	KCNF1	-0.97	-5.2	2.9e-06	0.0038	4.4
22	SST	-1.5	-5.2	3e-06	0.0038	4.4
23	PPEF1	-0.99	-5.1	3.1e-06	0.0038	4.4
24	DDR2	0.96	5.1	3.8e-06	0.0045	4.2
25	LOC284988	0.74	5.1	4e-06	0.0045	4.1
26	HBQ1	-0.85	-5.1	4.2e-06	0.0046	4.1
27	FOXJ1	1.3	5	4.7e-06	0.0048	4
28	FAM107A	0.64	5	4.9e-06	0.0048	4
29	LOC388481	-0.49	-5	5e-06	0.0048	3.9
30	KCNA4	-0.28	-5	5.1e-06	0.0048	3.9
31	LOC144438	0.51	5	5.1e-06	0.0048	3.9
32	ANTXR1	1	5	5.8e-06	0.0049	3.8
33	MRPS25	-0.56	-5	5.8e-06	0.0049	3.8
34	ENTPD2	0.57	5	5.8e-06	0.0049	3.8
35	LOC649362	0.83	5	5.9e-06	0.0049	3.8
36	KRT17	-1.1	-5	6.2e-06	0.0049	3.7
37	APLNR	1.6	5	6.3e-06	0.0049	3.7
38	ITGB4	1.3	4.9	6.6e-06	0.005	3.7
39	SLC25A46	-0.48	-4.9	6.9e-06	0.005	3.6
40	SLC16A9	1.1	4.9	7e-06	0.005	3.6
41	TOB1	0.67	4.9	7.1e-06	0.005	3.6
42	NPM3	-0.47	-4.9	7.4e-06	0.005	3.6
43	SMAD5	0.51	4.9	7.5e-06	0.005	3.6
44	PLEC1	0.65	4.9	7.6e-06	0.005	3.6
45	ARHGEF9	-0.78	-4.9	8.6e-06	0.0055	3.4
46	PRKCB	-0.82	-4.9	8.8e-06	0.0055	3.4
47	DNALI1	0.77	4.9	8.9e-06	0.0055	3.4
48	FAM89A	0.92	4.8	9.7e-06	0.0057	3.3

. . .

Table A.1 – *Continued*

	Gene Symbol	log FC	t	p-Value	adj. p-Value	B
49	ABCA1	1.1	4.8	9.7e-06	0.0057	3.3
50	SLC38A2	0.65	4.8	9.8e-06	0.0057	3.3
51	AHNAK	1.1	4.8	1.1e-05	0.0059	3.3
52	CAPRIN2	-0.71	-4.8	1.1e-05	0.0059	3.2
53	NELL1	-1.3	-4.8	1.1e-05	0.0059	3.2
54	NLGN4X	-0.63	-4.8	1.2e-05	0.0059	3.2
55	PAPOLA	0.5	4.8	1.2e-05	0.0059	3.2
56	ARHGDIG	-1.1	-4.8	1.2e-05	0.0059	3.2
57	PTTG1IP	0.74	4.8	1.2e-05	0.0059	3.2
58	PALLD	0.92	4.8	1.2e-05	0.0059	3.1
59	TUBG2	-0.45	-4.8	1.2e-05	0.0059	3.1
60	PRPH2	-0.59	-4.8	1.3e-05	0.0059	3.1
61	TMEM163	-0.71	-4.8	1.3e-05	0.0059	3.1
62	FXYD7	-0.91	-4.8	1.3e-05	0.0059	3.1
63	TPI1	-0.52	-4.8	1.3e-05	0.0059	3.1
64	HSPB3	-0.77	-4.7	1.3e-05	0.006	3.1
65	TSPAN7	-1.3	-4.7	1.4e-05	0.0064	3
66	GNB2L1	0.2	4.7	1.5e-05	0.0066	3
67	ATP6V0D1	-0.55	-4.7	1.5e-05	0.0066	2.9
68	TARBP1	-0.94	-4.7	1.6e-05	0.0066	2.9
69	DCLK1	-0.85	-4.7	1.6e-05	0.0066	2.9
70	CORT	-0.98	-4.7	1.6e-05	0.0066	2.9
71	CPNE9	-0.76	-4.7	1.7e-05	0.0071	2.8
72	VKORC1L1	-0.58	-4.7	1.8e-05	0.0071	2.8
73	NRSN1	-1.7	-4.7	1.8e-05	0.0071	2.8
74	KANK2	0.49	4.7	1.8e-05	0.0071	2.8
75	LPP	0.73	4.7	1.8e-05	0.0071	2.8
76	C13orf36	-0.51	-4.6	1.9e-05	0.0072	2.7
77	PABPC1	0.68	4.6	1.9e-05	0.0073	2.7
78	TSC22D1	-0.33	-4.6	2e-05	0.0073	2.7
79	ME3	-0.79	-4.6	2e-05	0.0073	2.7

. . .

