PhD Thesis

Pathway-centric approaches to the analysis of high-throughput genomics data

Sonja Hänzelmann

Department of Experimental and Health Science
Functional Genomics Group, Research Programme on Biomedical Informatics (GRIB), IMIM-UPF

Supervisor: Robert Castelo

Barcelona 2012

For my family.

Agraïments a la Fundació Institut Mar d'Investigacions Mèdiques (IMIM) per finançar l'enquadernació d'aquesta tesi.

Preface

"Science is built up of facts, as a house is built of stones; but an accumulation of facts is no more a science than a heap of stones is a house."
– Henri Poincaré,
(Hypotheses in Nature, as translated by George Bruce Halsted (1913))

The discovery of the structure of the deoxyribonucleic acid (DNA) [*Watson and Crick*, 1953] revealed the properties of the blueprint that instructs the cell on how to acquire its fate. The entire collection of genetic material in a cell, is called the genome. The detection of the exact order of the bases in a strand of DNA is called sequencing. When efficient sequencing methods were introduced [*Maxam and Gilbert*, 1977; *Sanger et al.*, 1977], research in molecular biology started to accumulate increasing amounts of genetic data. The availability of whole-genome sequencing data led to open-ended research opportunities and to the rapidly advancing field of genomics [*Lederberg and Mccray*, 2001]. Transcriptomics tries to understand on a genome-wide scale how an organism responds to different stimuli, such as disease or environmental factors. Over the last decade, microarrays have been a valuable instrument to measure transcriptomic changes. Representative selections of expressed sequences have been used to build microarrays and allow for parallel analysis of gene expression in many samples. In October 1990 the Human Genome Project (HGP) was announced, the goal of the project was to index and sequence the entire human genome. The project made use of sequencing techniques that gave a big push to the field, which made it quite popular. A lot of excitement bubbled up [*Gisler et al.*, 2010] when the first draft of the human genome [*Venter et al.*, 2001; *Lander et al.*, 2001] was released in the year 2000. Until then various estimates of the number of genes were made, up to 100,000 [*Milner and Sutcliffe*, 1983], and in 1999 even more genes (142,634) were estimated [*Dickson*, 1999]. The pharmaceutical industry was hoping to use these estimated hundreds of thousands of genes to find large druggable subsets of proteins that could be modulated by a compound. Protein-coding genes presented a rich resource for potential new drug targets. Surprisingly, the number of genes was not as high as expected, around 23,000 protein coding genes. Furthermore, diseases are very complex, hence, more time is needed to translate these data into new medical treatments. Accordingly, Eric Lander summarized some insights of the HGP, as follows:

Even though, medicine has not yet been revolutionized as expected and hoped for, regardless the HGP set ground for future research in biology. The availability

of a common scaffold transformed biology because every element can be linked and put into context - Eric Lander [*Lander*, 2011].

Since then, many organisms have been sequenced and their genomes have been made available, for example, through genome browsers [*Kent et al.*, 2002; *Flicek et al.*, 2011]. Sequencing of organisms and the human microbiome, to name a few, are still ongoing genomics research projects. Different branches of genomics have evolved over the years, including, meta-genomics, phylogenomics, pharmacogenomics, and functional genomics. From the need to interpret and organize the ever-increasing piles of data the field of functional genomics emerged [*Hieter and Boguski*, 1997]. This field tries to make sense of the interactions of genes with each other and to describe gene functions with focus on transcription, translation, and protein-protein interactions.

With the reduction in sequencing costs, an increasing number of personal genomes have been sequenced, the first one from Craig Venter [*Levy et al.*, 2007] and many to follow. The Personal genomes project [*Church*, 2005] provides a platform where individual genomes can be deposited, freely available to the public. The field of personal genomics gained a lot of attention because many of the human diseases have a hereditary component and thus are expected to have a genetic blueprint for the disease. The hope is that complex diseases such as cancer, diabetes, and cardiovascular disease can be treated in a new, more powerful way, allowing for more effective treatments based on an individual's genetic make-up.

Abstract

In the last decade, molecular biology has expanded from a reductionist view to a systems-wide view that tries to unravel the complex interactions of cellular components. Owing to the emergence of high-throughput technology it is now possible to interrogate entire genomes at an unprecedented resolution. The dimension and unstructured nature of these data made it evident that new methodologies and tools are needed to turn data into biological knowledge. To contribute to this challenge we exploited the wealth of publicly available high-throughput genomics data and developed bioinformatics methodologies focused on extracting information at the pathway rather than the single gene level. First, we developed Gene Set Variation Analysis (GSVA), a method that facilitates the organization and condensation of gene expression profiles into gene sets. GSVA enables pathway-centric downstream analyses of microarray and RNA-seq gene expression data. The method estimates sample-wise pathway variation over a population and allows for the integration of heterogeneous biological data sources with pathway-level expression measurements. To illustrate the features of GSVA, we applied it to several use-cases employing different data types and addressing biological questions. GSVA is made available as an R package within the Bioconductor project.

Secondly, we developed a pathway-centric genome-based strategy to reposition drugs in type 2 diabetes (T2D). This strategy consists of two steps, first a regulatory network is constructed that is used to identify disease driving modules and then these modules are searched for compounds that might target them. Our strategy is motivated by the observation that disease genes tend to group together in the same neighborhood forming disease modules and that multiple genes might have to be targeted simultaneously to attain an effect on the pathophenotype. To find potential compounds, we used compound exposed genomics data deposited in public databases. We collected about 20,000 samples that have been exposed to about 1,800 compounds. Gene expression can be seen as an intermediate phenotype reflecting underlying dysregulatory pathways in a disease. Hence, genes contained in the disease modules that elicit similar transcriptional responses upon compound exposure are assumed to have a potential therapeutic effect. We applied the strategy to gene expression data of human islets from diabetic and healthy individuals and identified four potential compounds, methimazole, pantoprazole, bitter orange extract and torcetrapib that might have a positive effect on insulin secretion. This is the first time a regulatory network of human islets has been used to reposition compounds for T2D.

In conclusion, this thesis contributes with two pathway-centric approaches to important bioinformatic problems, such as the assessment of biological function and "in silico" drug repositioning. These contributions demonstrate the central role of pathway-based analyses in interpreting high-throughput genomics data.

Resum

En l'última dècada, la biologia molecular ha evolucionat des d'una perspectiva reduccionista cap a una perspectiva a nivell de sistemes que intenta desxifrar les complexes interaccions entre els components cel·lulars. Amb l'aparició de les tecnologies d'alt rendiment actualment és possible interrogar genomes sencers amb una resolució sense precedents. La dimensió i la naturalesa desestructurada d'aquestes dades ha posat de manifest la necessitat de desenvolupar noves eines i metodologies per a convertir aquestes dades en coneixement biològic. Per contribuir a aquest repte hem explotat l'abundància de dades genòmiques procedents d'instruments d'alt rendiment i disponibles públicament, i hem desenvolupat mètodes bioinformàtics focalitzats en l'extracció d'informació a nivell de via molecular en comptes de fer-ho al nivell individual de cada gen. En primer lloc, hem desenvolupat GSVA (Gene Set Variation Analysis), un mètode que facilita l'organització i la condensació de perfils d'expressió dels gens en conjunts. GSVA possibilita anàlisis posteriors en termes de vies moleculars amb dades d'expressió gènica provinents de microarrays i RNA-seq. Aquest mètode estima la variació de les vies moleculars a través d'una població de mostres i permet la integració de fonts heterogènies de dades biològiques amb mesures d'expressió a nivell de via molecular. Per il·lustrar les característiques de GSVA, l'hem aplicat a diversos casos usant diferents tipus de dades i adreçant qüestions biològiques. GSVA està disponible com a paquet de programari lliure per R dins el projecte Bioconductor.

En segon lloc, hem desenvolupat una estratègia centrada en vies moleculars basada en el genoma per reposicionar fàrmacs per la diabetis tipus 2 (T2D). Aquesta estratègia consisteix en dues fases: primer es construeix una xarxa reguladora que s'utilitza per identificar mòduls de regulació gènica que condueixen a la malaltia; després, a partir d'aquests mòduls es busquen compostos que els podrien afectar. La nostra estratègia ve motivada per l'observació que els gens que provoquen una malaltia tendeixen a agrupar-se, formant mòduls patogènics, i pel fet que podria caldre una actuació simultània sobre múltiples gens per assolir un efecte en el fenotipus de la malaltia. Per trobar compostos potencials, hem usat dades genòmiques exposades a compostos dipositades en bases de dades públiques. Hem recollit unes 20.000 mostres que han estat exposades a uns 1.800 compostos. L'expressió gènica es pot interpretar com un fenotip intermedi que reflecteix les vies moleculars desregulades subjacents a una malaltia. Per tant, considerem que els gens d'un mòdul patològic que responen, a nivell transcripcional, d'una manera similar a l'exposició del medicament tenen potencialment un efecte terapèutic. Hem aplicat aquesta estratègia a dades d'expressió gènica en illots pancreàtics humans corresponents a individus sans i diabètics, i hem identificat quatre compostos potencials (methimazole, pantoprazole, extracte de taronja amarga i torcetrapib) que podrien tenir un efecte positiu sobre la secreció de la insulina. Aquest és el primer cop que una xarxa reguladora d'illots pancreàtics humans s'ha utilitzat per reposicionar compostos per a T2D.

En conclusió, aquesta tesi aporta dos enfocaments diferents en termes de vies moleculars a problemes bioinformàtics importants, com ho són el contrast de la funció biològica i el reposicionament de fàrmacs "in silico". Aquestes contribucions demostren el paper central de les anàlisis basades en vies moleculars a l'hora d'interpretar dades genòmiques procedents d'instruments d'alt rendiment.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Summary

The reductionist approach to identify disease associated genes is rapidly changing to a systems view of disease due to the eruption of data generated by high-throughput technologies, such as microarrays and next generation sequencing. Single gene approaches to the analysis of these data have been used to derive gene signatures that describe disease phenotypes on the molecular level of nucleic acids. This has shown that genes with similar expression profiles tend to have similar functions. By grouping these genes into gene sets, the participating pathways can be identified. Further, if putative regulators of these genes are identified, pathways can be dissected into regulatory modules. The concerted changes of these modules contain information about the underlying processes that are disrupted in a disease phenotype. Therefore, pathway-centric approaches to the analysis of high-throughput genomics data, where the canonical unit of analysis are gene sets and regulatory modules, rather than single genes, have the potential to contribute to a better understanding of molecular mechanisms. Further, the omics revolution generates enormous amounts of heterogeneous data on a genome-wide scale, including DNA-variation, gene expression, mutations, and clinical data. It is imperative to integrate these data to successfully identify disease driving genes/modules. To address this need, we developed Gene Set Variation Analysis (GSVA) (Chapter 3.1), a method that allows for the estimation of coordinated changes of pathways and their integration with heterogeneous data sources.

We expand our strategy to identify disease modules that can potentially be affected by compounds (Chapter 3.2). The hierarchical and scale-free nature of many biological networks suggests that genes with similar functions tend to cluster together in the same neighborhood. The same has been observed for disease driving genes. Motivated by these observations, we infer a transcriptional regulatory network with the aim to identify regulatory disease modules in type 2 diabetes (T2D). The notion behind this study is that the molecular profile of complex diseases such as T2D reflects the disease phenotype that results from many small changes in several genes, grouped into disease driving modules. Subsequently, we find potential compounds that are able to perturb the disease phenotype. For this, we made use of the drug exposed genomics data deposited in public databases and integrated them with the T2D disease modules to find potential therapeutics that could have an effect on dysfunctional insulin secretion.

In a larger collaborative effort reported in Appendix A, we used co-expression analysis to find modules that are involved in the disease phenotype of T2D. This work has been conducted during this thesis and is thus included, but not as one of the main objectives. The main contribution to this work comprised

the bioinformatic analyses. The main focus of this study was to infer regulatory modules from gene expression data to identify a disease module present in T2D. We identified a module that is significantly correlated with T2D and enriched for interleukin-related genes. We prioritized the genes contained in this module and identified *SFRP4* as a new biomarker for T2D involved in glucose reduction via interleukin 1β. We also carried out a transcription factor binding site (TFBS) analysis to get a better understanding of the underlying mechanism. We found that *SFRP4* has several binding sites for NF-κB, which in turn is regulated by interleukin 1β.

1.2 The Flow of Genetic Information

1.2.1 From sequence to function

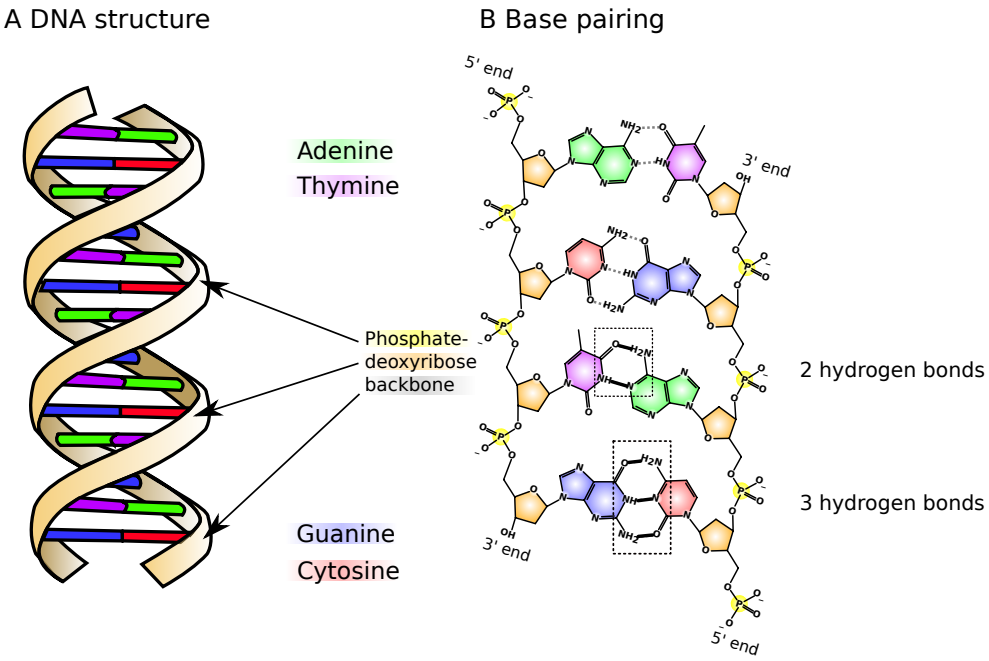


Figure 1.1: **Cartoon of the DNA.** The bases Adenine (A), Thymine (T), Cytosine (C) and Guanine (G) are attached to the sugar backbone of the DNA. The bases pair with each other via hydrogen bonds, giving the structure to the DNA. But only AT and GC can bind with each other, respectively. Figure adapted from Wikipedia.

The deoxyribonucleic acid (DNA) [*Watson and Crick*, 1953] is the blueprint that guides a cell to its fate. The DNA is built from the four bases Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). The backbone of the DNA is

composed of sugar and phosphates joined by ester bonds (Figure 1.1). A base is attached to the sugar deoxyribonucleic molecule on each side and pairs with the complementary base (base pairs). Pairs can only be formed by G with C and A with T. G binds with three hydrogen bonds to C and A with two hydrogen bonds to T (Figure 1.1). The binding of G and C is very strong because of the three hydrogen bonds and harder to break than the AT base pair.

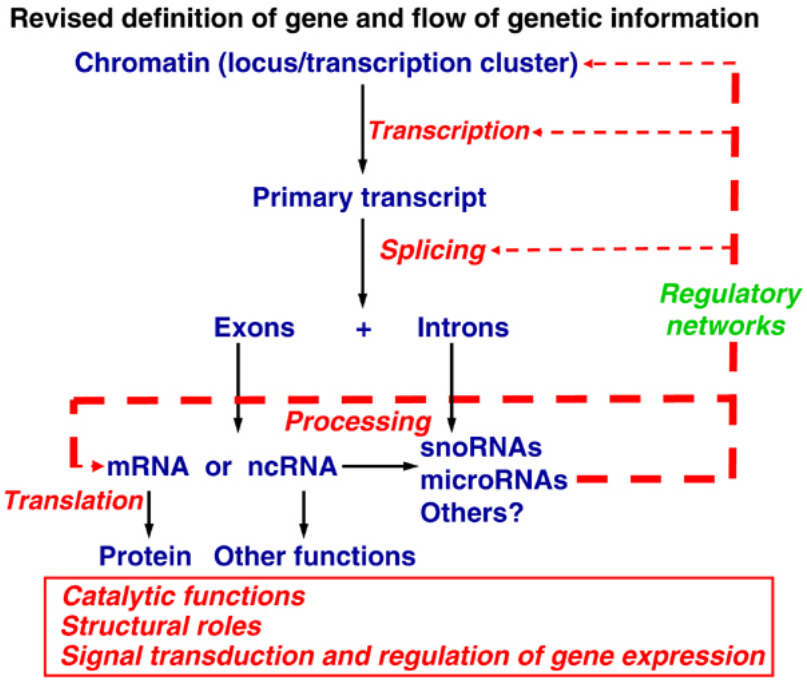


Figure 1.2: **Revised flow of genetic information.** With the discovery of non-coding RNAs a refined view on the regulation of gene expression is possible. On several steps, gene expression is influenced by non-coding RNA, suggesting a regulatory RNA network. Taken from [*Kenzelmann et al.*, 2006]

Recent evidence [*Birney et al.*, 2007; *Carninci et al.*, 2008] shows that as much as 93% of DNA in the human genome is transcribed into primary RNA transcripts. The transcribed RNA can be partitioned into non-coding RNA (ncRNA) and coding RNA (messenger RNA, mRNA). Non-coding RNAs include small interfering RNAs (siRNA) and micro-RNAs (miRNA), which participate in the coordination of gene regulation, chromatin remodeling and translation [*Kenzelmann et al.*, 2006]. Less than 1.5 % of the human genome codes for proteins. The information flow from transcription of the DNA into RNA, which is then translated into a protein is called the central dogma of molecular biology. It has been elaborated between 1941 [*Beadle and Tatum*, 1941] and 1958 [*Crick*, 1958] and revised in 1970 [*Crick*, 1970]. During the last two decades of research in molecular biology

it has been recognized that this model was over-simplified. Non-coding RNAs are key regulators of gene expression influencing the flow of information at several levels and contributing to the control of transcription and translation. These findings have been used to extend the dogma by another layer of regulation controlled by ncRNAs (Figure 1.2). In this thesis we focus on the transcriptional regulation by transcription factors.

1.2.2 Transcriptional regulation of gene expression

With the discovery of the *lac* operon it became evident that proteins can regulate gene expression. This observation resulted in a growing interest to study this mechanism [*Gilbert and Müller-Hill*, 1966]. These regulating proteins are called transcription factors (TF) and form the transcription initiation complex (Figure 1.3). This complex binds to the promoter region of the coding sequence, influencing RNA transcription. The promoter is the DNA sequence located 5' upstream of the transcription start site (TSS) and can be separated (in mammals) in conserved TATA-box enriched and CpG-rich [*Barrett et al.*, 2012]. Often, the promoter overlaps with the first coding sequence or includes it completely. Promoters initiate transcription and attract RNA polymerase II to the transcription start site (TSS). The complex consists of specific TFs, general TFs, co-factors, and RNA polymerase II [*Wasserman and Sandelin*, 2004]. Transcription is not only depending on co-factors and TFs and their ability to work together, but also the on the structure of the chromatin. Additional mechanisms that also control gene expression, are RNA interference, methylation and acetylation.

Methylation (addition of a methyl group) and acetylation (addition of an acetyl group) are types of histone (basic proteins) modifications that control gene expression. Different epigenetic factors (such as methyl groups) are added or removed from the histone tails determining whether genes are activated or not. DNA winds around histones to be more compact, this combination is called chromatin (Figure 1.3). Regions of the DNA that are wrapped around a histone are not accessible and thus cannot be transcribed. They become activated (unwound) when epigenetic factors bind to the histones.

The stretches of DNA where transcription factors are able to bind, are called transcription factor binding sites (TFBS) (Figure 1.3). TFBS can be far away from the TSS (distal TFBS) or close to the TSS (proximal TFBS). The delineation of the TSS is problematic because of different, overlapping transcripts belonging to the same gene and leading to multiple TSS. Finding TFBS is a crucial task, because of their importance to the expression of gene signatures [*Veerla et al.*, 2010] and the understanding of regulatory mechanisms.

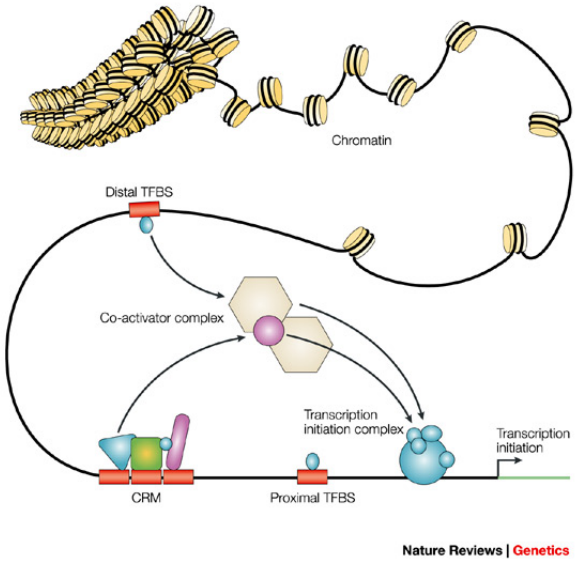


Figure 1.3: **Transcriptional Regulation**. To initiate transcription, several factors are necessary. These are part of the transcription initiation complex, consisting of Polymerase II, co-factors and transcription factors (TF). Transcription factor binding sites (TFBS) are bound by TFs, and can be far away from (distal) or close to (proximal) the TSS. The cis-regulatory module (CRM) is located on the same strand as the TSS and is bound by a complex of TFs. Figure from [*Wasserman and Sandelin*, 2004].

1.3 Transcription Factor Binding Site Detection

1.3.1 Representation of TFBS

In a simplified model, transcription can be partitioned into three steps, (1) TF binds to TFBS, (2) TF recruits polymerase II, (3) RNA is transcribed starting from the TSS. Discovery of TFBS *in silico* requires a representation of the binding site in the computer. Assuming that TFs tend to bind to similar DNA sites, a representation of a binding site collection can be built in two steps. First, a multiple sequence alignment of the TFBS is performed and second, the content of each column in the alignment is represented by the International Union of Pure and Applied Chemistry (IUPAC) notation. This representation gives a poor idea of the relative importance of each nucleotide, hence matrices have been introduced to represent the number (position frequency matrix; PFM) and probability (position weight matrix; PWM) of occurrences of the four possible nucleotides [*Stormo*, 2000]. Figure 1.4 illustrates the PFM and logo representation of a binding site. The logo or motif scales every nucleotide according to the total

number of bits of information. Sequence logos allow a fast visual estimate of the conservation of each position.

Although, Stormo *et al.* [*Stormo and Fields*, 1998] found that the calculated score agrees well with physical properties such as binding energy, other information contributing to the binding process, such as methylation and acetylation events should be considered [*Rego et al.*, 2012]. Acetylation opens the chromatin and eases the access for TFs to the DNA, whereas methylation inhibits acetylation. There are TFs that influence acetylation by recruiting acetylases.

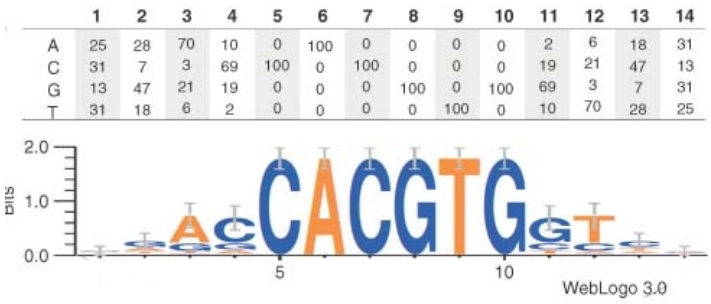


Figure 1.4: **Representation of TFBS in the computer.** A TFBS is shown represented as a position frequency matrix (a), stating the observed counts for each of the nucleotides (A, C, G or T). The frequencies are very similar in the ends of the matrix, but a clear motif evolves in the middle of the matrix. It becomes more evident when the matrix is transformed into a logo which is a visual aid facilitating the identification of the motif. Adapted from [*Kim and Park*, 2011].

1.3.2 Databases with Regulatory Elements

PWMs and sequence logos, along with annotation for every TFBS, are available in several databases. The databases differ in their contents, validation procedures and commercial nature. JASPAR [*Stormo*, 2000; *Wasserman and Sandelin*, 2004] is a database that contains motifs derived from literature. The collection is continuously extended and freely available at (http://jaspar.genereg.net/). The binding sites are not redundant and stored as PWMs and sequence logos. The TRANSFAC [*Wingender*, 2008] database (http://www.biobase-international.com) from the company Biobase is commercially available. A free version providing a limited set of motifs is available. The CisRED database contains putative TFBS [*Robertson et al.*, 2006], derived by employing neutral evolution, phylogenetic foot printing and homology based analyses. Regulon DB provides comprehensive data including, a regulatory network, binding sites and interactions between TFs for *Escherichia Coli* [*Gama-Castro et al.*, 2011].

1.3.3 Algorithms to find binding sites

Different approaches to find TFBS exist, for example using motif information (scanning tools) or *de novo* search [*Frith et al.*, 2004]. *De novo* detection of binding sites is done by comparing sequences putatively bound by the same TF with each other. These sequences are, for instance, derived from genes co-expressed in microarray experiments or from orthologous promoter sequences. *De novo* algorithms start off with an initial matrix or binding site and then iteratively refine the motif, using methods such as Hidden Markov Models, Expectation Maximization (EM) or Neural Networks. The first algorithm for the detection of binding sites has been introduced in 1985 by Galas *et al.* [*Galas et al.*, 1985]. Among the most popular methods for *ab initio* detection of regulatory motifs are Gibbs-sampling [*Casella and George*, 1992] and MEME (Multiple EM for Motif Elicitation) [*Bailey and Elkan*, 1995]. Other, more recent *de novo* algorithms include AnnSpec [*Workman and Stormo*, 2000] or Dispom [*Keilwagen et al.*, 2011].

Scanning tools use libraries of PWMs such as TRANSFAC or JASPAR. The underlying principle of tools that make use of PWMs is that they move the PWM along the DNA sequence and calculate a score for each position. The same procedure is repeated for background sequences (intergenic regions, CpG islands, etc.). The scores are then compared and the motif is evaluated to be significant or not. Often, sequences with enriched CpG islands (regions rich in CG dinucleotides) are used as a second reference sequence, because in these regions histones are acetylated. Tools that use PWMs to derive the presence of motifs include STORM [*Schones et al.*, 2007], TOUCAN [*Aerts et al.*, 2003], MotifViz [*Fu et al.*, 2004] or PEAKS [*Bellora et al.*, 2007].

1.4 High-throughput Gene Expression Profiling

1.4.1 Batch Effects

Experimental Design

Back in 1935 Ronald A. Fisher introduced a number of principles that should be applied when designing an experiment [*Fisher*, 1935]. These principles have also been suggested to be applied to high-throughput genomics experiments as technical variability and non-biological biases can be introduced when generating the data [*Kerr and Churchill*, 2001]. Microarray and RNA-seq experiments are used to identify altered gene expression profiles between treatment groups of interest.

By considering the three concepts of experimental design, Randomization, Replication and Blocking, confounding of the outcome with technical variation can be avoided or at least systematically measured and later adjusted or removed.

1. **Randomization.** Samples are randomly assigned to groups to reduce bias and to create homogeneous treatment groups without prior judgment/knowledge.
2. **Replication.** Experiments should be replicated to estimate variation and uncertainty in order to reliably measure the biological effect.
3. **Blocking.** The assignment of samples into groups or blocks (batches) is termed blocking. It reduces systematic variation that is not due to the treatment.

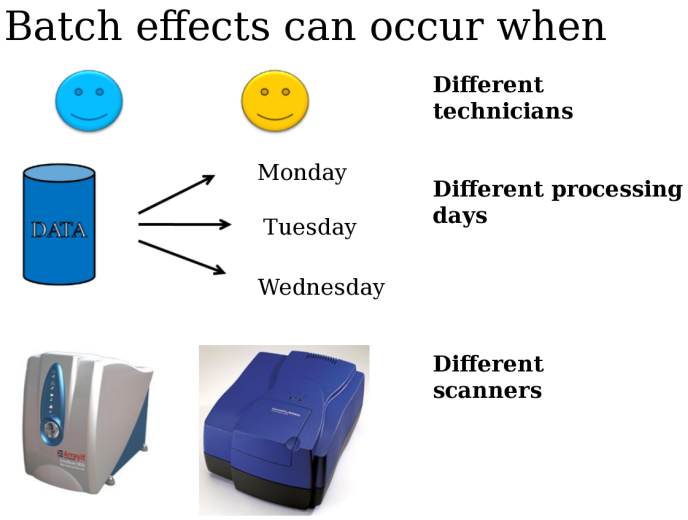


Figure 1.5: **Different sources of batch effects.**

In the high-throughput gene expression setting, a batch is defined as a set of microarrays grouped together for processing and the so-called 'batch effect' is a systematic bias introduced in a biological experiment. Batch effects are not limited to microarray experiments but also occur in other experimental designs [*Leek et al.*, 2010], however, most methods detecting and correcting batch effects are for microarrays. The main drawback is that biological information is influenced by technical information and consequently the results are confounded with the research question. Batch effects in microarrays have already been reported with the emergence of the first microarray experiments [*Lander*, 1999]. Typical

sources for batch effects are the laboratory where the data have been processed, the scan date or the personnel preparing the chip [*Leek et al.*, 2010] (Figure 1.5). Often, the microarray is hybridized as the samples come in, meaning that the period of finalizing the experiment can stretch over several years [*Benito et al.*, 2004]. These sources lead to increased variability and decreased power to detect the real biological signal.

Dealing with Batch Effects

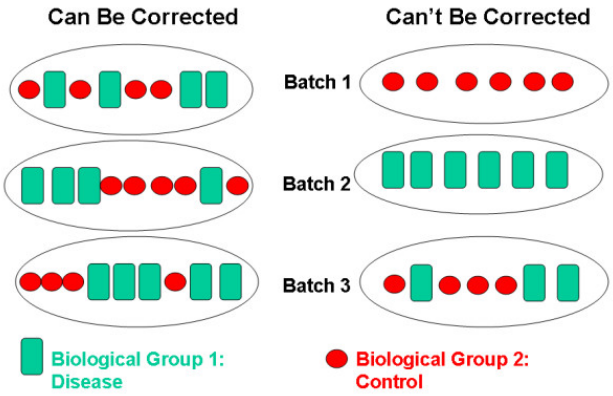


Figure 1.6: **Experimental design.** On the right side an experimental design is shown for which it is impossible to remove the batch effect. The first two batches contain one biological group per batch, and the third batch has a mix of the two groups. This distribution of samples is not enough to estimate whether the observed effect is due to the differences in the groups or the day the samples have been analyzed. The experimental design on the left can be corrected because a balanced number of samples has been assigned to each batch. Figure from [*Walker et al.*, 2008].

Many published studies are influenced by batch effects because their confounding factors have not been accounted for [*Chen et al.*, 2011]. As Baggerly *et al* point out: "Batch effects are common in large-scale expression studies, but are not commonly addressed" [*Baggerly et al.*, 2008]. It is imperative to avoid batch effects confounded with the outcome of interest. The right study design has to be chosen, such that the confounding factor can be removed. Figure 1.6 illustrates an experimental design in which the batch effect can be removed and a second design in which the batch effect cannot be removed. Optimally, the samples of each phenotype (e.g. case, control) are distributed equally into batches.

But even a perfect design will be affected by technical variation. Usually, arrays are normalized before the actual analysis begins, but normalization methods are not capable of removing batch effects [*Qiu et al.*, 2005; *Johnson et al.*,

2007; *Leek et al.*, 2010]. As explained earlier, RMA, the most used normalization method for Affymetrix chips, equalizes the distributions of the raw measurements for each array. But this is not accounting for batch effects as Figure 1.7 illustrates. The raw expression data are adjusted to follow the same distribution (Figure 1.7 b). Upon clustering, the samples show a clear pattern confounded with the scan dates (Figure 1.7 d). To actually remove the bias, batch effect removal methods have to be consulted. These either remove the bias by applying some statistic to the data or they adjust for the batch effect.

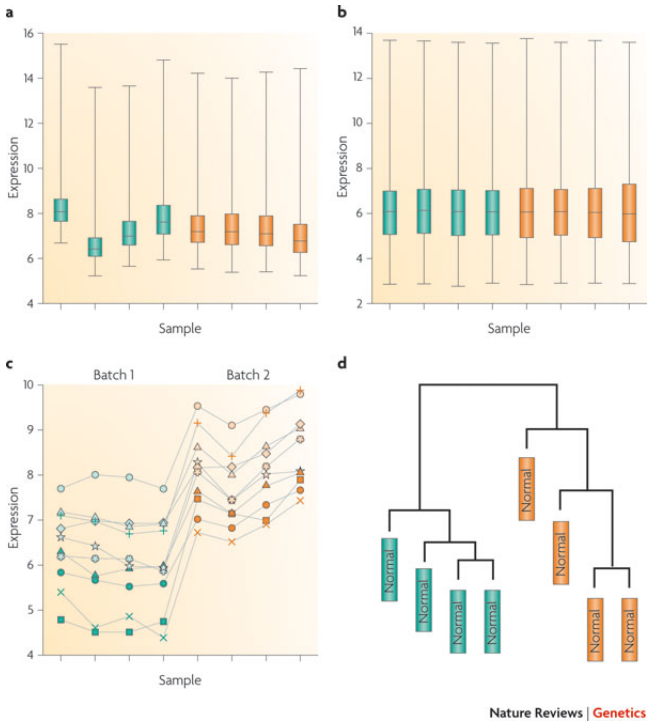


Figure 1.7: **Batch effect surviving normalization.** Boxplots of gene expression from healthy samples are shown. The color code represents the different processing dates of the samples. a) The expression levels are shown for each sample before normalization and b) after normalization with RMA. The data is normalized such that it follows the same distribution. However, when looking closer at individual genes c) or upon clustering d) of the samples, the batch effect becomes visible [*Leek et al.*, 2010].

Methods that can be used to address batch effects include single value decomposition (SVD) [*Alter et al.*, 2000], distance weighted discrimination (DWD) [*Benito et al.*, 2004], Empirical Bayes (ComBat) [*Johnson et al.*, 2007], and surrogate variable analysis (SVA) [*Leek and Storey*, 2007]. SVD and DWD require at least 25 samples within each batch to work properly. Both methods try to identify variables that explain the variation resulting from batch effects. DWD

finds a hyperplane separating two batches and then takes the mean of the batches which is then subtracted to correct for batch effect. Potential drawbacks of these methods are the number of samples required (25 samples per batch), biological variation may be confused with batch effect and DWD only works by pairwise analysis. To address these limitations the empirical Bayes method has been introduced [*Johnson et al.*, 2007], which can be used to remove batch effects on smaller sample sizes. Chen and colleagues compared six batch effect removal methods and found that ComBat outperforms the remaining five [*Chen et al.*, 2011].

Often, the technical variable influencing the gene expression measurement is not known [*Leek and Storey*, 2007; *Leek et al.*, 2010] or the data have not been recorded, for example the different technicians preparing the data. Thus, if the time stamp is not enough to remove non-biological variability, SVA [*Leek and Storey*, 2007] can be used to identify these hidden variables [*Irizarry et al.*, 2009]. The variables can then be used as main effects in a linear model for differential gene expression analysis, thus, taking the batch effect into account and adjust for it during the analysis to avoid confounding.

1.4.2 Microarrays

A microarray is a slide with a collection of DNA spots that are used to measure gene expression levels on a large scale. The concept of microarrays emerged in 1991 [*Fodor et al.*, 1991] and the first microarray was designed in 1995 [*Schena et al.*, 1995]. Over the last decades microarray data have been successfully used to identify gene signatures, detect important biomarkers and aid in cataloging the diverse molecular patterns underlying biological and physiological processes [*Hoheisel*, 2006; *Gresham et al.*, 2008; *Perou et al.*, 2000]. Different types of arrays have since been developed, each specializing on a certain biological aspect. These arrays include array-comparative genomic hybridization (aCGH) [*Lucito et al.*, 2003] to study copy-number variations, SNP (single nucleotide polymorphisms) arrays to detect polymorphisms [*LaFramboise*, 2009] in a sample population and ChIP-on-chip (ChIP: Chromatin-Immunoprecipitation) to find interactions between proteins and DNA [*Iyer et al.*, 2001].

The first gene expression arrays used a two channel technology (also known as comparative hybridization), hybridizing two samples together, i.e, control and disease. Each sample is distinguished by a different color, green (fluorochrome Cy3) and red (fluorochrome Cy5). When genes in the sample labeled with red are over-expressed, the spot on the microarray will appear red. The same applies to genes in the other sample labeled with green, when they are over-expressed

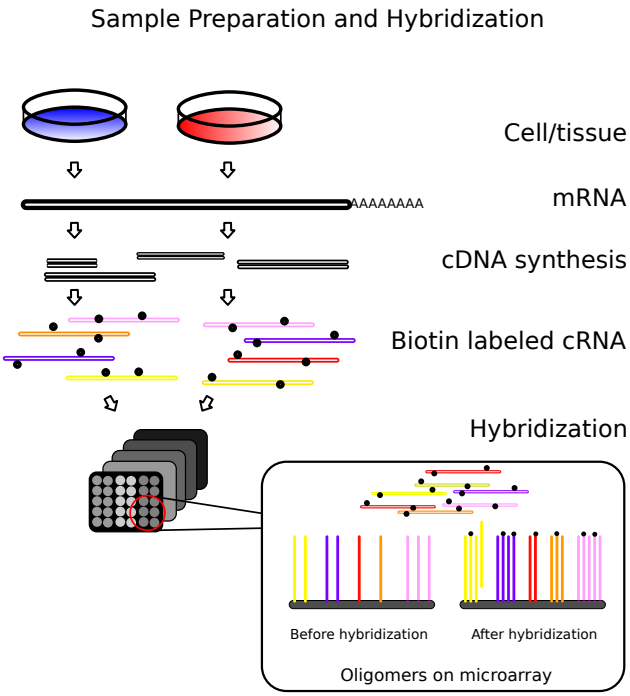


Figure 1.8: **A typical outline of a single-channel microarray experiment.** The sample preparation process begins with the isolation of RNA from cell lines grown under different conditions or tissue from different patient groups. Next, the mRNA is isolated and reverse transcribed into complementary DNA (cDNA). During *in vitro* transcription biotin labeled cRNA is produced from the cDNA. The cDNA is fragmented and applied onto the array. The fragments bind to the cDNA fragments on the array, this hybridization process is then detected by a laser. Relative RNA expression is measured by comparing the same gene between the different arrays that are used for the different conditions. Figure adapted from [*Schulze and Downward*, 2001] and [*Bolstad*, 2004].

the spot appears green. If the ratio of gene expression between both samples is the same, the spot appears orange or yellow.

After dual channel, single channel microarrays were introduced. Single channel microarrays use one color (fluorochrome), hence, each sample is applied on a separate chip. The expression ratio of the chips hybridized with the different samples are then calculated to estimate gene expression. One of the most popular microarray platforms are high-density oligonucleotide arrays produced by Affymetrix [*Lockhart et al.*, 1996] (single channel). Also, the data used throughout this thesis was measured on Affymetrix chips. Hence, we discuss the underlying technology further.

Microarray experiments consist of several steps, which can be roughly divided into three parts, the actual microarray experiment, the low-level pre-processing

step and the subsequent high-level analysis. In the following, the preparation of the microarray experiment and low-level pre-processing steps are outlined. The high-level analysis includes clustering, differential gene expression analysis, over-representation analysis (e.g. Gene Ontology enrichment or Gene Set Enrichment (GSE) Analysis) and classification. GSE analysis is discussed in a later chapter.

Microarray Experiment

An Affymetrix microarray contains millions of oligonucleotides (probes). The whole of probes interrogating a gene is referred to as probe set or feature. The design of the chip and its collection of probes follows a specific annotation of genes on the genome that was available when the chip was manufactured. The probes are designed to be representative of the genome and refer to a part of a gene. For each gene several probes are available interrogating the gene.

Figure 1.8 highlights the sample preparation and hybridization. To begin the experiment, RNA is extracted from the samples and reverse-transcribed into complementary DNA (cDNA), which is then mixed with biotin for labeling. The prepared cDNA fragments are given onto the array and bind to their respective counter probes (pairing or hybridization). Next, the array is washed with fluorescent dye binding to the biotin and washed again to remove the non-binding RNA. Arrays are placed into a scanner, in which a laser shines light onto the array causing the fluorescent to excite. The excitation is measured and its intensity is proportional to the amount of RNA that binds to the probe. From this image the Affymetrix software computes the probe intensities. The intensities build the basis for most of the gene expression analysis workflows.

Pre-processing of Microarrays

The pre-processing of the measured raw data comprises several steps, image quantification, quality control, background correction, normalization and summarization. The pre-processing step is very important, because technical noise can influence the results of the analysis.

Quality Control is carried out on the probe level and can be done using AffyPLM [*Bolstad*, 2004] and the Relative Log Expression (RLE) or Normalized Unscaled Standard Error (NUSE) diagnostics. Chips that do not pass the quality controls are removed from the analysis. AffyPLM fits models on probe set level to identify the chips of lower quality. RLE is calculated by subtracting the median value for each gene over the sample population from the expression estimate for a gene on an array, thus comparing probe expression on each array against the

median expression across all arrays. NUSE is calculated by dividing the standard error of the gene expression value of a gene on an array by the median of all the standard errors per array, thus standard error estimates are obtained for each gene and standardized across arrays.

One of the most popular methods are the MAS 5.0 algorithm [*Irizarry et al.*, 2003a] and the Robust Multichip Average (RMA) algorithm [*Irizarry et al.*, 2003b, a; *Bolstad et al.*, 2003]. RMA background corrects, normalizes and summarizes the raw intensities and scales the expression values to fall on a proper scale. Background correction adjusts for cross hybridization resulting from non-specific binding of the fluorophore on the array. Normalization gives the same distribution to each chip. This is done by sorting the values per sample in ascending order. Next, the average of each gene across samples is calculated. The value for each gene is then replaced by the average value. Then, the values are brought back to the original order for each sample. Finally, the data are scaled by setting the same mean value to each chip (quantile normalization [*Bolstad et al.*, 2003]).

Summarization is necessary to compile the individual measurements into one signal. The probes match to a part of the sequence of a gene and are grouped into probe sets on the chip. To obtain the expression value, the probe sets are combined into one signal using one or multiple chips depending on the method. Single chip approaches include, the calculation of the median or mean of the background corrected and normalized probe intensity values. RMA uses an additive multi-chip approach, because the same probe sets respond similarly over different chips. Different probes on the same chip have a higher variability than the same probe on different chips [*Zilliox and Irizarry*, 2007]. The multi-chip model includes probe and chip response parameters, to account for the probe effect as well as the relationship of concentration and gene expression [*Bolstad*, 2004].

1.4.3 RNA-seq

Next generation sequencing of RNA (RNA-seq) [*Mortazavi et al.*, 2008] allows to study transcriptomes at a great depth. RNA-seq is not only used to measure transcript levels in an experiment, but also to study the transcriptome in more detail. For example, RNA-seq allows for the more comprehensive study of alternative splicing [*Sultan et al.*, 2008], host and pathogen interactions [*Mandlik et al.*, 2011], and has helped to characterize the importance of fusion genes in cancer [*Mitelman et al.*, 2007]. Several technologies (sequencers) are available, including those developed by 454 Life Sciences (Roche) [*Margulies et al.*, 2005] and Illumina (formerly Solexa sequencing) [*Bennett et al.*, 2005].

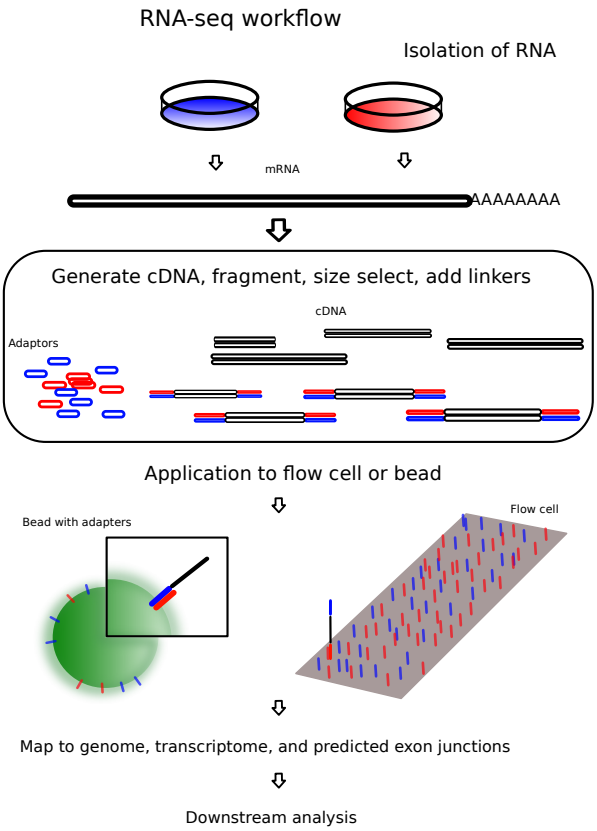


Figure 1.9: **RNA-seq workflow.** RNA is extracted from the sample cell lines or tissues and reverse transcribed into cDNA. Adapters (linkers) are added to the double stranded cDNAs, which are used to bind on the sequencing device (either flow cell or bead). Next, the fragments are binding to the complementary adapters and are amplified. Fluorescent nucleotides and polymerase are added. Nucleotides are incorporated when the polymerase transcribes the template. Upon laser exposure the nucleotides emit light that can be detected. Figure based on [*Marguerat and Bähler*, 2010] and [*Wang et al.*, 2009].

RNA-seq Experiment

A typical RNA-seq experiment consists of several steps briefly explained in the following (Figure 1.9). RNA is randomly fragmented and sheared to fit a certain length that the sequencer is able to sequence. If the fragment is too long and exceeds the maximum read length, it will not be accurately sequenced. Next, the RNA is reverse-transcribed into cDNA and adapters (linkers) are attached to the cDNA fragments. Adapters guide the fragment to the complementary adapter on the sequencing device. The sequencing device can be a bead or a flow cell, which are both covered in adapters. Then, the fragments are denaturated and poured over the sequencing device, where they are amplified. During amplification,

clusters of the same fragment are built. Each fragment is then sequenced in parallel by incorporation of fluorescent nucleotides emitting upon laser exposure. An image is recorded for each cycle, identifying the bases that bind in the clusters. Step by step each base is identified. Gene expression profiling with RNA-seq is an important task, therefore the following pre-processing outline is focused on the preparation for differential expression analysis [*Hansen et al.*, 2012].

Pre-processing of RNA-seq data

The output of all sequencers, independent of their manufacturer, are millions of short reads (25-300bp) from the cDNA (if read from both sides, the reads are called paired-end reads). The raw data (reads) have to be further processed. There are variations in the processing protocols because pipelines are still evolving. Depending on the focus of the study, for example variation detection or gene expression estimation, the pipeline changes. In the case of gene expression estimation, the reads have to be mapped to a reference genome or transcriptome and further condensed into a signal on the gene-, exon-, or transcript-level [*Oshlack et al.*, 2010]. The pre-processing steps are outlined in the following.

Read mapping. The mapping process is a core step in the processing pipeline, which is quite challenging. Reads are short and can match perfectly to several locations of the reference genome [*Trapnell et al.*, 2012] resulting in the multiple mapping problem. Aligners deal with the multiple mapping problem in different ways, either by discarding them, randomly allocating them, or by using alignment scores. Paired-end reads can overcome this problem, because both reads have to be mapped. Another challenge are sample-specific attributes such as single nucleotide polymorphisms (SNPs), insertions and deletions (indels), because each individual has specific variations in the genome. Variations can lead to complications when trying to map the reads. Gene expression estimates are made from transcripts, which adds another layer of complexity to the mapping, because the reads are obtained from a spliced transcript which does not contain intronic regions.

In addition to the biological challenges, there are also technical problems to be accounted for, such as sequencing errors. Read mapping is typically done by first using a heuristic search using either Burrows-Wheeler transformation (BWT) or hash tables. Then, the reads are aligned locally with, for example, the Smith-Waterman algorithm [*Homer et al.*, 2009]. Available aligners, partitioned according to their heuristic search algorithm, include: Burrows-Wheeler transformation: Bowtie [*Langmead et al.*, 2009], SOAP2 [*Li et al.*, 2009], Burrows-Wheeler Aligner (BWA) [*Li and Durbin*, 2009] and hash tables: BFAST [*Homer*

et al., 2009], SHRiMP [*Rumble et al.*, 2009], MAQ [*Li et al.*, 2008a], and SOAP [*Li et al.*, 2008b]. Not all of the aligners can be used for each sequencer [*Trapnell et al.*, 2012]. Aligners also differ in the type of the reads, e.g. transcript or genomic reads, that they can process into a signal. Also, some aligners need a reference genome for read mapping, whereas others map *de novo*. Further, not all aligners allow for indels or multiple mappings of reads and ignore them when mapping.

Normalization. To estimate gene expression levels, the simplest approach is to count the number of reads that overlap with the region of interest (sequencing depth). However, this method ignores any level of complexity (for instance, different isoforms) and introduces certain biases such as transcript length (the longer the transcript, the more reads will be associated with it) [*Oshlack and Wakefield*, 2009], sequencing depth (the higher sequencing depth, the higher counts) [*Tarazona et al.*, 2011], GC content bias (fragments are preferentially sequenced depending on their GC content) [*Pickrell et al.*, 2010] and differences in count distributions of the different samples under study. To address these RNA-seq biases, several normalization methods have been developed.

RPKM (Reads per Kilobase per Million) [*Mortazavi et al.*, 2008] is the first normalization method for RNA-seq. It divides the counts by the transcript length times the total number of millions of mapped reads addressing the length bias and sequencing depth problems. It has been found that sequencing depth is sample-specific, hence, more robust normalization methods have been introduced. TMM is the trimmed mean of M-values normalization [*Robinson and Oshlack*, 2010], which is based on the assumption that genes are similarly expressed throughout the sample population. Upper-quartile normalization divides the counts of transcripts with at least one read by the upper-quartile [*Bullard et al.*, 2010]. Conditional Quantile Normalization (CQN) [*Hansen et al.*, 2012] uses a generalized regression model to address systematic biases such as difference in distribution of counts per sample, gene length, sequencing depth bias, and GC content effect. The algorithm also uses quantile normalization to account for global distortions. For an overview of further normalization techniques and their (dis)advantages, the reader may be referred to [*Oshlack et al.*, 2010].

1.4.4 Microarrays versus RNA-seq

Gene expression levels are estimated in microarrays by measuring the fluorescence levels after hybridization and in RNA-seq as the number of reads mapping to a gene, exon or transcript. The value of RNA-seq lies in the larger dynamic

measurement range, the unbiased transcript discovery (sensitivity) due to the direct determination of the cDNA and the simultaneous sequencing of multiple samples. One of the main differences between microarray technology and RNA-seq is that RNA-seq does not depend on genome annotation. This advantage over microarrays allows for the detection of novel expressed regions, alternative isoforms, allele-specific expression or fusion genes. Furthermore, species that lack a reference genome can be sequenced [*Malone and Oliver*, 2011]. For splicing, splicing arrays have been used, but it has been shown recently that sequencing seems to detect exon/exon junctions more accurately [*Agarwal et al.*, 2010]. Other differences include that microarrays have biased signals due to cross-hybridization and a limited dynamic range because of the saturation of fluorescence signals [*Wang et al.*, 2009].