Table A.1 – *Continued*

	Gene Symbol	log FC	t	p-Value	adj. p-Value	B
80	CDK5	-0.83	-4.6	2e-05	0.0073	2.7
81	HPRT1	-1.2	-4.6	2e-05	0.0073	2.7
82	MYOT	1.1	4.6	2e-05	0.0073	2.7
83	JARID1A	0.68	4.6	2.1e-05	0.0073	2.7
84	SLC35F1	-1.1	-4.6	2.2e-05	0.0077	2.6
85	SEZ6L	-0.21	-4.6	2.3e-05	0.0078	2.6
86	METAP2	0.35	4.6	2.3e-05	0.0078	2.6
87	GLIS3	0.85	4.6	2.3e-05	0.0079	2.5
88	RIT2	-0.9	-4.6	2.4e-05	0.008	2.5
89	MAL2	-1.2	-4.6	2.5e-05	0.0081	2.5
90	RPL14	0.83	4.6	2.5e-05	0.0081	2.5
91	ADAM23	-0.93	-4.6	2.5e-05	0.0081	2.5
92	LOC90113	-0.72	-4.6	2.5e-05	0.0081	2.5
93	NPTX1	-0.43	-4.6	2.6e-05	0.0081	2.5
94	SYNC1	0.65	4.6	2.6e-05	0.0081	2.4
95	LOC653308	-0.45	-4.6	2.7e-05	0.0081	2.4
96	C10orf105	1.4	4.5	2.7e-05	0.0081	2.4
97	OCIAD1	-0.63	-4.5	2.7e-05	0.0081	2.4
98	MYST3	0.54	4.5	2.8e-05	0.0081	2.4
99	RIN2	0.71	4.5	2.8e-05	0.0081	2.4
100	SLC35A2	-0.38	-4.5	2.8e-05	0.0081	2.4
101	LOC100130148	-0.32	-4.5	2.8e-05	0.0081	2.4
102	GLRA2	-0.5	-4.5	2.9e-05	0.0081	2.4
103	SPHKAP	-0.96	-4.5	2.9e-05	0.0081	2.4
104	NAV2	0.81	4.5	2.9e-05	0.0082	2.3
105	SMPD3	-0.24	-4.5	3e-05	0.0084	2.3
106	RPH3A	-0.85	-4.5	3.1e-05	0.0085	2.3
107	TUBB2A	-1.5	-4.5	3.1e-05	0.0085	2.3
108	WBP5	0.41	4.5	3.2e-05	0.0085	2.3
109	SYTL4	1.1	4.5	3.2e-05	0.0085	2.3
110	SP1	0.6	4.5	3.2e-05	0.0085	2.3

. . .

Table A.1 – *Continued*

	Gene Symbol	log FC	t	p-Value	adj. p-Value	B
111	TSPYL2	-0.81	-4.5	3.3e-05	0.0086	2.2
112	OSBPL2	0.39	4.5	3.3e-05	0.0086	2.2
113	OSBPL10	-0.74	-4.5	3.3e-05	0.0086	2.2
114	SYT1	-2.1	-4.5	3.4e-05	0.0086	2.2
115	PIGZ	-0.68	-4.5	3.5e-05	0.0086	2.2
116	CFL1	-0.32	-4.5	3.5e-05	0.0086	2.2
117	GYPC	0.85	4.5	3.5e-05	0.0086	2.2
118	BMPER	-0.52	-4.5	3.5e-05	0.0086	2.2
119	EGR1	-1	-4.5	3.6e-05	0.0087	2.2
120	MDH2	-0.34	-4.5	3.6e-05	0.0087	2.2
121	DIRAS2	-1.4	-4.5	3.6e-05	0.0088	2.1
122	SPOP	0.36	4.5	3.8e-05	0.0089	2.1
123	CNTN4	-0.15	-4.5	3.8e-05	0.0089	2.1
124	OLFM1	-1.1	-4.4	3.8e-05	0.0089	2.1
125	RASL12	0.98	4.4	3.9e-05	0.009	2.1
126	LOC729513	0.72	4.4	4e-05	0.0091	2.1
127	NXPH1	-0.85	-4.4	4e-05	0.0092	2.1
128	CAPS	1	4.4	4e-05	0.0092	2.1
129	C4orf44	-0.37	-4.4	4.1e-05	0.0092	2
130	ZHX3	0.55	4.4	4.1e-05	0.0092	2
131	RGS4	-1.8	-4.4	4.3e-05	0.0094	2
132	FAM89A	0.84	4.4	4.3e-05	0.0094	2
133	NCALD	-1.2	-4.4	4.4e-05	0.0095	2
134	FRMPD2	-0.71	-4.4	4.4e-05	0.0095	2
135	IL13RA1	0.77	4.4	4.4e-05	0.0096	2
136	GLTSCR2	0.38	4.4	4.6e-05	0.0097	1.9
137	LOC647251	-1.1	-4.4	4.6e-05	0.0097	1.9
138	NRXN1	-0.59	-4.4	4.6e-05	0.0097	1.9
139	DGKI	-0.3	-4.4	4.7e-05	0.0098	1.9
140	HSPB8	0.81	4.4	4.7e-05	0.0098	1.9
141	BBX	0.56	4.4	4.8e-05	0.0098	1.9