However, gene expression microarrays are still an instrument of choice, because protocols for target probe preparation and downstream analysis are very well established and the costs are still lower than for RNA-seq. RNA-seq lacks for the moment development of appropriate data analysis tools and poses additional algorithmic and logistic challenges [*Baginsky et al.*, 2010]. Data analysis is very complex and can be a bottleneck. Next generation high-throughput technology requires data storage and computational infrastructure [*Schadt et al.*, 2010] that are not easily available for some laboratories. Microarray data analysis, on the other hand, can be carried out on a newer laptop. Over time, the protocols and procedures for RNA-seq will be more standardized and the technology will experience a reduction in cost approaching the maturity of microarray technology for gene expression.

1.5 Gene Set Enrichment Methods

Gene expression experiments typically end with a long list of differentially expressed genes (in form of a gene ID). To interpret and organize these genes is quite a challenging task. To help with this task, many statistical tests have been proposed that condense gene expression information into sets of related genes. Gene sets facilitate reproducibility between experiments [*Manoli et al.*, 2006]. Two main approaches are available that reduce and organize genes into more interpretable units (gene sets), referred to as over-representation methods and aggregate score methods [*Irizarry et al.*, 2009].

1.5.1 General outline of GSE methods

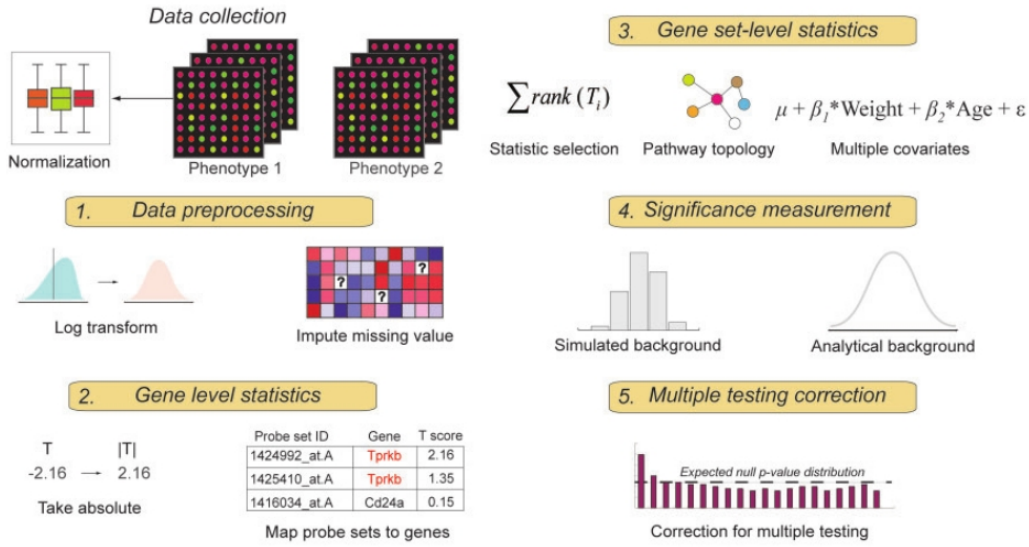


Figure 1.10: **Outline of gene set enrichment analysis.** Input are gene sets, which are formed beforehand and a ranked list of genes obtained for example from a microarray experiment. Phenotype 1 and 2 represent different experimental groups, e.g. healthy and diseased patients. 1. The microarray data are pre-processed including quality control, normalization and batch effect removal. 2. A gene level statistic is calculated to generate a ranked list of differential expression between the two phenotypes. 3. The genes are linked to the gene sets and a gene set level statistic is calculated to assign an aggregated score to the gene set. 4. The estimated scores are tested for significance compared to a background distribution, thus obtaining a p-value for each gene set. 5. The p-values have to be corrected for multiple testing. Figure source [*Hung et al.*, 2012].

The essential steps of GSE are outlined in Figure 1.10. The generalized GSE process is adapted from [*Hung et al.*, 2012] and explained in detail in the following.

1. **Data pre-processing.** The collected data have to undergo quality control checks and to be normalized using e.g. RMA. Next, the raw data have to be log-transformed to avoid bias towards high expression values and to increase statistical power. If there is a batch effect present in the data it should be identified and removed. RNA-seq data can be normalized using RPKM (Reads per Kilo base per million mapped reads) or CQN. Missing values can be imputed.

2. **Gene-level statistics.** On microarrays different probe sets can be linked to the same gene, which complicates the determination of the expression

level of the gene. One solution to this problem is that the probe set that varies the most over the sample population is assigned to the gene while the other probe sets are discarded. Taking the probe set with the highest variation ensures that the most information about the gene is captured. To calculate the differential expression value, t-test, fold change or linear models can be used. For RNA-seq data a similar problem arises when mapping the reads to a genetic region. Often the reads map to multiple regions. Strategies to map to the best gene are still discussed. There are several approaches available for differential expression analysis for RNA-seq data, such as NOISeq [*Tarazona et al.*, 2011], edgeR [*Robinson et al.*, 2010], baySeq [*Hardcastle and Kelly*, 2010], DESeq [*Anders and Huber*, 2010] or DEGseq [*Wang et al.*, 2010].

3. **Gene set-level statistic.**

   The gene set-level statistic describes whether a group of genes is able to distinguish the phenotypes under study. Two calculate the gene set-level statistic, different gene set tests are available that can also be partitioned into self-contained and competitive tests [*Goeman and Bühlmann*, 2007]. This distinction has also been termed focused and battery gene set testing [*Wu et al.*, 2010]. The main difference between both tests is the formulation of their null hypothesis. The self-contained or focused gene set test only considers genes inside the gene set and calls the gene set enriched if any of them is differentially expressed (DE). If any gene is indeed DE, the null hypothesis of no gene in the gene set being differentially expressed, is rejected. The competitive or battery test compares the DE genes inside the gene set with the DE genes outside the gene set and evaluates if the genes inside are at least as often DE as the genes outside the gene set. Focused gene set tests have an increased statistical power over battery tests, hence, they are especially used for specific gene sets that are particularly interesting for the experiment. When a gene set contains at least one gene and that gene is DE, the self-contained test would call the gene set enriched. This demonstrates the power of the test, but also its non-conservativeness. Methods for focused gene set testing are described in [*Goeman et al.*, 2004; *Tian et al.*, 2005; *Wu et al.*, 2010] and for battery testing described in [*Mootha et al.*, 2003; *Subramanian et al.*, 2005; *Dørum et al.*, 2009].

   Gene set tests usually do not include clinical data or additional information about the phenotype, such as gender. Those parameters, however, influence gene expression and can lead to a confounding result. Batch effect removal methods can be used to account for these effects. But it is more

beneficial when the information is included when calculating the gene set enrichment statistic, for example by using linear regression models [*Jiang and Gentleman*, 2007; *Wu and Smyth*, 2012]. Further, only few GSE methods account for correlations and interdependence between genes in a gene set [*Wu and Smyth*, 2012] or consider the topology of the pathway.

4. **Significance Measurement.**

   Once the gene set-level statistic or enrichment score is calculated, it has to be evaluated whether it is truly significant. To do so, a background distribution has to be estimated to which the gene set can be compared. Tian *et al.* [*Tian et al.*, 2005] describe two possible background distributions obtained by either shuffling the genes (but keeping the gene set sizes) or by shuffling the samples. Using the shuffled genes as background distribution tests if a real gene set has a more significant enrichment score than expected by chance, i.e. comparing to a randomly generated gene set. The shuffling of samples evaluates whether the gene set contributes to the phenotype or not. Usually, the shuffling of samples is employed to estimate significance because the gene set contents are preserved and the gene set/phenotype association can be estimated.

5. **Multiple test correction.**

   After a p-value is obtained for each gene set, multiple test correction should be performed to ensure that the number of false positives is controlled. The Bonferroni correction is more conservative than Benjamini and Hochberg's false discovery rate (FDR). Hence, FDR is preferred. For the corrected p-values (also called q-values) that pass a certain FDR threshold, the number of false positives can be estimated.

1.5.2 Databases containing Functional Units of Genes

Gene sets are groups of genes that share a certain property, such as a molecular function or take part in the same cellular component [*Irizarry et al.*, 2009]. Gene sets are derived from published information about biochemical pathways or co-expression in previous experiments. Several sources to obtain gene sets are available, including the Molecular Signatures Database (MSigDB) [*Subramanian et al.*, 2005], the Kyoto Encyclopedia of Genes and Genomes (KEGG) [*Ogata et al.*, 1999; *Kanehisa and Goto*, 2000], BioCarta (http://www.biocarta.com/) and Reactome [*Joshi-Tope et al.*, 2003]. The Gene Ontology (GO) project [*Ashburner et al.*, 2000] provides groups of genes that are involved either in the

same molecular function, the same biological process or the same cellular component. The gene set and the GO ontology databases are different. The gene set databases relate genes by a common property in no specific order, whereas the gene ontology project represents gene sets (GO terms) as directed acyclic graphs (DAG) with parent-child relationships. The categories are organized in a hierarchy, from more general terms to more specific terms. GO terms are derived from different sources, including experimental, computational or unknown [*Rhee et al.*, 2008].

1.5.3 Over-representation Methods

The first generation of condensation methods for microarray experiments are so called over-representation methods. Originally GO terms were used for such an analysis, but any gene set database can be used. A list of GO enrichment tools can be found here: `http://www.geneontology.org/GO.tools_by_type.term_enrichment.shtml`. Typically a hyper-geometric test, binomial distribution or chi-square test are used to assess whether a gene set is enriched in a given set of selected (e.g. differentially expressed) genes. The genes are selected based on their passing of a certain cut-off applied to a gene expression experiment. Thus, only genes that are differentially expressed are considered in the analysis [*Irizarry et al.*, 2009; *Goeman and Bühlmann*, 2007]. Only genes that are differentially expressed are included in the analysis and hence pathways with moderate changes are ignored. Also, the value (e.g. gene expression measurement, fold change) associated to each gene does not contribute to the resulting score, assuming that genes contribute equally and independently [*Khatri et al.*, 2012].

1.5.4 Aggregate score Approach

Over-representation methods assume independence of genes, do not take values associated to genes (such as their fold-change) into account and ignore genes that are moderately changing. To overcome these shortcomings, aggregate score methods have been introduced. There are over 60 aggregate score or enrichment methods [*Merico et al.*, 2010]. Gene set enrichment (GSE) analysis is a widespread method, which condenses information of several gene expression profiles into a pathway with respect to a phenotype. Thus, emphasis is given to grouped alterations rather than alterations of individual genes.

For this, a ranking of genes has to be obtained, typically derived from a microarray experiment that studies gene expression changes between two groups.

The genes are mapped into the pre-defined gene sets and their gene expression profile is summarized into one enrichment score for each gene set. The seminal work of [*Mootha et al.*, 2003; *Subramanian et al.*, 2005; *Tamayo et al.*, 2011] uses the Kolmogorov-Smirnov statistic, which walks along the gene list and increases if a gene is in the gene set and decreases if a gene is not in the gene set. The maximum deviation from zero is then the enrichment score for a particular gene set. Lastly, the significance of the enrichment score is assessed by comparing it to enrichment scores derived from randomizing sample labels. Many variations to the original method exist that differ in their statistical methodology to assess gene set enrichment. These methods use for example z-statistic [*Irizarry et al.*, 2009], mean of gene expression [*Choi and Kendziorski*, 2009] or principal component analysis and other measures [*Jiang and Gentleman*, 2007].

1.5.5 Comparison of GSE Methods

One simple, but very popular and widely used, aggregate score statistic is the z-statistic. To calculate the z-statistic, a ranked list of genes describing the difference in expression between different sample groups is required (e.g. case and control). A subset of these genes is linked to a pre-defined gene set. The z-statistic evaluates the difference between the expression change of genes in the gene set compared to the remaining genes in the ranked list. This shift in mean approach has been suggested by [*Tian et al.*, 2005] and has been used subsequently with differing strategies to evaluate the significance [*Creighton*, 2008; *Lee et al.*, 2008]. The major drawback of the z-statistic is that the correlation of genes are not taken into account, which might lead to an increased number of false-positive gene sets [*Tamayo et al.*, 2011]. Other methods addressing this problem have been developed, also quantifying the difference in expression between genes within a gene set and genes outside a gene set [*Tenenbaum et al.*, 2008; *Wu et al.*, 2010; *Jung et al.*, 2011; *Subramanian et al.*, 2005; *Tamayo et al.*, 2011].

All these GSE methods require exactly two groups to be able to calculate the difference in expression of the genes and condense this information to derive significant gene sets. Sometimes, a more complex phenotype, such as time course data or cancer data without control samples, is available and the case-control approach is not sufficient to obtain interesting gene sets. For this reason, a single sample enrichment method has been developed [*Barbie et al.*, 2009]. This method estimates independently for each sample the empirical cumulative density for genes that are inside the gene set and compares it to the genes that are outside the gene set. The result is an enrichment score for each gene set and individual sample. This approach has been used successfully, for example, to infer

subtypes in brain cancer [*Verhaak et al.*, 2010]. The drawbacks of this method include neglecting to exploit gene variation over the sample population and its sensitivity to gene-specific effects [*Zilliox and Irizarry*, 2007], which change the baseline expression.

ASSESS [*Edelman et al.*, 2006] addresses the short-comings of the single sample GSE method. For each sample and gene set, ASSESS calculates a gene set enrichment statistic based on the variation inherent to the sample population. ASSESS has been applied to different biological questions, including the analysis of a time-course series [*Andrechek et al.*, 2008], which demonstrates the use of a method that is able to capture gene set changes for each sample representing a time point. However, ASSESS also only works in the case-control paradigm and cannot be applied to more complex data such as in The Cancer Genome Atlas (TCGA) (`http://cancergenome.nih.gov/`) project where large patient cohorts with multiple phenotypes are currently being profiled.

1.5.6 GSE for RNA-seq

Since traditionally GSE methods are designed to analyze microarray data, there is no readily available method that is suited to assess GSE from RNA-seq experiments. The major challenge lies in the discrete nature inherent to RNA-seq data, which precludes the direct application of GSE methods designed for continuous microarray data. Recently, a method for GO enrichment analysis for RNA-seq (GOseq) data has been developed [*Young et al.*, 2010], which can also be used with user defined gene sets or the MSigDB collection. GOseq is used, in principle, in the same manner as over-representation methods for microarrays but with the difference that it accounts for the length bias present in RNA-seq data. GO terms containing an abundance of long transcripts will be called enriched although they might not be relevant to the phenotype. GOseq works by first identifying a list of differentially expressed genes, which are then mapped into their corresponding GO terms. For each GO term the significance of being overrepresented is calculated based on the comparison of differentially expressed genes within the term and the genes outside of the term. Genes that are not considered as differentially expressed are removed from the analysis. Therefore, subtle, but coordinated changes, especially in phenotypes that are similar in their molecular profile, are not detected by this type of enrichment analysis.

1.6 Network Biology

1.6.1 History of Graph Theory

A network can be represented as a labeled graph G and is formed by an ordered pair G=(V,E) of vertices V or nodes, and edges E or links [*Barabasi and Oltvai*, 2004]. The graph is represented by an adjacency matrix, also called connection matrix. Figure 1.11 shows on the left panel a representation of a network in graph form and on the right side the representation of this graph as adjacency matrix. Each row and column combination represents a different node. If two nodes form an edge in G, the cell in the corresponding row and column in the matrix will be filled with a 1 and else with a 0. A node that has a self loop is marked by a 1 in the diagonal, otherwise the diagonal contains only zeros.

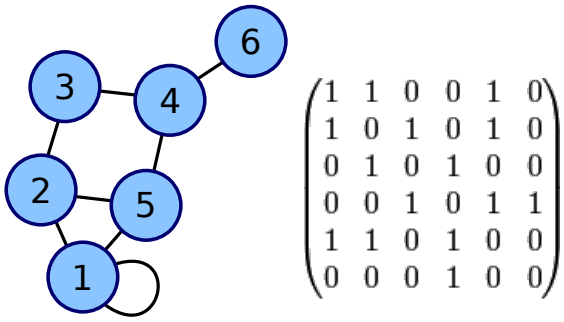


Figure 1.11: **Graph representation.** Representation of a graph in matrix format (Figure from: `http://en.wikipedia.org/wiki/Adjacency_matrix`).

The fundamentals of graph theory have been laid out by Euler [*Euler*, 1736] and the Königsberg problem. The seven bridges of Königsberg and the landmarks they are connecting are represented as a graph. The bridges serve as edges and the landmarks are nodes. The idea was to mathematically describe, how to cross all the seven bridges (Figure 1.12) without using a bridge more than once and omitting any bridge at the same time. Additionally, the return and starting point have to be the same. Euler proved that this is not possible. Landmarks with an uneven number of bridges enforce that the walker is not able to return without using one bridge twice. But when the landmark is connected with an even number of bridges, then the walker can return to the starting point by using different bridges. Mathematically this can be explained by the node degree. The degree or connectivity of a node is described as the number of links the node has to other nodes [*Barabasi and Oltvai*, 2004]. Euler showed that the possibility of walking through the graph depends on the node degree. The conclusion of Euler's work is that "a graph has a path traversing each edge exactly once if exactly two vertices have an odd degree" [*Euler*, 1736].

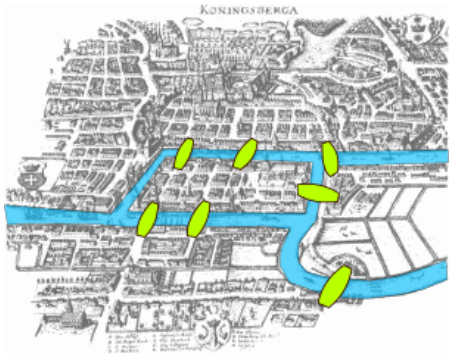


Figure 1.12: **Illustration of the Königsberg Problem.** It is not possible to go over all bridges using each bridge only once (Figure source: `http://en.wikipedia.org/wiki/Seven_Bridges_of_Koenigsberg`).

1.6.2 Definition of a Network

A network is a good model to understand the biological function of a cell and to illustrate the underlying interactions. A network describes relationships among entities to understand their functions and properties at once [*Arrell and Terzic*, 2010]. The entities are linked when they show an interaction. Such entities can be genes whose expression levels are correlated with each other because of coordinated changes in expression [*Eisenberg et al.*, 2000] or proteins that interact physically. Depending on the kind of biological network, the edges and nodes represent different molecules in the cell, for example, in a protein-protein interaction network the nodes are proteins and the edges are the interactions of these proteins. In a transcriptional regulatory network, the nodes are genes and proteins and the edges represent a transcription factor regulating a gene. Different kinds of biological networks help to answer different questions and can be distinguished at the molecular level. They include cell signaling networks [*Ma'ayan et al.*, 2005], protein-protein interaction [*Bossi and Lehner*, 2009; *Giot et al.*, 2003], pathway cross-talk [*Li et al.*, 2008c], transcriptional regulatory networks [*Basso et al.*, 2005], and gene-disease networks [*Goh et al.*, 2007].

1.6.3 Network Models

There are two main types of graph, directed or undirected graphs. Undirected means that the edges between two nodes do not indicate a direction, whereas directed means that the edges indicate a direction. This direction represents the flow of information in the graph, for example, a transcription factor regulates a gene. Nodes in a directed graph have an out-degree and an in-degree. The in-and

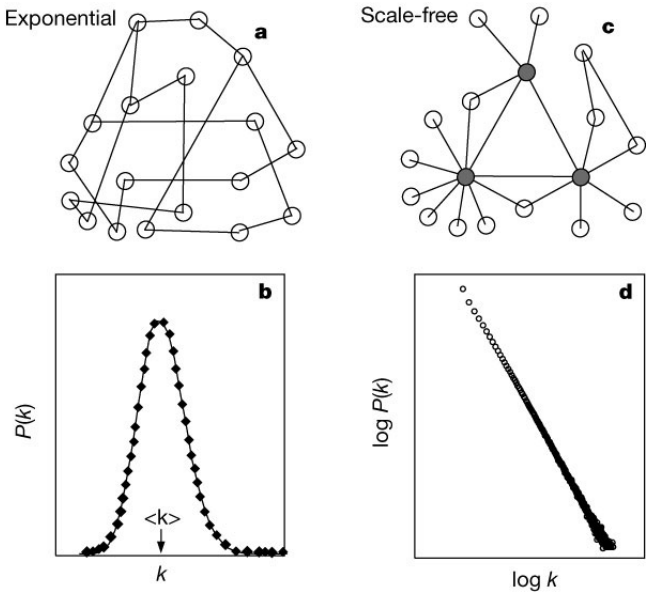


Figure 1.13: **Properties of random (exponential) and scale-free networks.** a) Shown is the typical structure of a random (exponential) network in which each link has the same likelihood to be present by chance than any other link. b) The probability of connection between any two nodes in random networks follows a Poisson distribution. c) Scale-free networks have few nodes that are highly connected and many nodes that have only few connections. d) The probability distribution decays into a power law distribution. k is the number of connections and γ is the slope. Taken from [*Jeong et al.*, 2000].

out-degrees describe the number of edges pointing at the node and the number of edges pointing away from the node. The degree of a node is used to calculate the degree distribution [*Diestel*, 2010]. For this, the number of nodes with the same degree are counted and divided by the total number of nodes. Thus, when there are n nodes in total in a network and nk of them have connectivity of k, the distribution can be calculated as $P(k) = nk/n$ [*Albert*, 2005]. The degree distribution is a probability distribution of the degrees over the whole network and can be used to classify a network.

The first description of the mathematical properties of networks were made by Pál Erdős and Alfréd Rényi [*Erdoes and Renyi*, 1960]. They modeled networks as random networks assuming that everything is connected by chance, that is, each node has the same likelihood of being connected to another node. There is a maximum number of possibilities of forming the graph and one is chosen at random (Figure 1.13 a). Hence, the node degree distribution $P(k)$ has a uniform character and follows the Poisson distribution [*Jeong et al.*, 2000]. Most of the node degrees k are approximately equal to the average degree $< k >$ of all nodes in the graph (Figure 1.13 b).

In the late 1990's Barabási and Albert found that many real networks are scale-free, that is, most of the nodes have very few links, whereas very few nodes (hubs) have many links (Figure 1.13 c). Scale-free describes that there is a high diversity in node degrees and no node alone can be used to characterize the network [*Albert*, 2005]. Scale-free networks have the small world property, which means that any path between two nodes is short. The degree distribution $P(k)$ (proportion of nodes that have connectivity k) for a scale-free network follows approximately a power-law distribution $P(k) \approx k^{-\gamma}$ with γ between 2 and 3 [*Barabasi and Albert*, 1999; *Milo et al.*, 2002] (Figure 1.13 d).

1.6.4 Organizing Principles

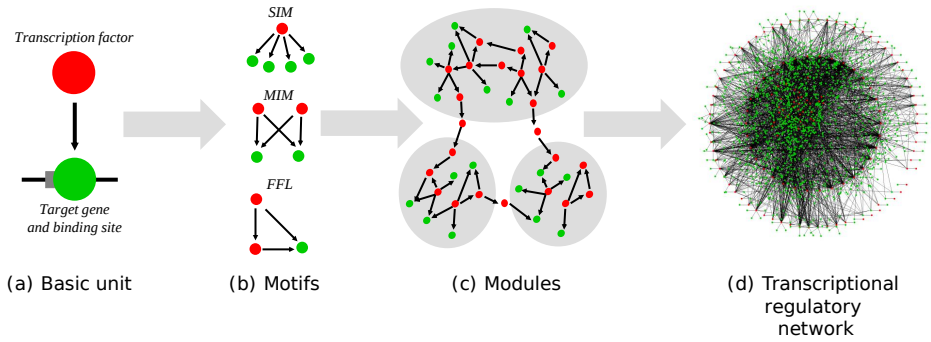


Figure 1.14: **Topological properties of scale-free networks.** a) A basic interaction of two nodes representing, for example, a transcription factor regulating a target gene. b) Recurring, small circuits between transcription factors and target genes are either feed-forward loops (FFL), single-input motifs (SIM) or multiple input motifs (MIM, dense overlapping regulons) also called bi-fan. c) Hubs are highly connected nodes in a biological network. d) The typical hairball representation of a network, in this case a transcriptional regulatory network. Taken from [*Babu et al.*, 2004].

Many Biological networks exhibit certain properties derived from network theory, which make it possible to identify modules or subnetworks that are involved in a certain disease [*Loscalzo and Barabasi*, 2011]. In the following the properties are explained. The most basic interaction in a network is between two nodes (Figure 1.14 a). This can, for example, represent a transcription factor regulating a target gene. Further, there are occurring patterns (motifs) in complex networks with predictable functional consequences [*Milo et al.*, 2002]. Recurring, small circuits between transcription factors and target genes form the simplest motifs of network architecture [*Babu et al.*, 2004; *Ma'ayan*, 2009] (Figure 1.14 b). These circuits are made of three nodes that are either feed-forward loops (FFL) [*Milo et al.*, 2002], single-input motifs (SIM) or multiple input motifs

(MIM, dense overlapping regulons) also called bi-fan [*Milo et al.*, 2002]. FFL have the property of acting as filters for transient signals. SIM are for example one TF regulating many genes and MIM can be a group of genes sharing a small number of regulators [*Ma'ayan*, 2009]. Basic units that are highly connected are also referred to as hubs or modules (Figure 1.14 c). Hubs are highly connected nodes in a biological network and most of them are essential. Removal or modification of hub nodes *in utero* leads to embryonic lethality [*Loscalzo et al.*, 2007; *Barabasi et al.*, 2011].

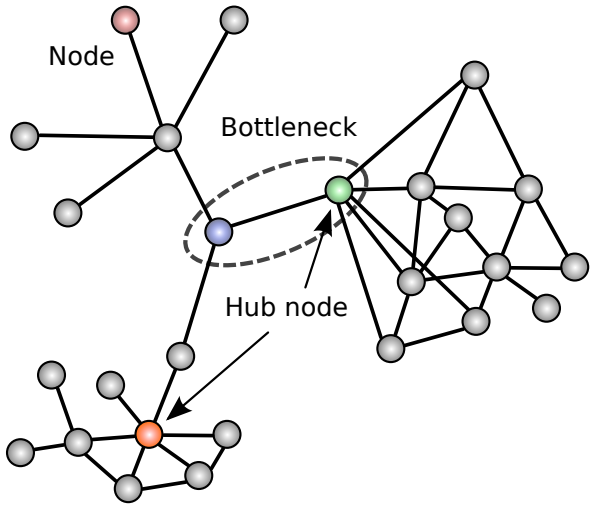


Figure 1.15: **Bottleneck principle**. The principle of bottlenecks is illustrated here. Four different node types can be distinguished, red is a peripheral node, orange is a hub node, blue and green form together the bottleneck, but only the green node is also a hub node. Nodes with a high betweenness (number of shortest paths crossing them) are called bottlenecks. Figure adapted from [*Yu et al.*, 2007].

To measure the centrality of a node, that is, how influential that node is in the network [*Freeman*, 1977], Freeman defined in 1977 three different measures for centrality, degree, closeness, and betweenness. The shortest paths [*Dijkstra*, 1959] between all node pairs is used to measure the farness [*Sabidussi*, 1966] and closeness of the node. Farness is the sum of all the shortest paths (i.e. the sum of the distances to every other node) and closeness is the reciprocal of farness (i.e the inverse of the sum) [*Opsahl et al.*, 2010]. The number of shortest paths that go through a certain node from any other node, is called betweenness [*Opsahl et al.*, 2010]. Nodes with a high betweenness are so called bottlenecks (Figure 1.15) [*Yu et al.*, 2007]. In biological networks, bottlenecks connect different hubs and thus control the information flow [*Loscalzo and Barabasi*, 2011]. Bottleneck nodes are also of importance to identify key information flow directors, which could be potential drug targets [*Yu et al.*, 2007].

1.6.5 Disease pathways

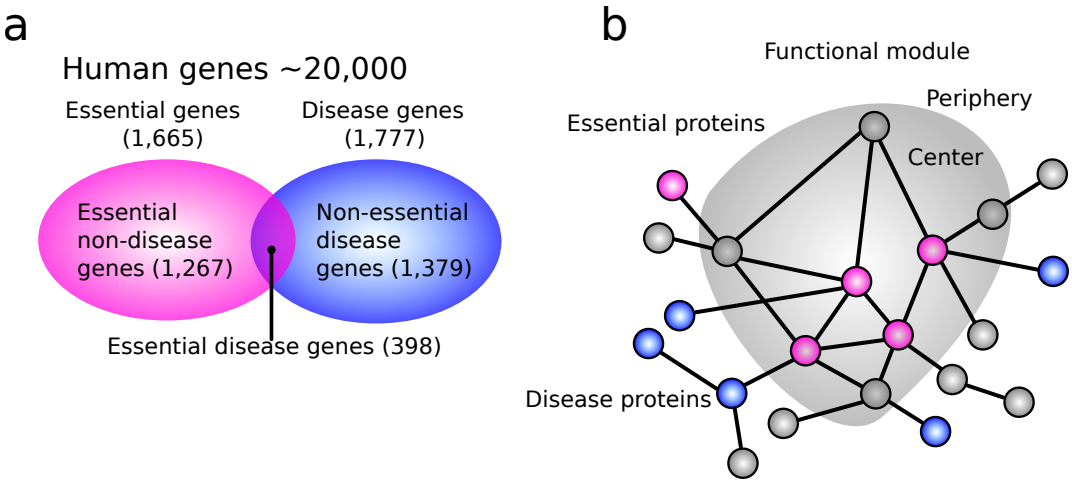


Figure 1.16: **Classification of disease genes**. a) Of the about 20,000 human genes, 1,665 are essential, that means, if they were mutated it would be fatal. 1,777 genes are non-essential but they are disease genes. The overlap of essential genes that are also disease genes comprises 398 genes. This demonstrates that essential genes are less likely to be disease driving genes. b) In a network, disease genes encoding disease proteins are not in the center of a functional module but rather in the periphery. The red nodes represent essential proteins and the blue nodes disease proteins. Essential proteins tend to be in the center of a module. Taken from [*Barabasi et al.*, 2011].

Most biological networks exhibit a scale-free topology. This topology is advantageous when it comes to inclusions and deletions of nodes (proteins or genes) as it happens for example during evolution when genes are duplicated or during alternative splicing events. It has been shown in several model organisms that the system is still viable upon deletion or inclusion of new connections or proteins [*Isalan et al.*, 2008; *Pastor-Satorras and Vespignani*, 2002]. In yeast and human it has been observed that hubs are encoded by essential genes and that these genes are expressed in many tissues [*Jeong et al.*, 2001; *Albert et al.*, 2000; *Han et al.*, 2004; *Goh et al.*, 2007].

The deletion or mutation of hub proteins or non-hub proteins has different consequences. Modification of hub proteins can lead to lethality whereas modification of non-hub proteins creates variation [*Loscalzo et al.*, 2007]. Thus, hub proteins are essential for the survival of the organism while the organism is able to live with mutations in non-hub proteins. Most of the disease driving genes are non-essential genes located in the periphery of a complex network [*Goh et al.*, 2007; *Hopkins*, 2008; *Barabasi et al.*, 2011; *Loscalzo et al.*, 2007]. Figure 1.16a shows the distribution of essential genes, disease genes, essential non-disease

genes and non-essential non-disease genes, while Figure 1.16b illustrates the notion of the different classes of the aforementioned node types and their positions in a functional module.

Disease driving genes tend to group together and reside in the same neighborhood [*Goh et al.*, 2007]. This grouping of disease genes is also called module, pathway, gene set or subnetwork. The hierarchical connectivity structure of biological networks accounts for the high modularity and interconnections of genes within the same cluster [*Barabasi and Oltvai*, 2004]. Genes in the same module contribute to a disease phenotype when one or more members of the module are dysfunctional [*Loscalzo and Barabasi*, 2011; *Park et al.*, 2009]. The principle of disease driving modules is based on studies of protein interaction networks or genome-wide networks that are robust against random deletions or inhibitions of individual nodes [*Hopkins*, 2008]. The single gene knock-out is not affecting the phenotype, however, the simultaneous knock-out of two genes results in lethality or sickness (synthetic lethality and synthetic sickness) [*Ooi et al.*, 2006]. Multiple proteins have to be targeted to achieve an effect on the disease phenotype. These observations imply that it is of importance to target a disease module and the non-essential genes therein and be careful that essential genes are not affected as this can lead to severe side effects [*Zhu et al.*, 2008].

The identification of disease pathways has the goal to pinpoint to the disease driving genes and to find potential drug targets. The extensive analysis of a pathway in breast cancer revealed that the "gang of four", (EGF-R) ligand eregulin (EREG), the cyclo-oxygenase COX2 and the matrix metalloproteases 1 and 2 (MMP1, MMP2), is essential for lung metastasis [*Eltarhouny et al.*, 2008; *Gupta et al.*, 2007]. A combination of available drugs targeting these protein classes, i.e. growth factor antibody and COX2 and metalloprotease inhibitors, was chosen as treatment with the result of metastasis reduction. This study is a good example for the power of pathway-centric analyses that guide drug administration and the identification of disease driving genes [*Hopkins*, 2008; *Pujol et al.*, 2010].

The simplest approach to find disease modules is to use pre-defined pathways (gene sets). These are derived from gene signatures in different diseases or biochemical experiments and available from MSigDB. The gene sets can also be used to assess whether groups of genes are involved in a pathophenotype [*Alles et al.*, 2009]. While gene sets are useful to recover concerted, often subtle changes of known groupings of genes, network models allow for the identification of new disease modules and give information about the network topology. Network derived pathways are crucial when no relevant gene sets are available for the disease under study.

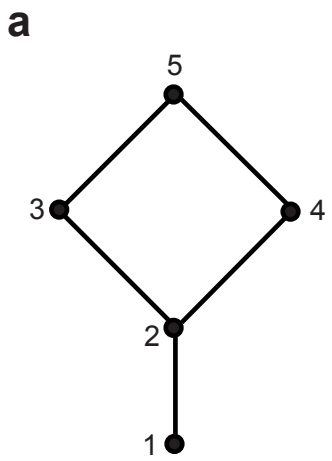
1.6.6 Network Inference

High-throughput techniques, such as microarrays, can be used to build network models and estimate the interactions in the cell. A simple approach to estimate a relation of genes with each other, is clustering [*Eisen et al.*, 1998]. Genes, grouped together in the same cluster, are more likely to have a similar function and to be co-expressed. The grouping is done by calculating a pairwise coefficient, such as Pearson correlation or Euclidean distance, and subsequent application of a clustering method on that measure. Clustering is mostly used in gene expression analysis to infer molecular signatures. Pearson correlations are also used as the most basic technique to infer networks. A certain threshold is set and gene connections passing the threshold are considered to be linked to each other. This approach has first been used by [*Butte et al.*, 2000]. Later, mutual information has been used to generate networks. Mutual information is based on concepts from information theory initially developed by Claude Shannon [*Shannon*, 1949]. A popular reverse-engineering tool using mutual information is ARACNE [*Margolin et al.*, 2006].

Weighted gene co-expression network analysis (WGCNA) [*Zhang and Horvath*, 2005] is another popular method that can be used to construct a co-expression network from microarray data. Briefly, the algorithm starts by calculating absolute values of the Pearson correlation $|\rho_{i,j}|$ for all pairs of genes. A value β is chosen by testing which $|\rho_{i,j}|^\beta$ leads to a network with a scale-free topology. Finally, average linkage hierarchical clustering coupled with a dissimilarity measure is used to define a dendrogram to identify modules.

However, these methods are not able to distinguish direct and indirect (spurious) correlations. For this, an appealing formalism are Gaussian graphical models (GGMs) [*Lauritzen*, 1996] and partial correlations that estimate the associations between two variables considering the remaining ones. Partial correlations are calculated by inverting the covariance matrix obtaining the so-called concentration matrix. Figure 1.17 illustrates the relationship between an undirected graph G, a covariance matrix and a concentration matrix in the framework of a GGM. The covariance matrix corresponding to G contains the σ_{ij} (covariance) for each pair of variables (vertices in the graph) (Figure 1.17b). Therefore, $\sigma_{ii} = \sigma_i^2$ with σ_i (the standard deviation of variable X_i). The elements in the diagonal of Σ are hence the variances of each variable. Figure 1.17c depicts the inverse of the covariance matrix with zeros where two vertices (i, j) in G are disconnected. Thus, two variables are conditionally independent when they have a zero in the concentration matrix. The network can then be modeled by connecting pairs of variables that do not have any zeros in the concentration matrix.

For microarray data the number of genes (random variables, p) is much larger than the number of samples (observations, n) describing the *small n large*



b

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & \sigma_{35} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & \sigma_{45} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55} \end{pmatrix}$$

c

$$\Sigma^{-1} = \begin{pmatrix} \kappa_{11} & \kappa_{12} & 0 & 0 & 0 \\ \kappa_{21} & \kappa_{22} & \kappa_{23} & \kappa_{24} & 0 \\ 0 & \kappa_{32} & \kappa_{33} & 0 & \kappa_{35} \\ 0 & \kappa_{42} & 0 & \kappa_{44} & \kappa_{45} \\ 0 & 0 & \kappa_{53} & \kappa_{54} & \kappa_{55} \end{pmatrix}$$

Figure 1.17: **Gaussian Graphical Model**. a) An undirected graph representation of the underlying Gaussian graphical model is shown. The corresponding b) covariance matrix and c) inverse of the covariance matrix (concentration matrix) of the graph are also depicted. The zero pattern contained in the concentration matrix matches the missing edges in the graph shown in a).

p problem (p $\gg$ n). Due to this, the sample covariance matrix cannot be inverted and hence the partial correlations cannot be calculated. To overcome this problem, statistical strategies have been developed that estimate the covariance matrix under p $\gg$ n, including a Bayesian approach with sparsity inducing prior [*Dobra et al.*, 2004], graphical lasso [*Friedman et al.*, 2008] and limited-order partial correlations [*de la Fuente et al.*, 2004; *Castelo and Roverato*, 2006, 2009]. One of the latter approaches is implemented in the Bioconductor package qpgraph, which provides a reverse-engineering method tailor-made for microarray data based on limited-order partial correlations. The procedure estimates the structure of the network by calculating a quantity called the non-rejection rate (NRR), which is based on partial correlations of order $q < (n-2)$ [*Castelo and Roverato*, 2006]. The NRR gives an estimate of the strength of a direct interaction between two genes and can be understood as a linear measure of association over all marginal distributions of size q that is calculated for every gene pair.

1.7 Genomic-based Drug Discovery and Repositioning

1.7.1 De novo Drug Development

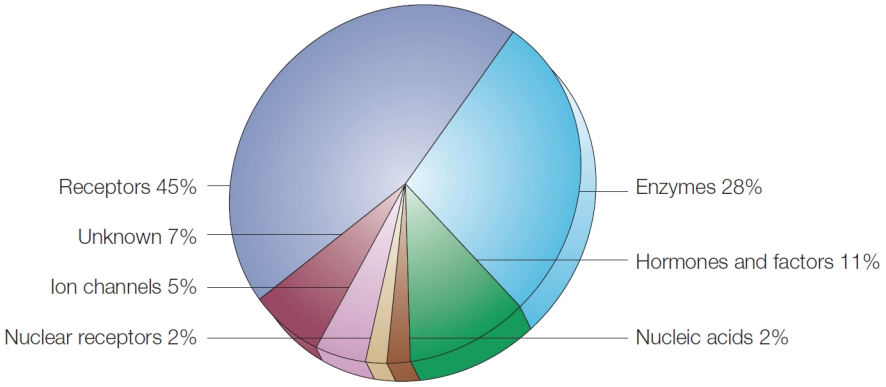


Figure 1.18: **Target classes**. Depicted are the different target classes. The two main classes are enzymes and receptors. Figure source: [*Bleicher et al.*, 2003].

Drug discovery tries to find new drugs with few side effects that can be used to treat a disease. Most companies employ a similar pipeline to find new drugs (Figure 1.19 A). The pipeline begins by identifying a target that is a protein involved in a disease. The most common classes of druggable proteins are receptors and enzymes [*Bleicher et al.*, 2003] (Figure 1.18). The goal is to find compounds that inhibit enzyme activity or block a receptor [*Schadt et al.*, 2009]. Once a target has been identified and classified, compounds are tested for their activity against the target. High-throughput screening techniques are used to test thousands of compounds in the lead discovery step [*Bleicher et al.*, 2003]. Only compounds that show activity against the target (hits) are considered further.

Knowledge of the target and the compound are used to select a fewer number of active compounds to be tested further (focused screening). This knowledge based decision making has to be done as early as possible to reduce the number of possible failures and to save money [*Bleicher et al.*, 2003]. This knowledge comprises chemical properties, activity profiles, similarity to other compounds and solubility. To further narrow down the list of compounds that are potential leads, SAR (structure activity relationships) and ADME (absorption, distribution, metabolism, excretion) [*Butina et al.*, 2002] of the compounds are considered. Promising compounds from this phase move on to the lead optimization stage.

The potentially ideal compound is then designed based on several rounds of biochemical assay data. The candidate compounds progress further from pre-clinical development (tests on cell lines and in animal models) up to clinical trials.

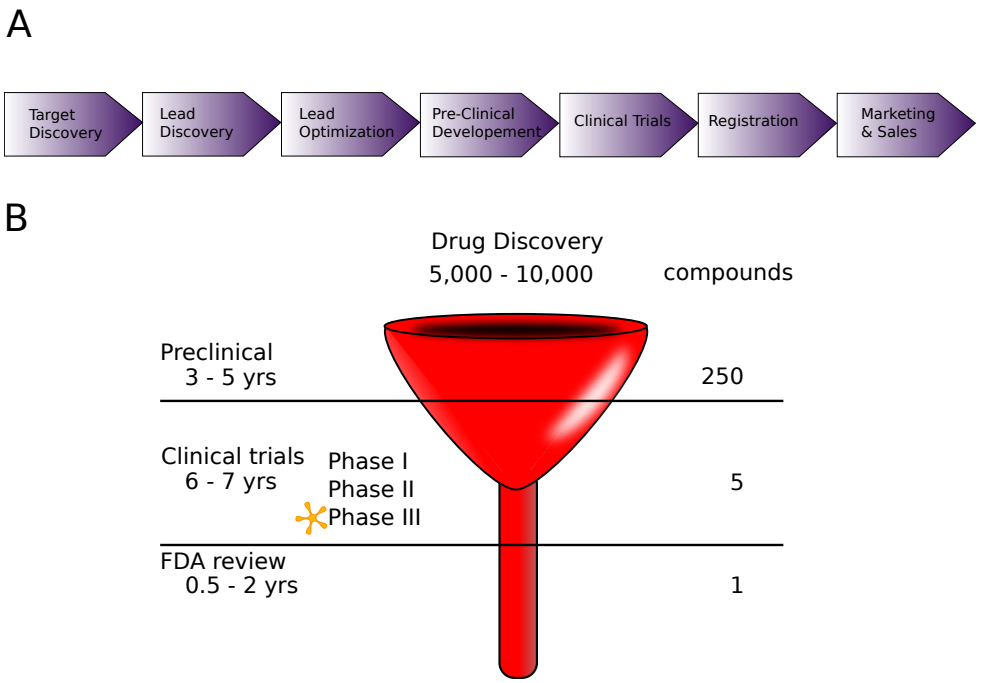


Figure 1.19: **Drug discovery process.** Panel A depicts the typical *de novo* drug discovery pipeline. First, a target (protein that is involved in a disease) is defined. Then, all compounds available to a company are screened against that target (lead discovery). Active compounds are then further selected according to available knowledge, such as side effect profiles or chemical properties. The leads are further optimized and tested in several biological assays before entering pre-clinical trial. Panel B shows the number of compounds that are put in the pipeline. The funnel structure represents the large number of compounds entering the pipeline and the decreasing number of drugs that make it through each step. In the end, maybe one drug is suitable and on average 15 years have passed. The reason for this is, that with the increasing number of participants included in each phase of a clinical trial, the drug exhibits severe side effects. Most drugs fail in phase III. Figure source: `http://www.alzdiscovery.org` and [*Phillips et al.*, 2006].

Figure 1.19 B illustrates the funnel structure of the entire drug discovery process. First, all compounds that are available to a company are tested (5,000 - 10,000) in the high-throughput screening. Of these, maybe 250 compounds go into pre-clinical trial and, 5 into clinical trial and, if lucky, 1 compound is suitable and can go for review to the Food and Drug Administration (FDA). The entire process takes 10 to 17 years and costs about US$800 million [*Chong and Sullivan*, 2007]. Only 11% of drug candidates make it to the market because

they show efficacy and are not toxic [*Kola and Landis*, 2004]. This low outcome is due to the diversity of the human population, every individual reacts differently to a drug due to different epi/genetic make-ups [*Lum et al.*, 2009]. Even after the drug has been marketed, and thus applied to a larger population, there can be individuals for which the drug does not show any efficacy. Also, adverse drug reactions (ADR) can appear. ADRs are classified in two types, A and B, with A being recorded during clinical trial, thus accounted for, whereas type B is completely random. ADRs can appear at any time and the underlying cause is unknown [*Wilke et al.*, 2007; *Scott and Thompson*, 2011].

For example, rofecoxib and cerivastatin have been withdrawn from the market due to unexpected side effects [*Scheiber et al.*, 2009]. The lack of efficacy and the emergence of severe side effects are the main reasons for a drug to fail. The extensive cost and low outcome coupled with the high expectations of finding more effective and safer drugs, has led to the productivity problem (Figure 1.20) [*Ashburn and Thor*, 2004]. One solution to this problem is to use drugs that are already approved by the FDA and find new diseases to use the drug for. This way, the development cost can be reduced to US$17 million and the estimated time of a clinical trial phase II is two years [*DiMasi et al.*, 2003]. These estimates change depending on disease, compound and drug discovery strategy.

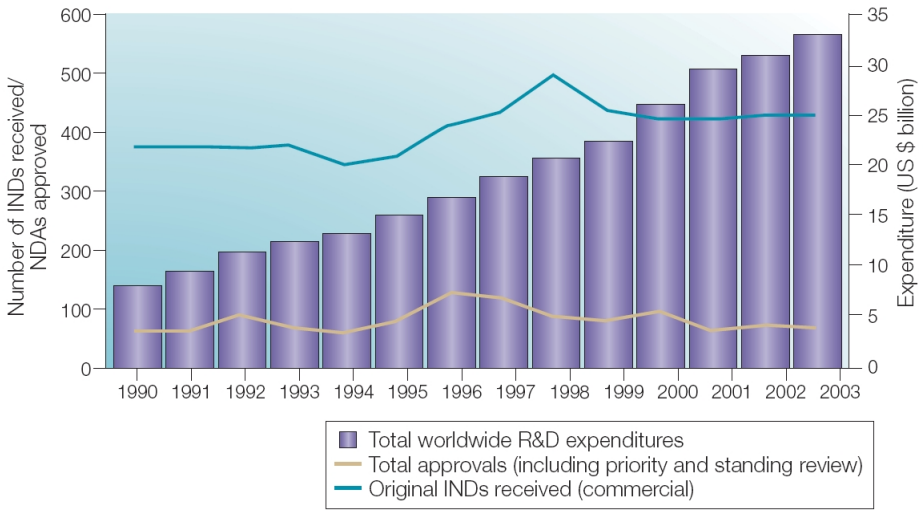


Figure 1.20: **Productivity Problem.** The outcome of new drugs (productivity) has decreased since the 90's. On the y-axis the number of New Drugs Approved (NDA) and the number of Investigatory New Drugs (IND) submitted to the FDA per dollar spent are shown. The x-axis shows the years. Figure from [*Ashburn and Thor*, 2004].

1.7.2 Drug repositioning

The reuse of existing drugs facilitates the drug discovery process. It has been observed that drugs act unexpectedly due to off-target effects. Off-target means that a drug binds to different targets in addition to the actual target it was designed for. This effect can have harmful consequences but can as well be beneficial and lead to new therapeutic indications for drugs. Finding new uses for drugs is termed drug redirecting, repositioning, re-purposing or re-profiling [*Ashburn and Thor*, 2004]. Drug repositioning significantly accelerates the drug finding process and decreases costs [*Chong and Sullivan*, 2007]. The big advantage is that these drugs have already passed the drug discovery pipeline and have been administered to a diverse population of patients reducing the pharmacokinetic uncertainty. Also because the drug passed already certain steps of the pipeline, extensive safety and pharmacokinetic profiles are available [*Ashburn and Thor*, 2004]. Side effects are the key to finding new uses for existing drugs. In the following, two examples for successful drug-repositioning are explained. Both are taken from [*Ashburn and Thor*, 2004]. First, Thalidomide (Contergan) was prescribed for morning sickness during pregnancy 1957 - 1961. The drug led to severe skeletal birth defects and has been taken off the market. In 1964 the interest for Thalidomide revived, because it is the only drug that can treat erythema nodosum leprosum (ENL), a painful complication of leprosy. Second, a rather famous example, is Sildenafil (Viagra). It was designed to relax coronary arteries to allow greater blood flow, but did not show any effect (1991). However, its side effect, penile erection, ensured that it went to into trial in 1993-1995 and by 2003 the annual sales amounted to US$1,88 billion.

1.7.3 Exploiting the Drug-Target Space

Spanning a network with drugs and their target genes (genes encoding proteins that are modulated by a drug), shows that drugs, although designed to be specific, are able to modulate different proteins (promiscuity) [*Paolini et al.*, 2006]. This promiscuity is sometimes intrinsic to the drug's therapeutic efficacy [*Keiser et al.*, 2009], but also opens the road for further investigation of additional uses for the drug. The drug-target space describes the potential of a drug to be effective for another target. When examining the drug-target space it becomes clear that it is not fully exploited and leaves room for new discoveries.

The drug-target space has been used by different research groups to find candidate drugs. Such strategies include side effect similarity [*Campillos et al.*, 2008], ligand similarity [*Keiser et al.*, 2009] and promiscuity [*Keiser et al.*, 2007]. Side effect similarity groups target genes by their side effects, for example, two

drugs causing blurred vision are grouped together. This approach led to a number of new indications that have been experimentally verified [*Campillos et al.*, 2008]. An entire database is available, called Sider [*Kuhn et al.*, 2010], which provides drugs repositioned on a large scale according to their side effects.

A more chemical based approach is the classification of targets by the similarity of their ligands [*Keiser et al.*, 2007; *Hert et al.*, 2008]. The authors developed an algorithm similar to blast [*Altschul et al.*, 1990], which finds additional drug-targets that were predicted to show similar affinities to the ligands as the already known target/ligand interactions. In a second publication the authors describe the validation of the newly found targets and confirmed their *in vitro* efficacy [*Keiser et al.*, 2009]. These studies show that drugs and their targets form a highly interconnected relationship that has potential to be further exploited.

1.7.4 Gene signatures for drug repositioning

Beyond screening libraries of approved compounds against potential targets, drug repositioning nowadays also employs genome-based approaches. As the productivity problem shows, the outcome of the drug pipeline is very low, whereas the costs are very high. It became more and more evident that the underlying mechanisms of a disease also have to be considered, as well as the physiology of test animals [*Lum et al.*, 2009]. With the advent of high-throughput gene expression profiling it became more feasible to integrate genetic changes into the drug finding process. Genomic profiling enables the systematic assessment of molecular changes underlying disease phenotypes [*Lukk et al.*, 2010]. A better understanding of the link between gene expression signatures and physiological profiles became possible [*Lum et al.*, 2009]. The notion underlying genome-based drug-repositioning is that each biological state can be reflected by its genomic molecular profile [*Lukk et al.*, 2010; *Dudley et al.*, 2009; *Harrison*, 2011]. Clinical phenotypes can be described by gene expression as an intermediate phenotype [*Lussier and Chen*, 2011]. Thus, gene expression changes occurring in response to compound exposure can be directly measured and estimates about potential disease indications can be made as has been recently demonstrated in different cancers [*Shigemizu et al.*, 2012; *Jin et al.*, 2012].