. . .

Table A.1 – *Continued*

	Gene Symbol	log FC	t	p-Value	adj. p-Value	B
142	ITGB5	0.75	4.4	4.8e-05	0.0098	1.9
143	SLN	-0.58	-4.4	4.8e-05	0.0098	1.9
144	SIX5	0.83	4.4	4.9e-05	0.0098	1.9
145	BRE	-0.39	-4.4	4.9e-05	0.0098	1.9
146	C1QTNF5	0.92	4.4	4.9e-05	0.0098	1.9
147	ZBTB40	0.55	4.4	5e-05	0.0099	1.9
148	FAM19A1	-0.88	-4.4	5.1e-05	0.0099	1.8

A.2 Methylation Data Preprocessing

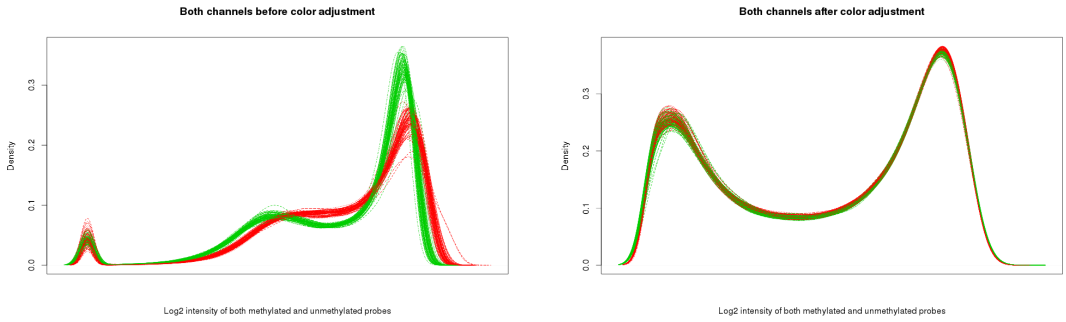


Figure A.1: Density of methylated and unmethylated probes before and after processing