One of the first large scale approaches was the connectivity map [*Lamb et al.*, 2006]. The connectivity map (cmap) provides a database of gene expression data from cell lines exposed to different drugs. The user uploads a gene signature and the cmap algorithm compares the uploaded disease signature to the drug signatures present in the database in order to find a potential drug for the disease signature. The Cancer Cell Line Encyclopedia [*Barretina et al.*, 2012] represents

a heterogeneous resource of cancer data such as gene expression, chromosomal copy number and sequencing data from 947 human cancer cell lines. The Gene Expression Omnibus (GEO) [*Barrett et al.*, 2011] has collected over the last decade all kinds of microarray studies including drug perturbation studies. This repository provides the basis to construct drug-disease relationships based on genomic information. This strategy was recently employed with the goal of finding new drug-disease associations that were validated for inflammatory bowel disease and lung adenocarcinoma [*Sirota et al.*, 2011; *Dudley et al.*, 2011].

1.7.5 Networks in disease

All the aforementioned approaches are based on single gene measurements. Gene expression signatures have been demonstrated to be able to classify subtypes of cancer [*Perou et al.*, 2000]. The underlying pathway changes, however, are not understood entirely. Also, for some diseases such as type 2 diabetes, glioblastoma or coronary artery disease, it has been shown that the disease phenotype is caused due to small changes in many genes and not striking changes in few genes [*Schadt et al.*, 2009]. Therefore, it is necessary to investigate the pathway changes that lead to the disease [*Lussier and Chen*, 2011]. Several co-expression analysis tools have emerged in the recent years that identify differentially co-expressed genes [*Choi and Kendziorski*, 2009] or that can be used to identify groups of co-expressed genes [*Zhang and Horvath*, 2005; *Oldham et al.*, 2006; *Tesson et al.*, 2010; *Langfelder and Horvath*, 2008]. For a more detailed review of different network methods the reader may be referred to [*de la Fuente*, 2010].

Gene network enrichment analysis (GNEA) [*Liu et al.*, 2007] is a network based enrichment method that integrates a protein interaction network with gene expression data. Further studies on differential network analysis use graphical models to integrate the disease state [*Orešič et al.*, 2011; *Sysi-Aho et al.*, 2011; *Pietiläinen et al.*, 2011]. Last, machine learning methods are used to find disease driving genes [*Nitsch et al.*, 2010].

Genome-based networks can also be utilized to find potential drug targets. The goal is to find the subnetwork that is directly affected by a drug. High-throughput measurements deliver data that help to identify this group of genes affected by a perturbation (gene signature) [*Iorio et al.*, 2010]. A genome-based network strategy that has been previously proposed is able to identify this particular subset of genes. A network is constructed from gene expression data that is able to distinguish directly and indirectly affected transcripts [*Gardner et al.*, 2003; *Bernardo et al.*, 2005]. The latter algorithm employs samples with different perturbations to construct a training network. Next, a sample exposed to a drug of interest is used as a test set. The training network is then filtered with the test data and the direct (potential) targets of the compound can be identified.

CHAPTER 2

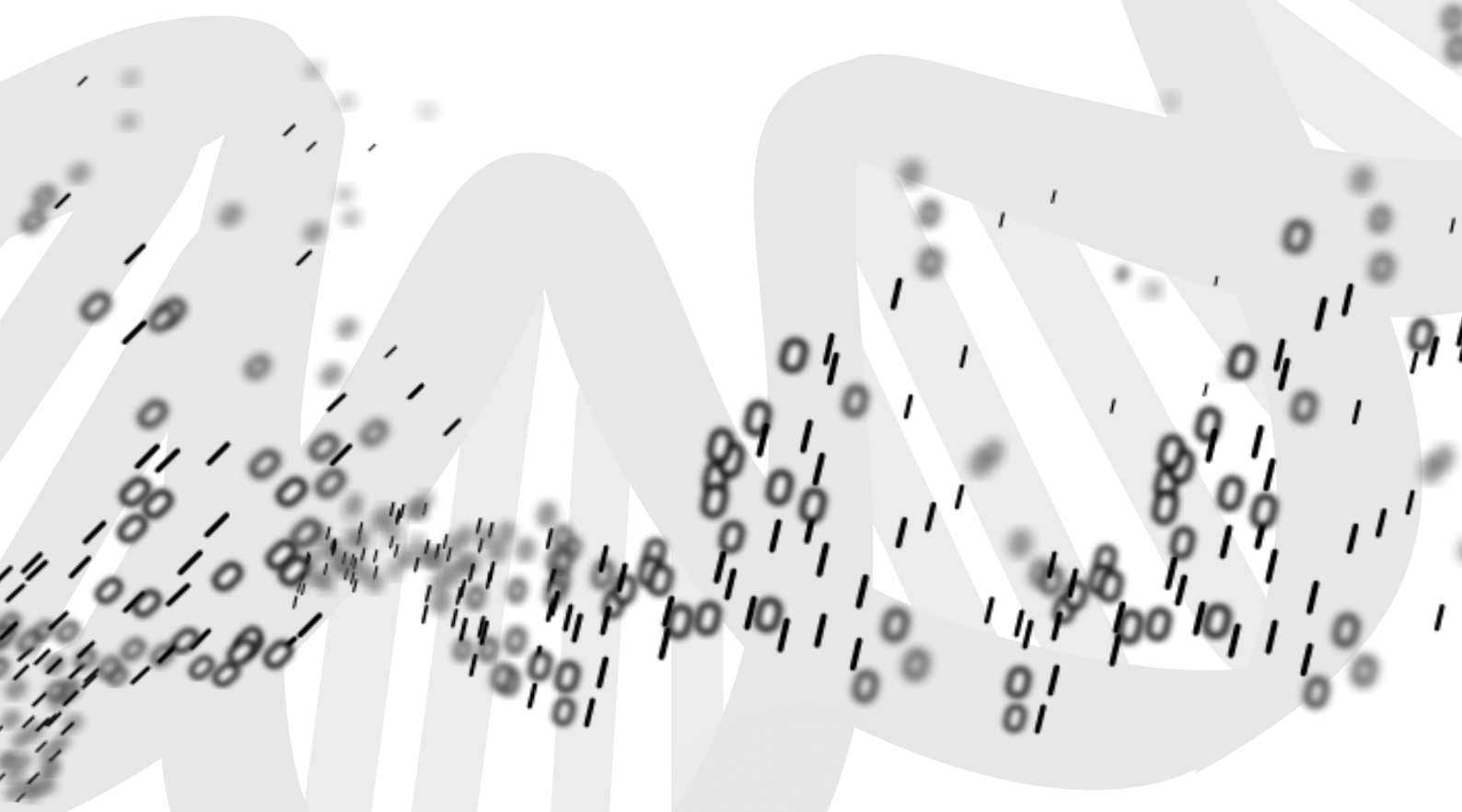
OBJECTIVES

The previous chapter gave an introductory overview of the different principles that have been used to achieve the following objectives:

- Developing a bioinformatic method that enables pathway-centric analyses of high-throughput genomics data by assessing sample-wise pathway variation in an unsupervised manner. The method should be applicable to RNA-seq and microarray data equally.
- Developing a pathway-centric strategy to reposition drugs based on high-throughput genomics data.
- Identifying disease driving pathways in type 2 diabetes using co-expression analysis and transcription factor binding site analysis to find a new biomarker for insulin secretory dysfunction.

CHAPTER 3

RESULTS



3.1 GSVA - Gene Set Variation Analysis

This contribution was under review at the time the thesis was submitted.

Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics. 2013 Jan 16;14(1):7.

Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. Supplementary material. BMC Bioinformatics. 2013 Jan 16;14(1):7.

3.2 Pathway-centric genome-based drug-repositioning

This contribution was in preparation and the compounds were being validated at the time the thesis was submitted.

Pathway-centric genome-based drug repositioning in human islet cells in type 2 diabetes.

Sonja Hänzelmann[1,2], Emre Güney[2], Robert Castelo[1,2] and Anders Rosengren[3*]

1 Research Program on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM)
2 Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain
3 Lund University Diabetes Centre, Lund University, SE-20502 Malmö, Sweden
* Corresponding author

1 Abstract

The clinical picture of type 2 diabetes (T2D) is characterized by the dysfunction of islet cells to secret insulin. This deficiency results from poly-genetic alterations driving the disease phenotype. A pathway-centric view enables the modeling of these underlying molecular mechanisms and helps to identify disease driving modules. The principle of disease driving modules is based on genome-wide networks that are robust against random deletions or inhibition of individual nodes as has been demonstrated by several gene knock-out studies. Hence, disease genes tend to group together in the same neighborhood forming disease modules and multiple genes have to be targeted simultaneously to attain an effect on the pathophenotype. These observations motivated us to develop a pathway-centric genome-based strategy to reposition drugs that would be able to restore insulin secretion in T2D. This strategy consists of two steps, first we infer a regulatory network from 64 gene expression samples of human islets that is used to identify disease driving modules. Second, we collected about 20,000 microarrays that have been exposed to about 1,800 compounds. Gene expression can be seen as an intermediate phenotype reflecting underlying dysregulatory pathways in a disease. Hence, genes contained in the disease modules that elicit similar transcriptional responses upon drug exposure are assumed to have a therapeutic effect. We identified four potential compounds, methimazole, pantoprazole, bitter orange extract and torcetrapib that might have a positive effect on insulin secretion. This is the first time a regulatory network of human islets has been used to reposition drugs for T2D.

2 Introduction

The reuse of existing drugs facilitates the drug discovery process. It has been observed that drugs act unexpectedly due to off-target effects. Off-target means that a drug binds to different targets in addition to the actual target it was designed for. This effect can have harmful consequences but can as well be beneficial and lead to new therapeutic indications for drugs. Finding new uses for drugs is termed drug repositioning [*Ashburn and Thor*, 2004]. Drug repositioning significantly accelerates the drug finding process and decreases costs [*Chong and Sullivan*, 2007]. The big advantage is that these drugs have already passed the drug discovery pipeline and have been administered to a diverse population of patients reducing the pharmacokinetic uncertainty and providing extensive safety and pharmacokinetic profiles [*Ashburn and Thor*, 2004].

Beyond screening libraries of approved compounds against potential targets, drug repositioning nowadays also employs genome-based approaches [*Dudley*

et al., 2011; *Sirota et al.*, 2011; *Lussier and Chen*, 2011; *Sanseau et al.*, 2012]. It became more and more evident that the underlying mechanisms of a disease also have to be considered, as well as the physiology of test animals [*Lum et al.*, 2009]. With the advent of high-throughput gene expression profiling techniques, the integration of genomic information into the drug finding process is facilitated. Genomic profiling enables the systematic assessment of molecular changes of nucleic acids underlying disease phenotypes [*Lukk et al.*, 2010]. A better understanding of the link between gene expression signatures and physiological profiles became possible [*Lum et al.*, 2009]. The notion underlying genome-based drug-repositioning is that each biological state can be reflected by its genomic molecular profile [*Lukk et al.*, 2010; *Dudley et al.*, 2009; *Harrison*, 2011]. Clinical phenotypes can be described by gene expression as intermediate phenotype [*Lussier and Chen*, 2011]. Thus, gene expression changes occurring in response to compound exposure can be directly measured and estimates about potential disease indications can be made. The largest, systematically designed resource for this type of data for human cell lines is the connectivity map [*Lamb et al.*, 2006]. Over the last decade, the Gene Expression Omnibus (GEO) [*Barrett et al.*, 2011] has collected a vast amount of microarray studies including drug perturbation studies. These repositories provide the basis to construct drug-disease relationships based on genomic information, which can be used to find new uses for drugs [*Sirota et al.*, 2011; *Dudley et al.*, 2011; *Shigemizu et al.*, 2012]. A straightforward method employed for this purpose has been the assessment of differential expression independently for each gene. The single-gene approach has been successfully used to characterize many physiological pathways and monogenic disorders but seems to be limited in delineating the mechanisms underlying complex disorders [*Pujol et al.*, 2010; *Barabasi et al.*, 2011].

In diabetes, genome-wide association studies indicate that multiple genes undergo small changes rather than a few affected genes that show significantly altered gene expression profiles [*Schadt et al.*, 2009]. These observations led to a strategy that considers genes interacting with each other in the cell and tries to build and work with network models of the underlying molecular mechanism. Networks have been employed for drug discovery as they are able to predict side effects as well as potential escape routes and interaction pathways for genes [*Iorio et al.*, 2010]. Many biological networks follow a scale-free power-law distribution [*Barabasi and Albert*, 1999] and are small world [*Watts and Strogatz*, 1998]. The characteristic of the power-law distribution is that most genes have very few links, whereas very few genes have many links (hubs). The small world property implies that genes are connected to each other with shorter paths than what would be expected in a random network. It has been found that hubs are mainly encoded by essential genes, whereas disease genes are on the periphery of modules[*Jeong et al.*, 2001; *Albert et al.*, 2000; *Han et al.*, 2004; *Goh et al.*, 2007].

For disease networks it has been demonstrated that disease driving genes indeed tend to group together into modules and reside in the same neighborhood [*Goh et al.*, 2007]. Further, disease driving modules have been identified that react to perturbations [*Schadt et al.*, 2009]. The principle of disease driving modules is based on genome-wide networks that are robust against random deletions or inhibition of individual nodes as has been demonstrated by several gene knock-out studies [*Hopkins*, 2008; *Pujol et al.*, 2010]. Under this model a single gene knock-out may not affect the phenotype, however, the simultaneous knock-out of two genes results in lethality or sickness (synthetic lethality and synthetic sickness) [*Ooi et al.*, 2006]. This implies that multiple proteins may have to be targeted to achieve an effect on the pathophenotype and at the same time essential genes should not be targeted.

Thus, understanding the pathophenotype on the molecular level plays a crucial role for developing effective therapeutic strategies. Here, we used a pathway-centric genome-based drug repositioning approach to identify disease driving modules that are likely to be involved in the development of type 2 diabetes (T2D). The clinical picture of T2D is characterized by the reduced ability of pancreatic β-cells to elevate insulin secretion in response to increased blood glucose levels [*Kaufman*, 2011]. Many studies aim at identifying compounds that help to reverse the mechanisms that lead to β-cell deficiency and therewith T2D. We focus on a specific region of the pancreas that is responsible for insulin secretion, human islets. The molecular profiles exhibit a gradient nature from healthy to disease. Hence, to detect disease modules specific to islets, we applied a network reverse-engineering algorithm [*Castelo and Roverato*, 2006, 2009] to microarray expression data of 64 human islet samples from both healthy and diabetic donors. We mapped the members of these disease modules onto a human protein interaction network that is associated to T2D and prioritized modules that contain genes encoding disease proteins related to diabetes. We downloaded gene expression data that have been exposed to a wide range of compounds to identify drug signatures that are present in the disease modules. The potential new therapeutics that we identified are methimazole, pantoprazole, bitter orange extract and torcetrapib. We have found independently documented evidence in the literature that these compounds are potential drug candidates because they potentially have a direct or indirect effect on insulin secretion.

3 Results

We aimed at using a pathway-centric approach to reposition compounds in type 2 diabetes (T2D) that are potentially targeting disease driving groups of genes (modules). For this, we constructed a transcriptional regulatory network from

human islet gene expression data consisting of 19 diabetic and 45 healthy individuals. This network is a molecular snapshot of T2D in human islets. In addition, we downloaded a large amount of data sets containing compound exposed microarray data sets from public resources. A schematic overview of our strategy is shown in Figure 1. Rather than defining a disease signature of human islets based on independent gene expression changes, we employed Gaussian graphical model theory [*Lauritzen*, 1996] to reverse-engineer a transcriptional regulatory network by means of limited-order partial correlations [*Castelo and Roverato*, 2006, 2009]. We identified potential disease driving modules in T2D and assessed their expression changes in microarray data of samples exposed to compounds. The idea behind our strategy is that the disease state can be reflected by its genomic molecular profile (that is, disease driving modules) and that if a module is affected by a compound (derived from compound exposed gene expression profiles), the compound can potentially perturb the disease phenotype.

3.1 Regulatory Network Inference

We inferred a regulatory network from microarray data and then derived modules from the network. The use of the network is motivated by the rarity of pre-defined pathways specific for human islets. We used diabetic (Type 2 Diabetes,T2D) and healthy (Non-diabetic, ND) samples together to have sufficient variability across the 64 samples. This is necessary because of the subtle gene expression changes between the two phenotypes and to be able to uncover the largest possible fraction of the underlying molecular regulatory network. We reverse-engineered the network using the qpgraph package [*Castelo and Roverato*, 2009] and then selected a NRR cut-off that provided modules of sufficient size (at least ten genes) needed for the remainder of the analysis (see methods for details). These regulatory modules consist of transcription factors or open chromatin genes [*Gaulton et al.*, 2010] as central genes and co-expressed genes connected to the central ones. The resulting network illustrates the modular structure of the network and interconnectedness of the modules (Figure 2). Central genes are colored in yellow and co-expressed genes in purple.

3.2 Module Prioritization

We prioritized the disease modules that contain genes encoding proteins related to diabetes in order to identify potential compounds capable of modulating these proteins. The prioritization was done by employing GUILD [*Guney and Oliva*, 2011], a network-based disease-gene prioritization framework, which calculated T2D-association scores using protein-protein interactions and known T2D-gene annotations (see methods). In order to prioritize the 90 modules based on their

significance for T2D, we used a protein-protein interaction network. The guilt-by-association principle suggests that proteins in the neighborhood of known disease proteins are also implicated in the disease. We use a protein-protein interaction network constructed from several public sources, identify proteins known to be associated with T2D and use different distance measurements from these source proteins to neighboring proteins to calculate association scores. The proximity from the source (disease) protein to the other proteins is recorded. After calculating the proximity score for each protein in the network, we linked the proteins to the genes contained in the 90 modules and calculated an average association score per module using the proximity scores. Then, the modules were shuffled by randomly assigning genes and association scores to the modules, of the same size as the original module. This was repeated 10,000 times and used as a measure of false positives detection. Modules, containing genes with stronger associations to diabetes were selected by passing a cut-off of 0.05, leaving 18 modules (Table 7). We chose the modules controlled by the following transcription factors SLC30A8, G6PC2 and GNAO1, because their member genes have been associated with T2D or important metabolic pathways. The modules are highlighted in Figure 2 and their members are shown in Table 2.

3.3 Compound Prioritization

To identify potential compounds targeting the prioritized disease modules, we downloaded data of about 20,000 microarray chips from the Gene Expression Omnibus (GEO) [*Barrett et al.*, 2011], the connectivity map (Cmap) [*Lamb et al.*, 2006], a rat study [*Tamura et al.*, 2006] from Array Express [*Brazma et al.*, 2003] and a big rat liver study [*Natsoulis et al.*, 2008] that have been systematically exposed to a wide range of compounds (in total about 1,800 compounds). During the selection of the drug experiments, we had to discard studies due to unsolvable batch effect issues or bad quality chips leaving a rather small set of studies. Of about 1,800 compounds and 20,000 samples we were left with 884 compounds and 10,331 samples.

Because we wanted to assign new purposes to existing compounds, we focused on data sets that are not investigating diabetes but other diseases, such as different cancer types, bowel or skin diseases. For each of the 90 modules, we applied a standardized z-statistic and assessed the enrichment of the module in the compound profile. The compound profile was calculated by differential expression between treated and untreated samples (see methods). We selected compounds where disease driving modules were enriched at a FDR < 0.01. We found that the 90 modules are enriched for 834 compounds. To reduce the number of compounds, we only looked at the 18 modules prioritized by GUILD, which are targeted by 289 compounds (Figure 3). To narrow the list of candidate

compounds further down, we selected three of the 18 modules that are related to insulin secretion. The three modules are enriched for 78 compounds (Figure 4). To filter this list further, we discarded compounds that 1) are already repositioned for T2D, 2) unlikely to have an effect on insulin secretion, 3) are toxic. These criteria narrowed our list down to 21 compounds. Of these we selected four that have a potential to restore insulin secretion dysfunction. We searched literature to study the mechanisms of the modules and their member genes as well as the potential effects that the compounds could have on them.

3.4 A closer look at the disease modules

The genes of the same module are grouped together because they are co-regulated and functionally similar or involved in the same processes. We focused on pathways that could potentially be involved in glucose homeostasis and insulin secretion. This mechanism is triggered by elevated glucose levels in the blood. Glucose is transported into the β-cells by glucose transporters, which leads to the depolarization of the membrane and influx of calcium. The increased concentration of calcium leads to insulin secretion. In the following we highlight some genes that are members of the disease modules that we prioritized. The modules are highly interconnected indicating the existence of cross-talk and back-up pathways, which when targeted by the compounds might lead to the perturbation of the pathophenotype. In pancreatic islets, insulin production is linked with zinc transporters encoded by SLC30A8 (Solute carrier family 30 (zinc transporter), member 8), which is one of the open chromatin genes. Co-regulated genes contained in the same module include SCGN (Secretagogin), a secreted calcium-binding protein, CASR is a calcium-sensing receptor which senses extracellular levels of calcium ions and DPP6 (Dipeptidyl aminopeptidase-like protein 6), which binds and alters voltage-gated potassium channels, which are involved in the regulation of intracellular calcium homeostasis. G6PC2 is an enzyme found in pancreatic islets and a major target in T2D and is connected to the SLC30A8 module. GAD2 (Glutamic acid decarboxylase 2 gene) over expression impairs insulin secretion in animals. GLP1R (Glucagon-like peptide 1 receptor) is known to be expressed in pancreatic β-cells. Activated GLP1R stimulates the adenylyl cyclase pathway which results in increased insulin synthesis and release of insulin. Consequently GLP1R has been suggested as a potential target for the treatment of diabetes.

3.5 A closer look at the compounds

The compounds that we identified methimazole, pantoprazole, bitter orange extract and torcetrapib as potential therapeutics. They are interesting candidates

due to their biological effects and because they have not been thoroughly studied for insulin secretion. In the following, we comment why we think that these compounds are potential candidates for T2D. Methimazole is an organic compound approved by the FDA. It is a thyroid peroxidase inhibitor that is administered to reduce thyroid hormone production. We selected this drug because it is easy to acquire and has a well-studied safety profile. Moreover, a study has recently investigated methimazole in a rat model and suggested that it can help to delay the onset of T2D [*Hwang et al.*, 2009]. Elevated levels of thyroid hormones lead to elevated blood glucose levels because the pancreas is reducing its insulin production. It has been hypothesized that targeting tissue-specific thyroid hormone action can help to treat T2D [*Crunkhorn and Patti*, 2008]. Thus, we want to study if there is a direct interaction between insulin signaling and thyroid hormone function.

Pantoprazole is a proton pump inhibitor (*PPI*) drug used for short-term treatment of erosion and ulceration of the esophagus caused by gastroesophageal reflux disease. The reason we think that *PPIs* could be of use for T2D is that they block the last step of gastric acid production causing gastrin levels to rise which stimulate the pancreas to produce insulin [*Rooman et al.*, 2002]. However, since we have found that pantoprazole potentially affects the modules in islets cells and not via gastrin, we propose to test whether it has a direct effect on insulin secretion. Moreover, in a pure observational study it has been reported that *PPI* reduces blood sugar levels [*Crouch et al.*, 2012], which supports the hypothesis that pantoprazole is an interesting drug candidate.

Torcetrapib is a cholesteryl ester transfer protein (CETP) inhibitor that has been developed by Pfizer to treat elevated cholesterol levels to prevent cardiovascular disease. However, phase III clinical trial was halted due to severe side effects of the drug. T2D is associated with low levels of HDL, that can be raised by inhibiting CETP. Torcetrapib has been suggested as a potential drug to treat T2D because it showed a positive effect in a mouse model of insulin resistant, obese mice [*Briand et al.*, 2011]. Further, it has been shown to improve glycemic control in atorvastatin-treated patients with T2D [*Barter et al.*, 2011]. A more detailed understanding of the mechanism of torcetrapib is necessary. We hope to determine a mechanism directly influencing human islets.

Bitter orange extract enriched in flavanones [*Giménez-Bastida et al.*, 2009] has been studied in human colon fibroblasts where it exhibits anti-inflammatory properties by repressing the plasminogen activator inhibitor-1. In a different study, citrus extract showed hypoglycaemic activity and an anticholinesterase effect [*Conforti et al.*, 2007]. Based on this observation, we decided to investigate the bitter orange extract as a potential bioactive compound for T2D.

4 Discussion

The reductionist view of biology has changed over the last two decades into a more systems-level view on biology. A change of entity from single target to multiple targets in form of pathways is leading to new exciting possibilities. Here, we used this idea and presented a pathway-centric genome-based drug repositioning strategy based on high-throughput genomics data. To our knowledge, a human islet specific regulatory network has not been studied before and used for systematic drug repositioning. We inferred this network from gene expression data of human islets (healthy and diabetic samples) and identified gene regulatory modules relevant to T2D. Next, we tested whether these modules were significantly altered in drug-exposed microarray data to identify compounds that would potentially target the entire module and thus restore the dysfunction of insulin secretion. T2D is a poly-genetic disease showing alterations in many different genes driving the disease phenotype. Hence, a network model seemed to be the most appropriate to model the disease and provide a snapshot of the underlying changes that have the potential of being targeted by a compound. We began with a long list of compounds for which gene expression data was available and shortened it by module prioritization to 78 compounds that target the three modules SLC30A8, G6PC2, GNAO1. Of these 78 compounds we selected four that promise to have the most potential to be beneficial for T2D patients. We carried a literature research out to validate the potential of these compounds. Further, we identified disease modules rather than gene signatures, hoping to target a group of disease genes that will perturb the disease phenotype upon compound administration.

Our network is a molecular snapshot of the interactions underlying insulin secretion and could be used in the future to identify genes that could be potential drug targets. Several network topology measures such as betweenness (number of shortest paths crossing a node) can help to identify so called bottleneck nodes. These nodes have a high betweenness, hence controlling the information flow and could thus be potential drug targets [*Yu et al.*, 2007; *Loscalzo and Barabasi*, 2011]. This network could be further used to identify drug associations to help estimatating multi-drug and possible side-effects.

5 Methods

5.1 Drugs

The soluble extract from bitter orange kindly provided by Zoster S.A. (Murcia, Spain) was composed of flavonoids (4555%), water (35%), proteins (1113%),

pectins (1214%), cellulosic material (01%), ashes (23%), vitamins (10001200 ppm vitamin C and traces of vitamins B1, B2, and B6), -carotene (35 ppm), and other components (35%). On arrival, the extract (a hygroscopic brown powder, 2.1% humidity) was kept in a tightly closed container within a desiccator at room temperature. Fresh solutions of the extract were prepared by dissolving 0.1 g of the powder extract in 50 mL of water (0.2%). Pepsin from porcine stomach mucosa (pepsin A, EC 3.4.23.1), pancreatin from porcine pancreas (4 g/L, 1.6 103 U.S. Pharmacopeia (USP) specifications), and bile extract mixture (25 g/L, containing a mixture of sodium cholate and sodium deoxycholate) were all from Sigma (Steinheim, Germany). Naringin (naringenin-7-O-neohesperidoside), hesperidin (hesperetin-7-O-rutinoside), neohesperidin (hesperetin-7-O-neohesperidoside), naringenin (4,5,7-trihydroxyflavanone), hesperetin (3,5,7-trihydroxy-4-methoxy flavanone), and isosakuranetin (5,7-dihydroxy-4-methoxyflavanone) were purchased from Extrasynthese (Genay, France). Human recombinant TNF- was from Sigma (Steinheim, Germany). All other chemicals were of analytical/HPLC grade. Ultrapure Millipore water was used for all solutions.

Torcetrapib was used at 3 uM (micro-molar) dissolved in DMSO and water Methimazole was used at 1 mM (milli-molar) dissolved in water. Pantoprazole was used at 100 uM (micro-molar) dissolved in water.

5.2 Open Chromatin and Transcrtiption factors

The transcription factors used in this study are obtained from TRANSFAC [*Matys et al.*, 2003]. Open chromatin genes [*Gaulton et al.*, 2010] are available from the ENCODE project (`http://genome.ucsc.edu/ENCODE/downloads.html`).

5.3 Obtaining ranked drug lists

In total, we used about 20,000 samples and about 1,800 compounds for the analysis, obtained as follows.

5.3.1 Manually collected and rat data sets

The following data sets were downloaded from GEO and used to generate the ranked lists of studies exposed to compounds: GSE10299, GSE10669, GSE10770, GSE10904, GSE11393, GSE11578, GSE11670, GSE11792, GSE11919, GSE12211, GSE12261, GSE12446, GSE12446, GSE12446, GSE12693, GSE12748, GSE12972, GSE13046, GSE13046, GSE13046, GSE13413, GSE1417, GSE14429, GSE14842, GSE14973, GSE15322, GSE15913, GSE15918, GSE15946, GSE15946, GSE15946, GSE15946, GSE15947, GSE16625, GSE16899, GSE17183, GSE17480, GSE18454, GSE18454, GSE18486, GSE1922, GSE19495, GSE20114, GSE20719, GSE23399,

GSE24468, GSE24824, GSE2638, GSE2639, GSE5007, GSE5741, GSE6721, GSE7538, GSE7648, GSE7807, GSE7868 and GSE9484.

Data analysis was performed using the R programming language [*Team R*, 2010] and the Bioconductor software packages. We selected chips which passed quality control using affyPLM [*Bolstad*, 2004]. AffyPLM fits models on probe set level to identify chips of lower quality. Relative Log Expression (RLE) values (comparing probe expression on each array against the median expression across all arrays) and Normalized Unscaled Standard Errors (NUSE) (standard error estimates obtained for each gene and standardized across arrays) are calculated and cut-offs applied that remove samples not falling into the margins. Each Affymetrix chip was background adjusted, normalized and log2 transformed using the Robust Multichip Averaging (RMA) algorithm [*Irizarry et al.*, 2003].

The two rat studies have been downloaded from GEO with the accession number GSE8858 [*Natsoulis et al.*, 2008] (5312 samples, 344 compounds) and from Array Express with the accession number E-MTAB-800 (8106 samples, 130 compounds) [*Tamura et al.*, 2006]. For all data, we used the matching compound and control pairs to calculate the ranked list, for this we used a moderated t-statistic [*Smyth*, 2004] and corrected for multiple testing using FDR [*Benjamini and Hochberg*, 1995]. Finally, we ranked each gene list by the t-value. The rat genes have subsequently been homology mapped with the Ensembl genome browser [*Flicek et al.*, 2010] to human genes.

5.3.2 The connectivity map

We downloaded the microarray data from the connectivity map (cmap) (build 02)[*Lamb et al.*, 2006], which contains 6100 samples of four cell lines treated with 1309 distinct small molecules. The cell lines are MCF7, a breast cancer cell line, prostate cancer epithelial cell line PC3, and HL60, a leukemia cell line. The data was measured on the production chip HT_HGU133A and HL60 also on the HGU133A chip. There is a batch effect present in the cmap data, for this We pre-processed the cell lines in a similar ways to [*Iskar et al.*, 2010]. In the high-throughput gene expression setting, a batch is defined as a set of microarrays grouped together for processing. The batch effect refers to a systematic bias that has to be removed or adjusted for to avoid that biological information is influenced by technical information [*Leek et al.*, 2010]. We removed the batch effect by following the filtering steps described in [*Iskar et al.*, 2010], who used a mean-centering approach. We followed their instructions and added some changes that suit our purposes better. First, we consider the drug exposed samples and group the data by cell line and chip type. Second, we discard samples contained in a batch of size less than 10 samples. This leaves us with 323 samples for HL60 HG-U133A, 885, 2576 and 1440 samples for HL60, MCF7 and PC3 samples for

the HT_HGU133A array, respectively. We use RMA [*Irizarry et al.*, 2003] to normalize these samples. For each probeset for each batch, cell line and array type combination we apply mean-centering to remove the batch effect [*Iskar et al.*, 2010]. Next, we remove control probe sets and map the remaining probe sets to their corresponding entrez gene ids. We repeat these steps with the vehicle samples. In order to carry differential expression (DE) analysis out, we removed all drug induced samples that did not have at least three replicates. For each compound we calculate a ranked list of gene expression levels for each cell line with limma [*Smyth*, 2004] using all the controls for the respective cell line as reference. We adjusted for multiple test correction with FDR [*Benjamini and Hochberg*, 1995].

5.4 Enrichment analysis

We retrieved a list of differentially expressed genes for each gene expression study. Assuming that the moderated t-statistics are approximately normally distributed, we use the following z-test [*Irizarry et al.*, 2009] to identify modules with a consistent shift in mean expression in compound-exposed microarray data. Let $t_S = \{t_1, \ldots, t_k\}$ denote the moderated t-statistics for each gene $1 \ldots k$ in a module S, the z-score statistic is defined as:

$$Z_S = \sqrt{|S|}\bar{t}_S, \tag{1}$$

where $|S|$ denotes the number of genes in a gene set and $\bar{t}_S$ the sample mean of the moderated t-statistics t_S. Assuming that the moderated t-statistics in t_S are independent under the null hypothesis a standard normal distribution can be used to calculate the corresponding p-values [*Irizarry et al.*, 2009].

5.5 Network Inference

We construct a regulatory network based on samples of healthy and diabetic patients from microarray data. We use the R package qpgraph [*Castelo and Roverato*, 2009] which is tailor made to learn an undirected Gaussian graphical model from microarray data by estimating q-order partial-correlations that lead to a quantity called the non-rejection rate (NRR). The NRR gives an estimate of the strength of the direct association between two genes by testing for zero partial correlations of order $q < n - 2$. Partial correlation is a measure of association between two genes that takes all the remaining observed genes into account, hence, reducing the number of spurious (that is, indirectly mediated by other genes) correlations. NRRs were calculated using different q-values equally

distributed from 1 to $n-2$ and then averaged, which has been shown to be a sensible strategy in order to avoid choosing a single *q*-value [*Castelo and Roverato*, 2009].

5.6 Prioritization of regulatory modules in relevance to T2D using protein-protein interactions

A detailed description of the proximity score calculation is given as follows. We used the weighted human binary protein-protein interaction network compiled by Guney and Oliva (Guney and Oliva, under revision). The interaction network was created using the biological interactions and network analysis (BIANA) tool [*Garcia-Garcia et al.*, 2010] and interactions from the following databases: DIP [*Salwinski*, 2004], HPRD [*Keshava Prasad et al.*, 2009], IntAct [*Kerrien et al.*, 2007], MIPS (MPACT) [*Güldener et al.*, 2006], BIOGRID [*Breitkreutz et al.*, 2008]. Using this interaction network and known gene associations for T2D in the Online Inheritance in Man (OMIM) database [*Amberger et al.*, 2009], we computed a T2D association score for each gene in the interaction network with the NetCombo prioritization algorithm provided in the GUILD framework [*Guney and Oliva*, 2011]. We link the genes in each module to the corresponding proteins in the protein interaction network. Then, we calculate an average association score for each module. Only genes that are in the network were taken into consideration while calculating T2D association scores for the module. Next, we evaluated the significance of the prioritized modules by comparing their association scores with the association scores for modules of the same size expected by chance. For each module of size k, we sampled k genes in the interaction network and calculated T2D association score for this random module. We repeated this procedure 10,000 times and counted the number of times that a random module had an association score equal or higher than the original module to yield a p-value. We, then, considered only the modules that were associated with T2D significantly with respect to network-based association scores.

Competing interests

The authors declare no conflict of interest.

Authors' contributions

Acknowledgements

S.H. and R.C. acknowledge support from an ISCIII COMBIOMED grant [RD07/0067/0001] and a Spanish MICINN grant [TIN2011-22826]. E.G. was supported by the Departament d'Educació i Universitats de la Generalitat de Catalunya i del Fons Social Europeu through an FI fellowship.

6 Figures

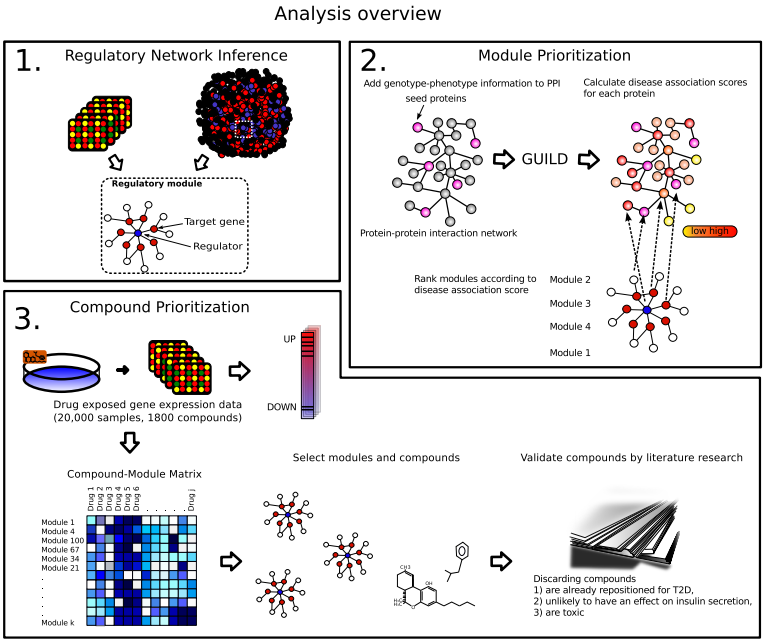


Figure 1: **Overview of module identification and drug signature enrichment.** Our pathway-centric genome-based drug-repositioning strategy consists of three steps: 1) Inference of the regulatory network and module identification, 2) Prioritization of the modules, 3) Compound prioritization. 1) We begin with the inference of a transcriptional regulatory network derived from gene expression data of 64 healthy and diabetic individuals. We apply a NRR cut-off to the network to obtain modules with sufficient sample size. 2) Then, we prioritize the modules using GUILD. GUILD is a network-based disease-gene prioritization framework, which calculates T2D association scores using protein-protein interactions and known T2D-gene annotations. We link each gene of the modules onto the network and calculate the average association score per module. The higher the score, the more interesting the module is. 3) Next, we downloaded 20,000 samples of compound-exposed gene expression data and calculated enrichment scores for each module. Finally, we select three modules from the compound-module matrix that are involved in insulin secretion. 78 compounds are targeting these three modules. We discarded compounds that 1) are already repositioned for T2D, 2) unlikely to have an effect on insulin secretion, 3) are toxic. Of the remaining 21 compounds, we select for that could potentially have an effect on insulin secretion based on literature research.

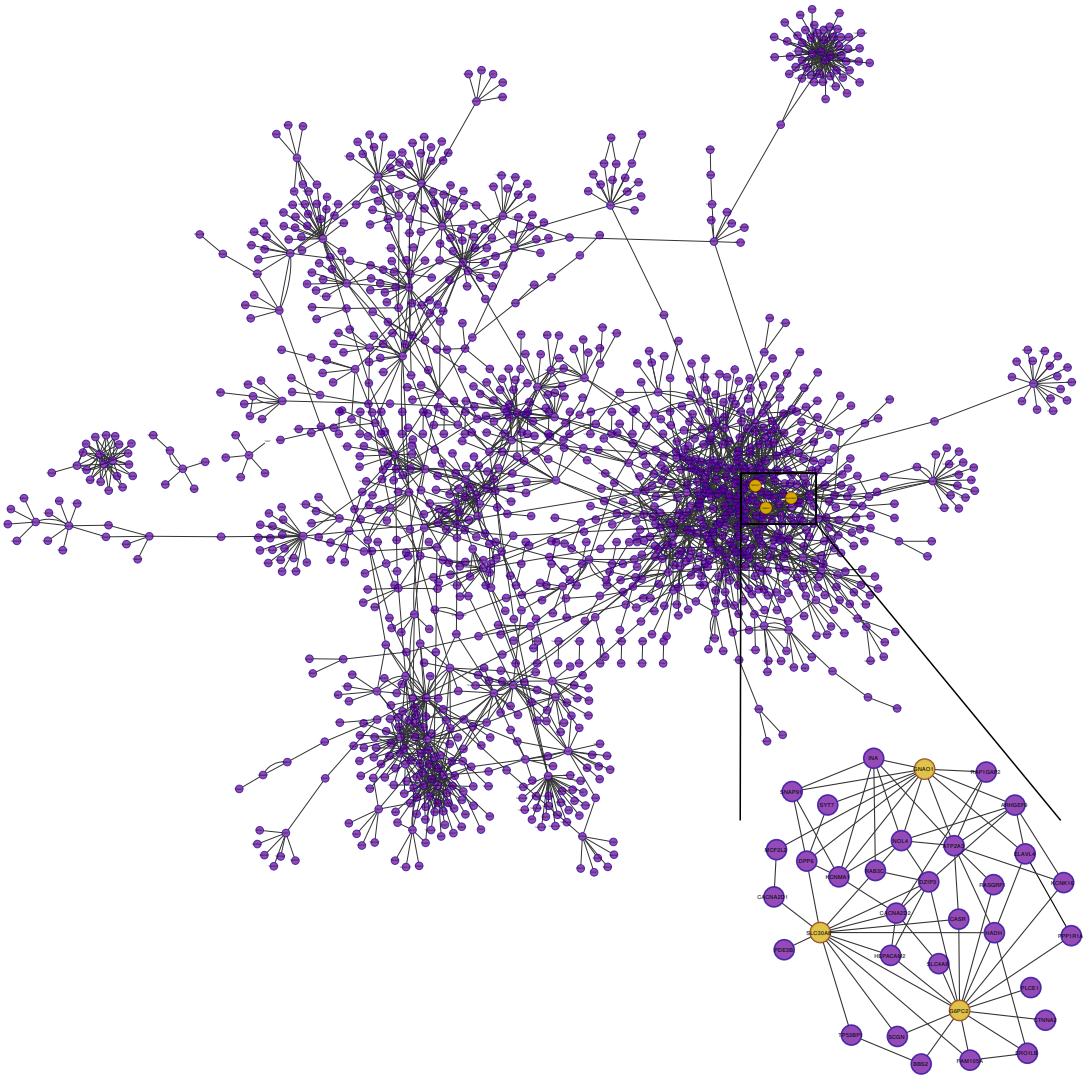


Figure 2: **Regulatory network of diabetes.** The shown network is derived from 64 human islet microarray samples. We used qpgraph to infer the network and applied a NRR cut-off of 0.9 to obtain a network with sufficient module sizes. The yellow circles represent central genes and the purple circles are genes that are co-expressed with the central genes thus forming the module. The three highlighted modules are prioritized because their members are involved in insulin secretion.

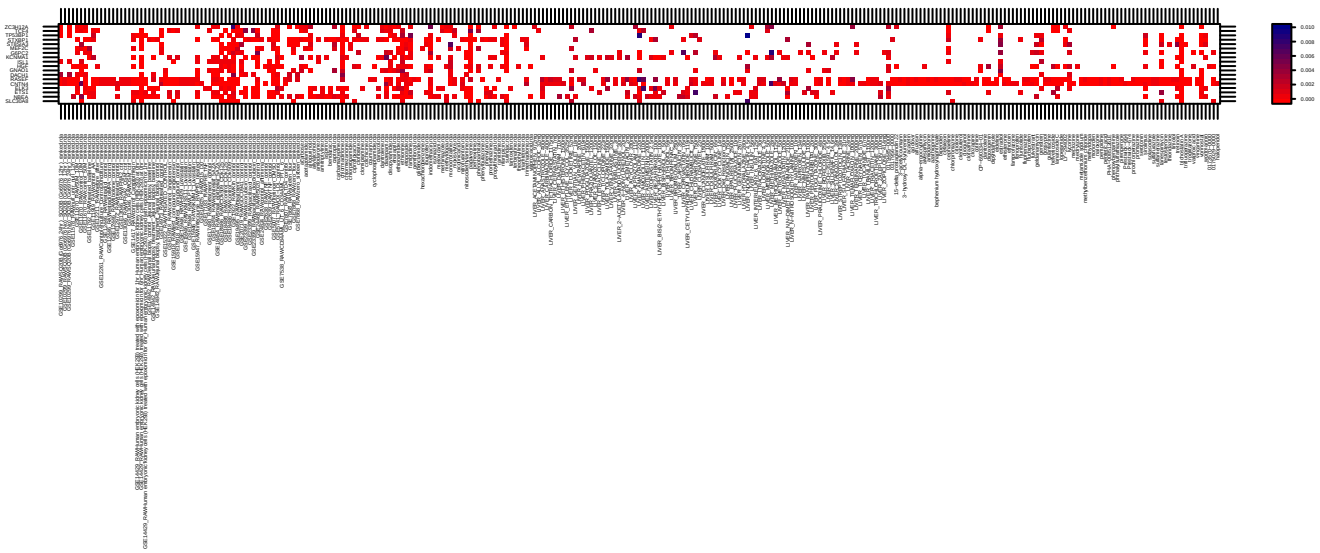


Figure 3: **Compound-Module matrix**. Shown are the 18 (out of 90) prioritized modules. The 18 modules are targeted by 289 compounds.

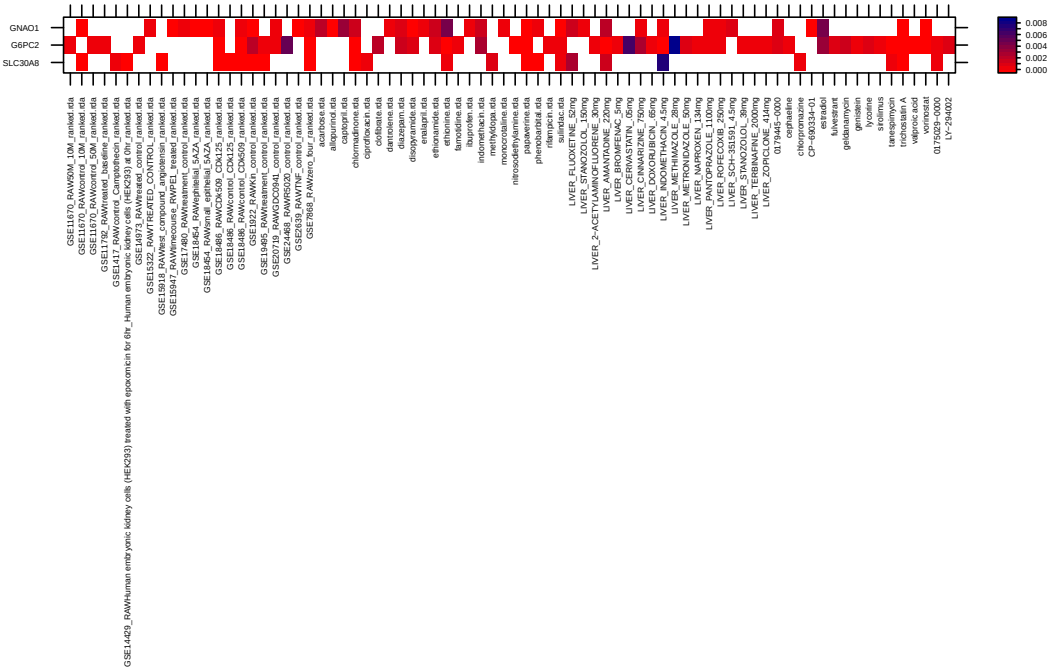


Figure 4: **Compound-Module matrix of the three prioritized modules.** The three modules are prioritized because they are most relevant for insulin secretion and T2D. The targeting 78 compounds are further reduced by discarding compounds that 1) are already repositioned for T2D, 2) unlikely to have an effect on insulin secretion, 3) are toxic.

7 Tables

Table 1: Modules represented by the central gene and their GUILD prioritization scores are shown. Only modules having a p-values < 0.05 are selected.

	real_scores	permuted_scores
TP53BP1	0.0922	0.0459
SLC30A8	0.0937	0.0476
G6PC2	0.0966	0.0484
RASEF	0.0976	0.0380
TCF4	0.1057	0.0296
GNAO1	0.1087	0.0327
CNTN4	0.1104	0.0315
STXBP1	0.1104	0.0118
NBEA	0.1138	0.0043
ELK3	0.1164	0.0048
ISL1	0.1326	0.0142
ZC3H12A	0.1405	0.0058
ETS1	0.1570	0.0027
ST8SIA3	0.1583	0.0011
KCNMA1	0.1672	0.0006
DACH1	0.1721	0.0012
HGF	0.1765	0.0020
MEF2C	0.2166	0.0027

Table 2: The three prioritized modules and their member genes are shown. Each central gene is co-expressed with a number of genes contained in the module.

Central gene	**Co-expressed genes**
SLC30A8	SCGN, RAB3C, DPP6, HEPACAM2, HADH, PDE3B, FAM105A, G6PC2, TP53BP1, CACNA2D1, CASR, CACNA2D2, DZIP3
G6PC2	DHRS2, TMEM200A, CTNNA2, SAMD3, ALCAM, BPIFC, GAD2, GLP1R, HADH, IAPP, PLCXD3, NKX6-1, PAM, ST8SIA3, FAM105A, ERO1LB, SYT13, VAT1L, ROBO2, SYP, SYT4, UCHL1
GNAO1	TSPAN2, ENAM, CPLX2, ACVR1C, DACH1, SPTSSB, ETV1, KCNJ3, MEIS2, NKX2-2, PAX6, PCSK2, PGR, SLC7A14, TM4SF4, MAFB

References

Albert, R., H. Jeong, and A. Barabási, Error and attack tolerance of complex networks, *Nature*, *406*, 378–382, 2000.

Amberger, J., et al., McKusick's online mendelian inheritance in man (OMIM(R)), *Nucleic Acids Research*, *37*, D793–D796, 2009.

Ashburn, T. T., and K. B. Thor, Drug repositioning: identifying and developing new uses for existing drugs, *Nature Reviews. Drug Discovery*, *3*, 673–683, 2004, PMID: 15286734.

Barabasi, A., and R. Albert, Emergence of scaling in random networks, *Science*, *286*, 509–512, 1999.

Barabasi, A., N. Gulbahce, and J. Loscalzo, Network medicine: a network-based approach to human disease, *Nat Rev Genet*, *12*, 56–68, 2011.

Barrett, T., et al., NCBI GEO: archive for functional genomics data sets–10 years on, *Nucleic Acids Research*, *39*, D1005–1010, 2011, PMID: 21097893.

Barter, P. J., et al., Effect of torcetrapib on glucose, insulin, and hemoglobin a1c in subjects in the investigation of lipid level management to understand its impact in atherosclerotic events (ILLUMINATE) trial, *Circulation*, 2011.

Benjamini, Y., and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Jour of the Royal Stat Soc*, *57*, 289–300, 1995.

Bolstad, B. M., Low-level analysis of high-density oligonucleotide array data: Background, normalization and summarization, Ph.D. thesis, University of Waikato, 2004.

Brazma, A., et al., ArrayExpress–a public repository for microarray gene expression data at the EBI., *Nucleic acids research*, *31*, 68–71, 2003.

Breitkreutz, B., et al., The BioGRID interaction database: 2008 update, *Nucleic Acids Research*, *36*, D637–D640, 2008, PMID: 18000002 PMCID: PMC2238873.

Briand, F., et al., CETP inhibitor torcetrapib promotes reverse cholesterol transport in obese insulin-resistant CETP-ApoB100 transgenic mice, *Clinical and Translational Science*, *4*, 414–420, 2011, PMID: 22212222.

Castelo, R., and A. Roverato, A robust procedure for gaussian graphical model search from microarray data with p larger than n, *J Mach Learn Res*, *7*, 2621–2650, 2006.

Castelo, R., and A. Roverato, Reverse engineering molecular regulatory networks from microarray data with qp-graphs, *J Comput Biol*, *16*, 213–27, 2009.

Chong, C. R., and D. J. Sullivan, New uses for old drugs, *Nature*, *448*, 645–646, 2007.

Conforti, F., et al., In vitro activities of citrus medica l. cv. diamante (Diamante citron) relevant to treatment of diabetes and alzheimer's disease, *Phytotherapy Research*, *21*, 427–433, 2007.

Crouch, M. A., I. N. Mefford, and E. U. Wade, Proton pump inhibitor therapy associated with lower glycosylated hemoglobin levels in type 2 diabetes, *The Journal of the American Board of Family Medicine*, *25*, 50–54, 2012.

Crunkhorn, S., and M. Patti, Links between thyroid hormone action, oxidative metabolism, and diabetes risk?, *Thyroid*, *18*, 227–237, 2008.