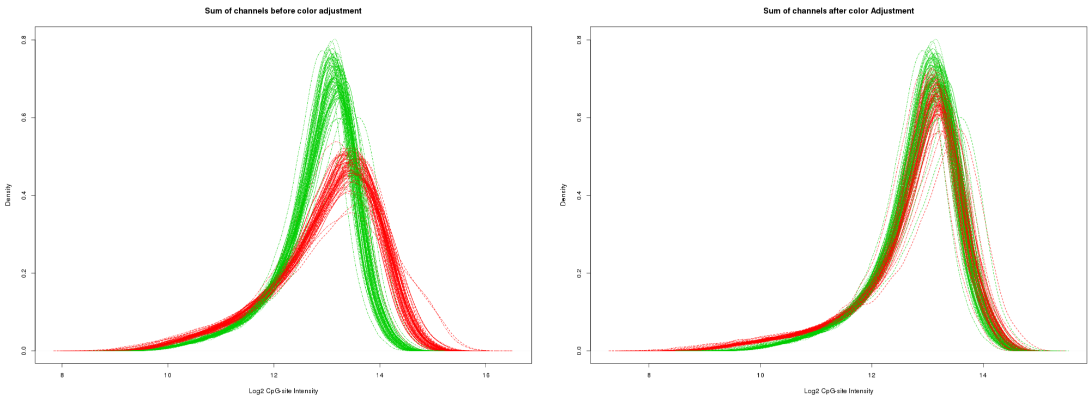


Figure A.2: Summed color bias before and after

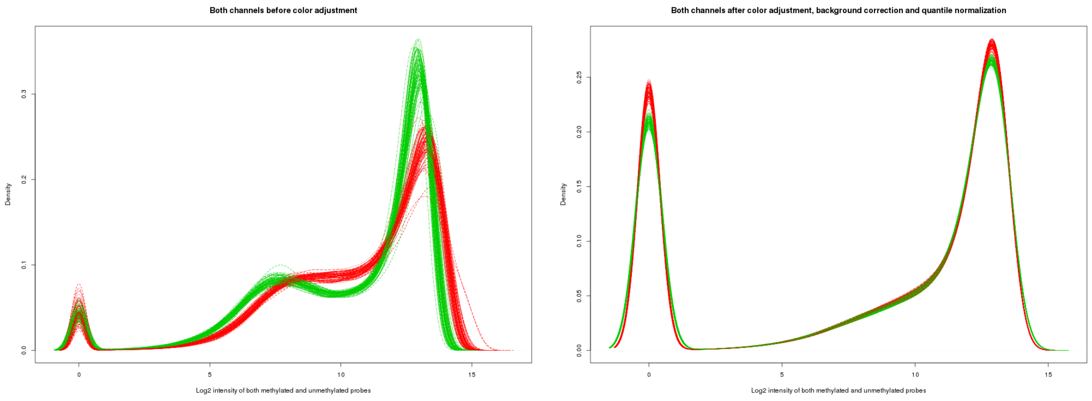


Figure A.3: Color bias for both channels before and after

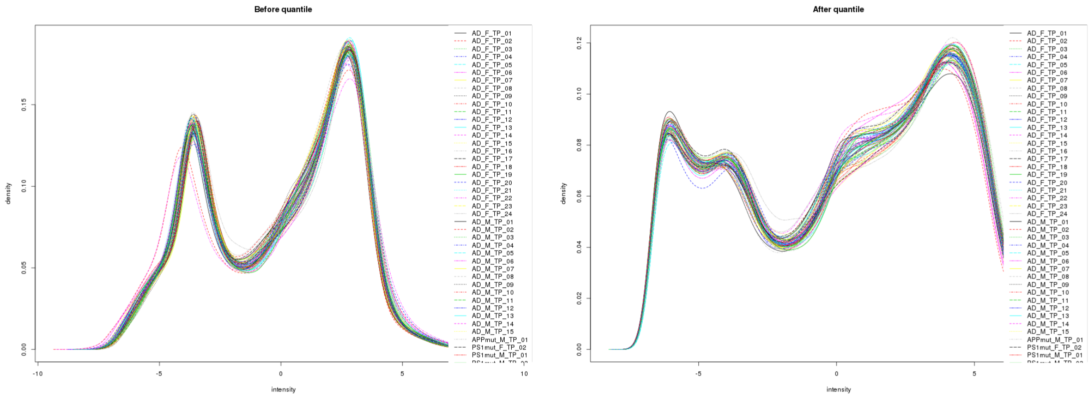


Figure A.4: Density before and after quantile normalization

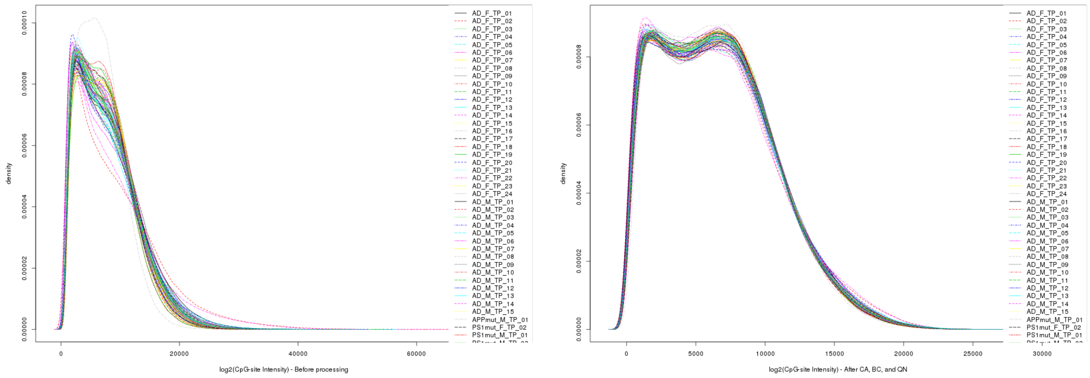


Figure A.5: CpG Intensity before and after processing