Dudley, J. T., et al., Disease signatures are robust across tissues and experiments, *Molecular Systems Biology*, *5*, 2009.

Dudley, J. T., et al., Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease, *Science Translational Medicine*, *3*, 96ra76, 2011.

Flicek, P., et al., Ensembl 2011, *Nucleic Acids Research*, *39*, D800–D806, 2010.

Garcia-Garcia, J., et al., Biana: a software framework for compiling biological interactions and analyzing networks, *BMC Bioinformatics*, *11*, 56, 2010, PMID: 20105306 PMCID: PMC3098100.

Gaulton, K. J., et al., A map of open chromatin in human pancreatic islets, *Nat Genet*, *42*, 255–259, 2010.

Giménez-Bastida, J. A., et al., A citrus extract containing flavanones represses plasminogen activator inhibitor-1 (PAI-1) expression and regulates multiple inflammatory, tissue repair, and fibrosis genes in human colon fibroblasts, *Journal of Agricultural and Food Chemistry*, *57*, 9305–9315, 2009, PMID: 19728713.

Goh, K., et al., The human disease network, *Proceedings of the National Academy of Sciences*, *104*, 8685–8690, 2007.

Güldener, U., et al., MPact: the MIPS protein interaction resource on yeast, *Nucleic Acids Research*, *34*, D436–441, 2006, PMID: 16381906.

Guney, E., and B. Oliva, Toward PWAS: discovering pathways associated with human disorders, *BMC Bioinformatics*, *12*, A12, 2011.

Han, J. J., et al., Evidence for dynamically organized modularity in the yeast protein-protein interaction network, *Nature*, *430*, 88–93, 2004.

Harrison, C., Translational genetics: Signatures for drug repositioning, *Nature Reviews Genetics*, *12*, 668–669, 2011.

Hopkins, A. L., Network pharmacology: the next paradigm in drug discovery, *Nature Chemical Biology*, *4*, 682–690, 2008, PMID: 18936753.

Hwang, I. K., et al., Effects of methimazole on the onset of type 2 diabetes in leptin receptor-deficient rats, *The Journal of Veterinary Medical Science / the Japanese Society of Veterinary Science*, *71*, 275–280, 2009, PMID: 19346693.

Iorio, F., et al., Discovery of drug mode of action and drug repositioning from transcriptional responses, *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 14,621–14,626, 2010, PMID: 20679242.

Irizarry, R. A., et al., Summaries of affymetrix GeneChip probe level data, *Nucleic Acids Res*, *31*, e15, 2003.

Irizarry, R. A., et al., Gene set enrichment analysis made simple, *Statistical Methods in Medical Research*, *18*, 565–575, 2009, PMID: 20048385.

Iskar, M., et al., Drug-induced regulation of target expression, *PLoS Computational Biology*, *6*, 2010, PMID: 20838579.

Jeong, H., et al., Lethality and centrality in protein networks, *Nature*, *411*, 41–42, 2001.

Kaufman, R. J., Beta-cell failure, stress, and type 2 diabetes, *The New England Journal of Medicine*, *365*, 1931–1933, 2011, PMID: 22087686.

Kerrien, S., et al., IntAct–open source resource for molecular interaction data, *Nucleic Acids Research*, *35*, D561–D565, 2007.

Keshava Prasad, T. S., et al., Human protein reference database–2009 update, *Nucleic Acids Research*, *37*, D767–772, 2009, PMID: 18988627.

Lamb, J., et al., The connectivity map: Using Gene-Expression signatures to connect small molecules, genes, and disease, *Science*, *313*, 1929 –1935, 2006.

Lauritzen, S. L., *Graphical Models*, Oxford University Press, 1996.

Leek, J. T., et al., Tackling the widespread and critical impact of batch effects in high-throughput data, *Nature Reviews Genetics*, *11*, 733–739, 2010.

Loscalzo, J., and A. Barabasi, Systems biology and the future of medicine, *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *3*, 619–627, 2011.

Lukk, M., et al., A global map of human gene expression, *Nature Biotechnology*, *28*, 322–324, 2010.

Lum, P. Y., J. M. Derry, and E. E. Schadt, Integrative genomics and drug development, *Pharmacogenomics*, *10*, 203–212, 2009.

Lussier, Y. A., and J. L. Chen, The emergence of Genome-Based drug repositioning, *Science Translational Medicine*, *3*, 96ps35, 2011.

Matys, V., et al., TRANSFAC: transcriptional regulation, from patterns to profiles, *Nucleic Acids Research*, *31*, 374–378, 2003, PMID: 12520026.

Natsoulis, G., et al., The liver pharmacological and xenobiotic gene response repertoire, *Molecular Systems Biology*, *4*, 175, 2008, PMID: 18364709.

Ooi, S. L., et al., Global synthetic-lethality analysis and yeast functional profiling, *Trends in Genetics: TIG*, *22*, 56–63, 2006, PMID: 16309778.

Pujol, A., et al., Unveiling the role of network and systems biology in drug discovery, *Trends in Pharmacological Sciences*, *31*, 115–123, 2010, PMID: 20117850.

Rooman, I., J. Lardon, and L. Bouwens, Gastrin stimulates ?-Cell neogenesis and increases islet mass from transdifferentiated but not from normal exocrine pancreas tissue, *Diabetes*, *51*, 686–690, 2002.

Salwinski, L., The database of interacting proteins: 2004 update, *Nucleic Acids Research*, *32*, 449D–451, 2004.

Sanseau, P., et al., Use of genome-wide association studies for drug repositioning, *Nature Biotechnology*, *30*, 317–320, 2012.

Schadt, E. E., S. H. Friend, and D. A. Shaywitz, A network view of disease and compound screening, *Nat Rev Drug Discov*, *8*, 286–295, 2009.

Shigemizu, D., et al., Using functional signatures to identify repositioned drugs for breast, myelogenous leukemia and prostate cancer, *PLoS Comput Biol*, *8*, e1002,347, 2012.

Sirota, M., et al., Discovery and preclinical validation of drug indications using compendia of public gene expression data, *Science Translational Medicine*, *3*, 96ra77, 2011.

Smyth, G. K., Linear models and empirical bayes methods for assessing differential expression in microarray experiments, *Statistical Applications in Genetics and Molecular Biology*, *3*, 2004.

Tamura, K., et al., PROFILING OF GENE EXPRESSION IN RAT LIVER AND RAT PRIMARY CULTURED HEPATOCYTES TREATED WITH PEROXISOME PROLIFERATORS, *The Journal of Toxicological Sciences*, *31*, 471–490, 2006.

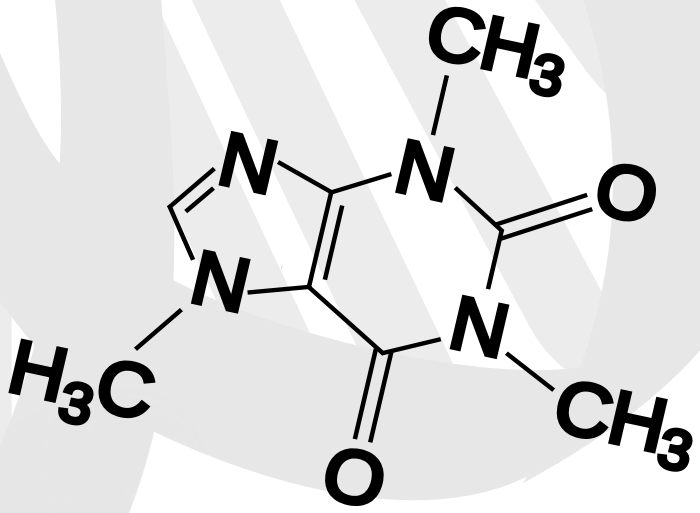
Team R, R. D. C., *R: A Language and Environment for Statistical Computing*, Vienna, Austria, 2010, ISBN 3-900051-07-0.

Watts, D. J., and S. H. Strogatz, Collective dynamics of 'small-world' networks, *Nature*, *393*, 440–442, 1998.

Yu, H., et al., The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics, *PLoS Comput Biol*, *3*, e59, 2007.

CHAPTER 4

DISCUSSION



Current high-throughput genomics technologies offer a systems-wide view on the interactions of molecules in complex biological processes. The generation of heterogeneous large-scale data sets requires new methods and strategies that are able to unveil pathways and networks that underlie complex physiological processes. It becomes evident that intermediate molecular phenotypes such as gene expression levels can be used to infer networks and pathways that are useful proxies for the disease phenotype. This thesis contributes with two pathway-centric approaches to the analysis of high-throughput genomics data. For this, we developed a new gene set enrichment method that is different from other GSE methods in that it constitutes a starting point to open-ended analyses. Further, we demonstrated how a pathway-centric view on disease combined with compound exposed transcriptomic data can be used to find new potential compounds for a complex polygenic disorder such as type 2 diabetes (T2D).

4.1 GSVA - Gene Set Variation Analysis

It becomes more and more important to be able to integrate different sources of data to complete the view on the complex processes which underlie a disease. Complex phenotypes are driven by groups of genes (gene set) working together in regulatory pathways. Method development has to keep up with the ever-increasing amounts of data being generated and assist in the downstream analyses. Further, data processing and interpretation have to be facilitated to extract new knowledge. For this, we developed Gene Set Variation Analysis (GSVA). The idea of analyzing coordinated changes in expression of gene sets is already around for some years. Many extensions and variations of the original method are available.

However, there is no method that allows for the analysis of individual samples by taking gene-specific effects into account and monitor the variation of the pathways per individual throughout the sample population for microarray and RNA-seq data equally. Thus, the motivation for GSVA lies in filling this particular gap. Because GSVA transforms gene expression profiles into pathway expression profiles, a pathway level analysis with different biological or clinical data is straightforward. In our publication, we demonstrate how pathways can be used as signatures to infer subtypes in brain cancer and how to identify pathways predictive of survival. Further, we show cross-talk associations of pathways within leukemia and ovarian cancer. The topology of these networks provides distinguishable structures that aid in dissecting disease pathways.

GSVA for RNA-seq

Many methods for network inference or gene set analysis are available for microarray data, but are not directly applicable to RNA-seq data. Reasons lie in the discrete nature of RNA-seq measurements or gene length bias, that is, current second generation sequencing technologies give longer transcripts higher counts at the same expression level. Normalization methods, such as Conditional Quantile Normalization (CQN) [*Hansen et al.*, 2012] address this shortcoming and correct for additional biases, such as GC content. Standard differential expression analysis strategies, as available for microarray data, are not completely established yet for RNA-seq, which adds another complication to downstream analysis tools, such as gene set enrichment. For these methods to work, they require an estimate of the gene expression level in the different conditions that are studied. GSVA is able to the estimate gene expression level statistics in analogous ways for microarrays and RNA-seq. For example, a RNA-seq experiment with two or more phenotypes can easily be studied in a pathway-centric setting by employing GSVA. For this, data pre-processing and normalization should be done before the use of GSVA. Expression changes are reflected in the pathway-level expression matrix providing GSVA scores per individual sample. The GSVA scores are approximately normally distributed and continuous, hence, conventional methods, such as limma can be applied to the GSVA scores to obtain differentially altered pathways. Moreover, all the features that GSVA offers are available in the exact same way for RNA-seq data, enabling the integration of diverse sources of biological data to establish a more complete picture of the underlying processes.

Limitations

Each gene set enrichment method relies on and is limited by the quality of the microarray data and the gene sets. Some innovations that could improve the sensitivity of the detection of pathway changes are suggested in the following. Microarrays lack sensitivity for genes expressed either at very low or very high levels because of saturation effects of the fluorescence signal. Therefore, microarrays have a much smaller dynamic range than next generation sequencing methods. The new technologies allow for a higher resolution of gene expression profiles but that also means that it would be a good idea to redo the gene set groupings. Some gene sets are derived from microarray experiments, thus the assignment of genes might change with the application of RNA-seq. Moreover, for some diseases there are simply no gene sets available and they have to be derived manually from literature. There is room for the extension of gene set libraries. Also, gene sets that are derived from ChiP-seq data, such as gene expression profiles that change due to methylation or acetylation, would be very interesting.

Future directions

The plethora of microarray data accumulating in public databases such as GEO and array express have enabled the characterization of the diverse molecular patterns underlying biological and physiological processes. With the rapid advance of array technologies and concomitant reduction in cost, novel efforts in this field have turned towards personalized medicine. Several preprocessing methods have been proposed, the most popular method being RMA. However, in a clinical setting data often has to be analyzed separately due to a lack of reference samples and varying admission times. For this purpose, frozen RMA [*McCall et al.*, 2010a] has been developed, a method, that is able to pre-process chips individually. In combination with the barcode algorithm [*Zilliox and Irizarry*, 2007; *McCall et al.*, 2010b] it is possible to estimate the expression state of the genes on the chip without a reference sample. The barcode distinguishes the expression status of a gene as either expressed (1) or not expressed (0).

Analogous to the barcode method a possible extension of GSVA includes the use of a pathway expression profile of a single patient. This personalized gene set variation analysis would assess the enrichment for each pathway on the basis of only one microarray. To be able to do this, a compendium of gene expression profiles has to be collected and normalized using fRMA. Then, these profiles are used to estimate a background distribution that can be compared to the newly obtained pathway expression profile.

4.2 Pathway-centric Genome-based Drug Repositioning

Molecular regulatory networks are sparse and many of them follow a power-law distribution. Power-law means that many nodes have few edges and few nodes have many edges (hubs). Essential proteins are encoded by hubs whereas disease associated proteins are encoded by peripheral genes. Further, disease genes tend to cluster in the same neighborhood which led to the model of disease associated modules. This hypothesis is supported by knock-out studies that showed that it is not enough to target one protein to perturb the disease phenotype, but that several should be targeted to attain a potential effect. This robustness intrinsic to biological networks is due to escape routes and fail-safe mechanisms (redundancy and diversity) [*Pujol et al.*, 2010] present in the cell. Further, disease states are reflected by intermediate phenotypes such as gene expression.

Based on these notions, we developed a pathway-centric genome-based drug-repositioning strategy. For this, we built a transcriptional regulatory network

from human islets (diabetic and healthy) to identify disease associated regulatory modules. To perturb the disease phenotype, we leveraged publicly available data sets measuring the transcriptional response to different compounds. We focused on a specific tissue that is responsible for insulin secretion. This is the first time that a large-scale data set on human islets was used to infer a network uncovering transcriptional regulatory processes leading to T2D. Our methodology differs from previous studies in that it uses disease associated modules instead of single genes as the disease signature. To achieve an effect in the perturbation of the pathophenotype, we aimed at compounds that target multiple members of the disease module.

What makes our strategy special is that it is comprehensive and unbiased. Unbiased means in this case that we do not rely on pre-defined modules (i.e. gene sets) but infer the modules directly from the gene expression measurements. The study is comprehensive because it uses all possible modules and compound exposed transcriptomics data that were publicly available.

Limitations

In the beginning of this study we had in mind to compare two networks derived from healthy and diabetic patients, respectively. Unfortunately, we did not have enough samples for the diabetic patients, and some of the healthy patients, might not have been healthy at all. Their molecular profile showed onset of T2D, which gave motivation to infer a joint network, in which we would identify modules present in both phenotypes. This allowed us to find modules that were already present in the healthy or T2D patients and the intermediates. However, it would also be nice to extend the data set by a few diabetic samples and sort out the intermediate cases to be able to compare modules and their changes. Especially, to detect which genes have changed modules upon disease onset.

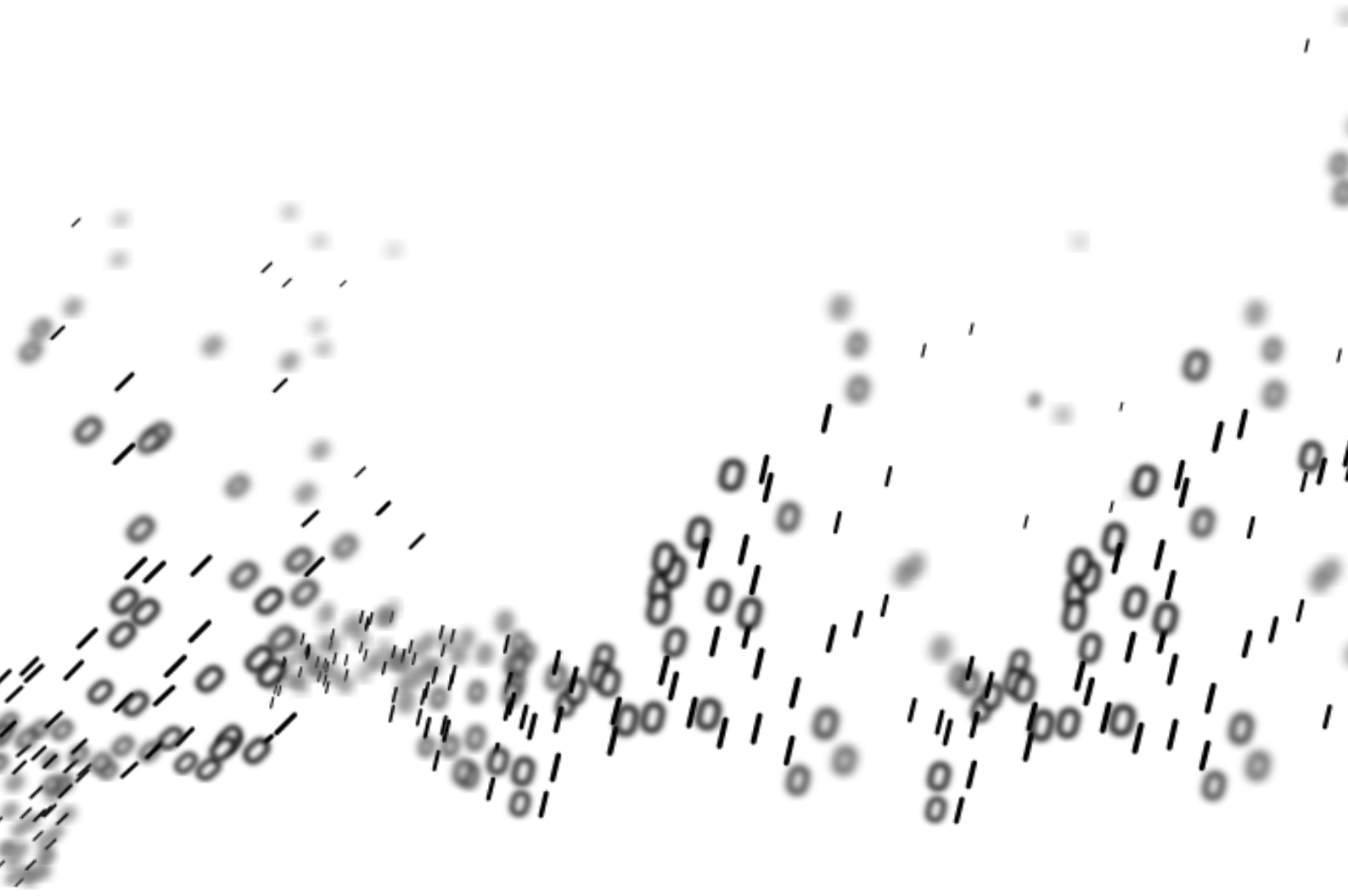
Our strategy does not identify the actual targets of the compounds, only the whole regulatory disease associated module. The following ideas are mentioned for follow-up investigations. The limiting factors of these ideas are usually the data that are not available or that still have to be generated. One logic next step would be to identify the genes that are directly targeted by the compounds. For this, a diverse range of data that have been perturbed in different ways, is necessary to train a network and for this a data set of the human islet data exposed to one drug as a test set would be required [*Bernardo et al.*, 2005]. With these data and the algorithm proposed in [*Bernardo et al.*, 2005], estimates of the direct targets of the drug could be made.

Future directions

The reductionist view of biology has changed over the last two decades into a more systems-level view on biology. A change of entity from single target to multiple targets in form of pathways or subnetworks has occurred leading to new exciting possibilities and revolutions in biology. The most likely applications for network biology include the prediction of drug networks and drug repositioning [*Pujol et al.*, 2010]. Drug networks describe the targets affected by a drug, which can be spread over an entire network and do not have to follow a biological pathway. For example, once the key players for a certain disease in a network are identified, drugs that target these can be administered. Also, networks could be further used to identify drug associations to help estimatating multi-drug and possible side-effects.

CHAPTER 5

CONCLUSIONS



The ultimate goal of this thesis was to provide methodologies that aid in the identification of concerted changes of pathways that underlie disrupted processes in a disease.

The main contributions of this thesis can be summarized as follows:

- We developed methodologies to extract information of a vast amount of high-throughput genomics data at the pathway-level.
- We developed Gene Set Variation Analysis (GSVA), a pathway-centric method that allows to integrate heterogeneous data derived from microarray and RNA-seq gene expression data.
  - GSVA facilitates the organization and condensation of information of genes into gene sets and enables pathway-centric downstream analyses.
  - We demonstrated how GSVA can be used to find disease pathways, infer subtypes in cancer and integrate clinical data with pathways.
  - GSVA is available as a Bioconductor package facilitating the integration into existing workflows.
- We developed a pathway-centric drug-repositioning strategy to identify disease associated modules in human islets and to find potential therapeutics that have a positive effect on insulin secretion.
  - We prioritized three modules related to the dysfunction of insulin secretion.
  - We found four drugs, methimazole, pantoprazole, bitter orange extract and torcetrapib, as potential therapeutics that could have an effect on insulin secretion.
- We have shown how, within a large collaborative effort, the application of co-expression analysis, TFBS analysis and gene regulatory module prioritization can help to identify a new potential therapeutic target for T2D.

Appendices

APPENDIX A

OVEREXPRESSION OF SECRETED FRIZZLED-RELATED PEPTIDE-4 CONTRIBUTES TO TYPE 2-DIABETES

This contribution was in revision at the time the thesis was submitted.

Mahdi T, Hanzelmann S, Salehi A, Muhammed SJ, Reinbothe TM, Tang Y, et al. Secreted frizzled-related protein 4 reduces insulin secretion and is overexpressed in type 2 diabetes. Cell Metab. 2012 Nov 7;16(5):625-33.

Mahdi T, Hanzelmann S, Salehi A, Muhammed SJ, Reinbothe TM, Tang Y, et al. Secreted frizzled-related protein 4 reduces insulin secretion and is overexpressed in type 2 diabetes. Supplemental information. Cell Metab. 2012 Nov 7;16(5):625-33.

BIBLIOGRAPHY

Aerts, S., et al., Toucan: Deciphering the Cis-Regulatory logic of coregulated genes, *Nucleic Acids Research*, *31*, 1753–1764, 2003.

Agarwal, A., et al., Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays, *BMC Genomics*, *11*, 383, 2010, PMID: 20565764 PMCID: 3091629.

Albert, R., Scale-free networks in cell biology, *Journal of Cell Science*, *118*, 4947–4957, 2005, PMID: 16254242.

Albert, R., H. Jeong, and A. Barabási, Error and attack tolerance of complex networks, *Nature*, *406*, 378–382, 2000.

Alles, M. C., et al., Meta-Analysis and gene set enrichment relative to ER status reveal elevated activity of MYC and E2F in the "Basal" breast cancer subgroup, *PLoS ONE*, *4*, e4710, 2009.

Alter, O., P. O. Brown, and D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling, *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 10,101–10,106, 2000, PMID: 10963673.

Altschul, S. F., et al., Basic local alignment search tool, *Journal of molecular biology*, *215*, 403–410, 1990, PMID: 2231712.

Anders, S., and W. Huber, Differential expression analysis for sequence count data, *Genome Biology*, *11*, R106, 2010.

Andrechek, E. R., et al., Patterns of cell signaling pathway activation that characterize mammary development, *Development*, *135*, 2403–2413, 2008.

Arrell, D. K., and A. Terzic, Network systems biology for drug discovery, *Clinical Pharmacology & Therapeutics*, *88*, 120–125, 2010.

Ashburn, T. T., and K. B. Thor, Drug repositioning: identifying and developing new uses for existing drugs, *Nature Reviews. Drug Discovery*, *3*, 673–683, 2004, PMID: 15286734.

Ashburner, M., et al., Gene ontology: tool for the unification of biology, *Nature Genetics*, *25*, 25–29, 2000.

Babu, M. M., et al., Structure and evolution of transcriptional regulatory networks, *Current Opinion in Structural Biology*, *14*, 283–291, 2004.

Baggerly, K. A., K. R. Coombes, and E. S. Neeley, Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer, *Journal of Clinical Oncology*, *26*, 1186–1187, 2008.

Baginsky, S., et al., Gene expression analysis, proteomics, and network discovery, *Plant Physiology*, *152*, 402–410, 2010.

Bailey, T. L., and C. Elkan, Unsupervised learning of multiple motifs in biopolymers using expectation maximization, in *Machine Learning*, p. 51–80, 1995.

Barabasi, A., and R. Albert, Emergence of scaling in random networks, *Science*, *286*, 509–512, 1999.

Barabasi, A., and Z. N. Oltvai, Network biology: understanding the cell's functional organization, *Nature Reviews. Genetics*, *5*, 101–113, 2004, PMID: 14735121.

Barabasi, A., N. Gulbahce, and J. Loscalzo, Network medicine: a network-based approach to human disease, *Nature Reviews Genetics*, *12*, 56–68, 2011.

Barbie, D. A., et al., Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1, *Nature*, *462*, 108–112, 2009.

Barretina, J., et al., The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity, *Nature*, *483*, 603–607, 2012.

Barrett, L. W., S. Fletcher, and S. D. Wilton, Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements, *Cellular and Molecular Life Sciences: CMLS*, 2012, PMID: 22538991.

Barrett, T., et al., NCBI GEO: archive for functional genomics data sets–10 years on, *Nucleic Acids Research*, *39*, D1005–1010, 2011, PMID: 21097893.

Basso, K., et al., Reverse engineering of regulatory networks in human b cells, *Nature Genetics*, *37*, 382–390, 2005.

Beadle, G. W., and E. L. Tatum, Genetic control of biochemical reactions in neurospora, *Proceedings of the National Academy of Sciences*, *27*, 499–506, 1941.

Bellora, N., D. Farré, and M. M. Albà, Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters, *BMC Genomics*, *8*, 459, 2007, PMID: 18078513 PMCID: PMC2249607.

Benito, M., et al., Adjustment of systematic microarray data biases, *Bioinformatics (Oxford, England)*, *20*, 105–114, 2004, PMID: 14693816.

Bennett, S. T., et al., Toward the 1,000 dollars human genome, *Pharmacogenomics*, *6*, 373–382, 2005, PMID: 16004555.

Bernardo, D. d., et al., Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks, *Nature Biotechnology*, *23*, 377–383, 2005.

Birney, E., et al., Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*, *447*, 799–816, 2007.

Bleicher, K. H., et al., Hit and lead generation: beyond high-throughput screening, *Nature Reviews Drug Discovery*, *2*, 369–378, 2003.

Bolstad, B., et al., A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics*, *19*, 185–193, 2003.

Bolstad, B. M., Low-level analysis of high-density oligonucleotide array data: Background, normalization and summarization, Ph.D. thesis, University of Waikato, 2004.

Bossi, A., and B. Lehner, Tissue specificity and the human protein interaction network, *Molecular Systems Biology*, *5*, 260, 2009, PMID: 19357639.

Bullard, J. H., et al., Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments, *BMC Bioinformatics*, *11*, 94, 2010.

Butina, D., M. D. Segall, and K. Frankcombe, Predicting ADME properties in silico: methods and models, *Drug Discovery Today*, *7*, S83–88, 2002, PMID: 12047885.

Butte, A. J., et al., Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks, *Proceedings of the National Academy of Sciences*, *97*, 12,182–12,186, 2000.

Campillos, M., et al., Drug target identification using side-effect similarity, *Science (New York, N.Y.)*, *321*, 263–266, 2008, PMID: 18621671.

Carninci, P., J. Yasuda, and Y. Hayashizaki, Multifaceted mammalian transcriptome, *Current Opinion in Cell Biology*, *20*, 274–280, 2008, PMID: 18468878.

Casella, G., and E. I. George, Explaining the gibbs sampler, *The American Statistician*, *46*, 167–174, 1992, ArticleType: research-article / Full publication date: Aug., 1992 / Copyright © 1992 American Statistical Association.

Castelo, R., and A. Roverato, A robust procedure for gaussian graphical model search from microarray data with p larger than n, *J Mach Learn Res*, *7*, 2621–2650, 2006.

Castelo, R., and A. Roverato, Reverse engineering molecular regulatory networks from microarray data with qp-graphs, *J Comput Biol*, *16*, 213–27, 2009.

Chen, C., et al., Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods, *PLoS ONE*, *6*, e17,238, 2011.

Choi, Y., and C. Kendziorski, Statistical methods for gene set co-expression analysis, *Bioinformatics (Oxford, England)*, *25*, 2780–2786, 2009, PMID: 19689953.

Chong, C. R., and D. J. Sullivan, New uses for old drugs, *Nature*, *448*, 645–646, 2007.

Church, G. M., The personal genome project, *Molecular Systems Biology*, *1*, 2005.

Creighton, C. J., Multiple oncogenic pathway signatures show coordinate expression patterns in human prostate tumors, 2008.

Crick, F., On protein synthesis, *Symposia of the Society for Experimental Biology*, *12*, 138–163, 1958, PMID: 13580867.

Crick, F., Central dogma of molecular biology, *Nature*, *227*, 561–563, 1970.

de la Fuente, A., From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases, *Trends in Genetics: TIG*, *26*, 326–333, 2010, PMID: 20570387.

de la Fuente, A., et al., Discovery of meaningful associations in genomic data using partial correlation coefficients, *Bioinformatics (Oxford, England)*, *20*, 3565–3574, 2004, PMID: 15284096.

Dickson, D., Gene estimate rises as US and UK discuss freedom of access, *Nature*, *401*, 311, 1999, PMID: 10517616.

Diestel, R., *Graph Theory*, 4th ed. 2010. corr. printing 2012 ed., Springer, 2010.

Dijkstra, E., A note on two problems in connection with graphs, *Numerische Mathematik*, *1*, 269–271, 1959.

DiMasi, J. A., R. W. Hansen, and H. G. Grabowski, The price of innovation: new estimates of drug development costs, *Journal of Health Economics*, *22*, 151–185, 2003, PMID: 12606142.

Dobra, A., et al., Sparse graphical models for exploring gene expression data, *Journal of Multivariate Analysis*, *90*, 196–212, 2004.

Dørum, G., et al., Rotation testing in gene set enrichment analysis for small direct comparison experiments, *Stat Apps in Gen and Mol Bio*, *8*, 2009.

Dudley, J. T., et al., Disease signatures are robust across tissues and experiments, *Molecular Systems Biology*, *5*, 2009.

Dudley, J. T., et al., Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease, *Science Translational Medicine*, *3*, 96ra76, 2011.

Edelman, E., et al., Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles, *Bioinformatics*, *22*, e108–116, 2006.

Eisen, M. B., et al., Cluster analysis and display of Genome-Wide expression patterns, *Proceedings of the National Academy of Sciences*, *95*, 14,863–14,868, 1998.

Eisenberg, D., et al., Protein function in the post-genomic era, *Nature*, *405*, 823–826, 2000.

Eltarhouny, S. A., et al., Genes controlling spread of breast cancer to lung "gang of 4", *Experimental Oncology*, *30*, 91–95, 2008, PMID: 18566569.

Erdoes, P., and A. Renyi, On the evolution of random graphs, *PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES*, p. 17—61, 1960.

Euler, L., Solutio problematis ad geometriam situs pertinentis, *Comment. Acad. Sci. U. Petrop*, *8*, 128–140, 1736.

Fisher, R., *The Design of Experiments*, Macmillan Pub Co, 1935.

Flicek, P., et al., Ensembl 2012, *Nucleic Acids Research*, *40*, D84–D90, 2011.

Fodor, S. P., et al., Light-Directed, spatially addressable parallel chemical synthesis, *Science*, *251*, 767–773, 1991.

Freeman, L., A set of measures of centrality based on betweenness, *Sociometry*, *40*, 35–41, 1977.

Friedman, J., T. Hastie, and R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, *Biostatistics (Oxford, England)*, *9*, 432–441, 2008, PMID: 18079126.

Frith, M. C., et al., Detection of functional DNA motifs via statistical Over-Representation, *Nucleic Acids Research*, *32*, 1372–1381, 2004.

Fu, Y., et al., MotifViz: an analysis and visualization tool for motif discovery, *Nucleic acids research*, *32*, W420–423, 2004, PMID: 15215422.

Galas, D. J., M. Eggert, and M. S. Waterman, Rigorous pattern-recognition methods for DNA sequences. analysis of promoter sequences from escherichia coli, *Journal of molecular biology*, *186*, 117–128, 1985, PMID: 3908689.

Gama-Castro, S., et al., RegulonDB version 7.0: transcriptional regulation of escherichia coli k-12 integrated within genetic sensory response units (Gensor units), *Nucleic Acids Research*, *39*, D98–105, 2011, PMID: 21051347.

Gardner, T. S., et al., Inferring genetic networks and identifying compound mode of action via expression profiling, *Science*, *301*, 102–105, 2003.

Gilbert, W., and B. Müller-Hill, Isolation of the lac repressor, *Proceedings of the National Academy of Sciences*, *56*, 1891–1898, 1966.

Giot, L., et al., A protein interaction map of drosophila melanogaster, *Science (New York, N.Y.)*, *302*, 1727–1736, 2003, PMID: 14605208.

Gisler, M., D. Sornette, and R. Woodard, Exuberant innovation: The human genome project, *arXiv:1003.2882*, 2010.

Goeman, J. J., and P. Bühlmann, Analyzing gene expression data in terms of gene sets: methodological issues, *Bioinformatics (Oxford, England)*, *23*, 980–987, 2007, PMID: 17303618.

Goeman, J. J., et al., A global test for groups of genes: testing association with a clinical outcome, *Bioinformatics*, *20*, 93 –99, 2004.

Goh, K., et al., The human disease network, *Proceedings of the National Academy of Sciences*, *104*, 8685–8690, 2007.

Gresham, D., M. J. Dunham, and D. Botstein, Comparing whole genomes using DNA microarrays, *Nature Reviews. Genetics*, *9*, 291–302, 2008, PMID: 18347592.

Gupta, G. P., et al., Mediators of vascular remodelling co-opted for sequential steps in lung metastasis, *Nature*, *446*, 765–770, 2007.

Han, J. J., et al., Evidence for dynamically organized modularity in the yeast protein-protein interaction network, *Nature*, *430*, 88–93, 2004.

Hansen, K. D., R. A. Irizarry, and Z. Wu, Removing technical variability in RNA-seq data using conditional quantile normalization, *Biostatistics*, 2012.

Hardcastle, T. J., and K. A. Kelly, baySeq: empirical bayesian methods for identifying differential expression in sequence count data, *BMC Bioinformatics*, *11*, 422, 2010.

Harrison, C., Translational genetics: Signatures for drug repositioning, *Nature Reviews Genetics*, *12*, 668–669, 2011.

Hert, J., et al., Quantifying the relationships among drug classes, *Journal of Chemical Information and Modeling*, *48*, 755–765, 2008, PMID: 18335977.

Hieter, P., and M. Boguski, Functional genomics: It's all how you read it, *Science*, *278*, 601–602, 1997.

Hoheisel, J. D., Microarray technology: beyond transcript profiling and genotype analysis, *Nature Reviews. Genetics*, *7*, 200–210, 2006, PMID: 16485019.

Homer, N., B. Merriman, and S. F. Nelson, BFAST: an alignment tool for large scale genome resequencing, *PLoS ONE*, *4*, e7767, 2009.

Hopkins, A. L., Network pharmacology: the next paradigm in drug discovery, *Nature Chemical Biology*, *4*, 682–690, 2008, PMID: 18936753.

Hung, J., et al., Gene set enrichment analysis: performance evaluation and usage guidelines, *Briefings in bioinformatics*, *13*, 281–291, 2012, PMID: 21900207.

Iorio, F., et al., Discovery of drug mode of action and drug repositioning from transcriptional responses, *Proceedings of the National Academy of Sciences*, *107*, 14,621–14,626, 2010.

Irizarry, R. A., et al., Summaries of affymetrix GeneChip probe level data, *Nucleic Acids Res*, *31*, e15, 2003a.

Irizarry, R. A., et al., Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics (Oxford, England)*, *4*, 249–264, 2003b, PMID: 12925520.

Irizarry, R. A., et al., Gene set enrichment analysis made simple, *Statistical Methods in Medical Research*, *18*, 565–575, 2009, PMID: 20048385.

Isalan, M., et al., Evolvability and hierarchy in rewired bacterial gene networks, *Nature*, *452*, 840–845, 2008.

Iyer, V. R., et al., Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF, *Nature*, *409*, 533–538, 2001.

Jeong, H., et al., The large-scale organization of metabolic networks, *Nature*, *407*, 651–654, 2000.

Jeong, H., et al., Lethality and centrality in protein networks, *Nature*, *411*, 41–42, 2001.

Jiang, Z., and R. Gentleman, Extensions to gene set enrichment, *Bioinformatics (Oxford, England)*, *23*, 306–313, 2007, PMID: 17127676.

Jin, G., et al., A novel method of transcriptional response analysis to facilitate drug repositioning for cancer therapy, *Cancer Research*, *72*, 33–44, 2012.

Johnson, W. E., C. Li, and A. Rabinovic, Adjusting batch effects in microarray expression data using empirical bayes methods, *Biostatistics (Oxford, England)*, *8*, 118–127, 2007, PMID: 16632515.

Joshi-Tope, G., et al., The genome knowledgebase: a resource for biologists and bioinformaticists, *Cold Spring Harbor Symposia on Quantitative Biology*, *68*, 237–243, 2003, PMID: 15338623.

Jung, K., et al., Comparison of global tests for functional gene sets in Two-Group designs and selection of potentially Effect-Causing genes, *Bioinformatics*, *27*, 1377–1383, 2011.

Kanehisa, M., and S. Goto, KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Research*, *28*, 27–30, 2000, PMID: 10592173.

Keilwagen, J., et al., De novo discovery of differentially abundant transcription factor binding sites including their positional preference, *PLoS Comput Biol*, *7*, e1001,070, 2011.

Keiser, M. J., et al., Relating protein pharmacology by ligand chemistry, *Nature Biotechnology*, *25*, 197–206, 2007, PMID: 17287757.

Keiser, M. J., et al., Predicting new molecular targets for known drugs, *Nature*, *462*, 175–181, 2009, PMID: 19881490.

Kent, W. J., et al., The human genome browser at UCSC, *Genome Research*, *12*, 996–1006, 2002.

Kenzelmann, M., K. Rippe, and J. S. Mattick, RNA: networks & imaging, *Molecular Systems Biology*, *2*, 2006.

Kerr, M. K., and G. A. Churchill, Experimental design for gene expression microarrays, *Biostatistics*, *2*, 183–201, 2001.

Khatri, P., M. Sirota, and A. J. Butte, Ten years of pathway analysis: Current approaches and outstanding challenges, *PLoS Comput Biol*, *8*, e1002,375, 2012.

Kim, T., and P. J. Park, Advances in analysis of transcriptional regulatory networks, *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *3*, 21–35, 2011.

Kola, I., and J. Landis, Can the pharmaceutical industry reduce attrition rates?, *Nature Reviews Drug Discovery*, *3*, 711–716, 2004.

Kuhn, M., et al., A side effect resource to capture phenotypic effects of drugs, *Molecular Systems Biology*, *6*, 343, 2010, PMID: 20087340.

LaFramboise, T., Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances, *Nucleic Acids Research*, *37*, 4181–4193, 2009.

Lamb, J., et al., The connectivity map: Using Gene-Expression signatures to connect small molecules, genes, and disease, *Science*, *313*, 1929 –1935, 2006.

Lander, E. S., Array of hope, *Nature Genetics*, *21*, 3–4, 1999.

Lander, E. S., Initial impact of the sequencing of the human genome, *Nature*, *470*, 187–197, 2011.

Lander, E. S., et al., Initial sequencing and analysis of the human genome, *Nature*, *409*, 860–921, 2001, PMID: 11237011.

Langfelder, P., and S. Horvath, WGCNA: an r package for weighted correlation network analysis, *BMC Bioinformatics*, *9*, 559, 2008.

Langmead, B., et al., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biology*, *10*, R25, 2009.

Lauritzen, S. L., *Graphical Models*, Oxford University Press, 1996.

Lederberg, J., and A. Mccray, The scientist : 'Ome sweet 'Omics– a genealogical treasury of words, *The Scientist*, *17*, 2001.

Lee, E., et al., Inferring pathway activity toward precise disease classification, *PLoS Comput Biol*, *4*, e1000,217, 2008.

Leek, J. T., and J. D. Storey, Capturing heterogeneity in gene expression studies by surrogate variable analysis, *PLoS Genet*, *3*, e161, 2007.

Leek, J. T., et al., Tackling the widespread and critical impact of batch effects in high-throughput data, *Nature Reviews Genetics*, *11*, 733–739, 2010.

Levy, S., et al., The diploid genome sequence of an individual human, *PLoS Biology*, *5*, e254, 2007, PMID: 17803354.

Li, H., and R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics (Oxford, England)*, *25*, 1754–1760, 2009, PMID: 19451168.

Li, H., J. Ruan, and R. Durbin, Mapping short DNA sequencing reads and calling variants using mapping quality scores, *Genome Research*, *18*, 1851–1858, 2008a, PMID: 18714091.

Li, R., et al., SOAP: short oligonucleotide alignment program, *Bioinformatics (Oxford, England)*, *24*, 713–714, 2008b, PMID: 18227114.

Li, R., et al., SOAP2: an improved ultrafast tool for short read alignment, *Bioinformatics (Oxford, England)*, *25*, 1966–1967, 2009, PMID: 19497933.

Li, Y., P. Agarwal, and D. Rajagopalan, A global pathway crosstalk network, *Bioinformatics*, *24*, 1442–1447, 2008c.

Liu, M., et al., Network-Based analysis of affected biological processes in type 2 diabetes models, *PLoS Genet*, *3*, e96, 2007.

Lockhart, D. J., et al., Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nature Biotechnology*, *14*, 1675–1680, 1996, PMID: 9634850.

Loscalzo, J., and A. Barabasi, Systems biology and the future of medicine, *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *3*, 619–627, 2011.

Loscalzo, J., I. Kohane, and A. Barabasi, Human disease classification in the postgenomic era: A complex systems approach to human pathobiology, *Molecular Systems Biology*, *3*, 124, 2007, PMID: 17625512 PMCID: PMC1948102.

Lucito, R., et al., Representational oligonucleotide microarray analysis: A High-Resolution method to detect genome copy number variation, *Genome Research*, *13*, 2291–2305, 2003.

Lukk, M., et al., A global map of human gene expression, *Nature Biotechnology*, *28*, 322–324, 2010.

Lum, P. Y., J. M. Derry, and E. E. Schadt, Integrative genomics and drug development, *Pharmacogenomics*, *10*, 203–212, 2009.

Lussier, Y. A., and J. L. Chen, The emergence of Genome-Based drug repositioning, *Science Translational Medicine*, *3*, 96ps35, 2011.

Ma'ayan, A., Insights into the organization of biochemical regulatory networks using graph theory analyses, *Journal of Biological Chemistry*, *284*, 5451–5455, 2009.

Ma'ayan, A., et al., Formation of regulatory patterns during signal propagation in a mammalian cellular network, *Science*, *309*, 1078–1083, 2005.

Malone, J. H., and B. Oliver, Microarrays, deep sequencing and the true measure of the transcriptome, *BMC Biology*, *9*, 34, 2011.

Mandlik, A., et al., RNA-Seq-Based monitoring of Infection-Linked changes in vibrio cholerae gene expression, *Cell Host & Microbe*, *10*, 165–174, 2011.

Manoli, T., et al., Group testing for pathway analysis improves comparability of different microarray datasets, *Bioinformatics*, *22*, 2500–2506, 2006.

Margolin, A. A., et al., ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC Bioinformatics*, *7 Suppl 1*, S7, 2006, PMID: 16723010.

Marguerat, S., and J. Bähler, RNA-seq: from technology to biology, *Cellular and Molecular Life Sciences: CMLS*, *67*, 569–579, 2010, PMID: 19859660.

Margulies, M., et al., Genome sequencing in microfabricated high-density picolitre reactors, *Nature*, *437*, 376–380, 2005.

Maxam, A. M., and W. Gilbert, A new method for sequencing DNA, *Proceedings of the National Academy of Sciences of the United States of America*, *74*, 560–564, 1977, PMID: 265521.

McCall, M. N., B. M. Bolstad, and R. A. Irizarry, Frozen robust multiarray analysis (fRMA), *Biostatistics*, *11*, 242 –253, 2010a.

McCall, M. N., et al., The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes, *Nucleic Acids Research*, *39*, D1011–D1015, 2010b.

Merico, D., et al., Enrichment map: A Network-Based method for Gene-Set enrichment visualization and interpretation, *PLoS ONE*, *5*, e13,984, 2010.

Milner, R. J., and J. G. Sutcliffe, Gene expression in rat brain, *Nucleic Acids Research*, *11*, 5497–5520, 1983.

Milo, R., et al., Network motifs: Simple building blocks of complex networks, *Science*, *298*, 824–827, 2002.

Mitelman, F., B. Johansson, and F. Mertens, The impact of translocations and gene fusions on cancer causation, *Nat Rev Cancer*, *7*, 233–245, 2007.

Mootha, V. K., et al., PGC 1alpha responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, *Nature Genetics*, *34*, 267–273, 2003.

Mortazavi, A., et al., Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nature methods*, *5*, 621–628, 2008, PMID: 18516045.

Nitsch, D., et al., Candidate gene prioritization by network analysis of differential expression using machine learning approaches, *BMC Bioinformatics*, *11*, 460, 2010.

Ogata, H., et al., KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Research*, *27*, 29–34, 1999, PMID: 9847135.

Oldham, M. C., S. Horvath, and D. H. Geschwind, Conservation and evolution of gene coexpression networks in human and chimpanzee brains, *Proceedings of the National Academy of Sciences*, *103*, 17,973 –17,978, 2006.

Ooi, S. L., et al., Global synthetic-lethality analysis and yeast functional profiling, *Trends in Genetics: TIG*, *22*, 56–63, 2006, PMID: 16309778.

Opsahl, T., F. Agneessens, and J. Skvoretz, Node centrality in weighted networks: Generalizing degree and shortest paths, *Social Networks*, *32*, 245–251, 2010.

Orešič, M., et al., Metabolome in schizophrenia and other psychotic disorders: a general population-based study, *Genome Medicine*, *3*, 19, 2011, PMID: 21429189 PMCID: PMC3092104.

Oshlack, A., and M. J. Wakefield, Transcript length bias in RNA-seq data confounds systems biology, *Biology Direct*, *4*, 14, 2009.

Oshlack, A., M. D. Robinson, and M. D. Young, From RNA-seq reads to differential expression results, *Genome biology*, *11*, 220, 2010, PMID: 21176179.

Paolini, G. V., et al., Global mapping of pharmacological space, *Nature Biotechnology*, *24*, 805–815, 2006, PMID: 16841068.

Park, J., et al., The impact of cellular networks on disease comorbidity, *Molecular Systems Biology*, *5*, 262, 2009, PMID: 19357641 PMCID: PMC2683720.

Pastor-Satorras, R., and A. Vespignani, Immunization of complex networks, *Physical Review E*, *65*, 2002.

Perou, C. M., et al., Molecular portraits of human breast tumours, *Nature*, *406*, 747–752, 2000, PMID: 10963602.

Phillips, K. A., S. V. Bebber, and A. M. Issa, Diagnostics and biomarker development: priming the pipeline, *Nature Reviews Drug Discovery*, *5*, 463–469, 2006.

Pickrell, J. K., et al., Understanding mechanisms underlying human gene expression variation with RNA sequencing, *Nature*, *464*, 768–772, 2010.

Pietiläinen, K. H., et al., Association of lipidome remodeling in the adipocyte membrane with acquired obesity in humans, *PLoS Biol*, *9*, e1000,623, 2011.

Pujol, A., et al., Unveiling the role of network and systems biology in drug discovery, *Trends in Pharmacological Sciences*, *31*, 115–123, 2010, PMID: 20117850.

Qiu, X., L. Klebanov, and A. Yakovlev, Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes, *Statistical Applications in Genetics and Molecular Biology*, *4*, Article34, 2005, PMID: 16646853.

Rego, T. G. d., et al., Inferring epigenetic and transcriptional regulation during blood cell development with a mixture of sparse linear models, *Bioinformatics*, 2012.

Rhee, S. Y., et al., Use and misuse of the gene ontology annotations, *Nature Reviews Genetics*, *9*, 509–515, 2008.

Robertson, G., et al., cisRED: a database system for genome-scale computational discovery of regulatory elements, *Nucleic Acids Research*, *34*, D68–73, 2006, PMID: 16381958.

Robinson, M. D., and A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data, *Genome Biology*, *11*, R25, 2010, PMID: 20196867.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth, edgeR: a bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, *26*, 139–140, 2010, PMID: 19910308 PMCID: PMC2796818.

Rumble, S. M., et al., SHRiMP: accurate mapping of short color-space reads, *PLoS Computational Biology*, *5*, e1000,386, 2009, PMID: 19461883.

Sabidussi, G., The centrality of a graph, *Psychometrika*, *31*, 581–603, 1966, PMID: 5232444.

Sanger, F., et al., Nucleotide sequence of bacteriophage phi x174 DNA, *Nature*, *265*, 687–695, 1977, PMID: 870828.

Schadt, E. E., S. H. Friend, and D. A. Shaywitz, A network view of disease and compound screening, *Nature Reviews Drug Discovery*, *8*, 286–295, 2009.

Schadt, E. E., et al., Computational solutions to large-scale data management and analysis, *Nature Reviews Genetics*, *11*, 647–657, 2010.

Scheiber, J., et al., Gaining insight into Off-Target mediated effects of drug candidates with a comprehensive systems chemical biology analysis, *J. Chem. Inf. Model.*, *49*, 308–317, 2009.

Schena, M., et al., Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, *270*, 467–470, 1995.

Schones, D., A. Smith, and M. Zhang, Statistical significance of cis-regulatory modules, *BMC Bioinformatics*, *8*, 19, 2007.

Schulze, A., and J. Downward, Navigating gene expression using microarrays \textbackslashtextbackslashtextbar[mdash]\textbackslashtextbackslashtextbar a technology review, *Nature Cell Biology*, *3*, E190–E195, 2001.

Scott, S., and J. Thompson, Adverse drug reactions, *Anaesthesia & Intensive Care Medicine*, *12*, 319–323, 2011.

Shannon, C., Communication theory of secrecy systems, *Bell System Technical Journal*, *28*, 656–715, 1949.

Shigemizu, D., et al., Using functional signatures to identify repositioned drugs for breast, myelogenous leukemia and prostate cancer, *PLoS Comput Biol*, *8*, e1002,347, 2012.

Sirota, M., et al., Discovery and preclinical validation of drug indications using compendia of public gene expression data, *Science Translational Medicine*, *3*, 96ra77, 2011.

Stormo, G. D., DNA binding sites: representation and discovery, *Bioinformatics (Oxford, England)*, *16*, 16–23, 2000, PMID: 10812473.

Stormo, G. D., and D. S. Fields, Specificity, free energy and information content in protein-DNA interactions, *Trends in biochemical sciences*, *23*, 109–113, 1998, PMID: 9581503.

Subramanian, A., et al., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 15,545–15,550, 2005.

Sultan, M., et al., A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome, *Science*, *321*, 956–960, 2008.

Sysi-Aho, M., et al., Serum lipidomics meets cardiac magnetic resonance imaging: Profiling of subjects at risk of dilated cardiomyopathy, *PLoS ONE*, *6*, e15,744, 2011.

Tamayo, P., et al., Gene set enrichment analysis made right, *arXiv:1110.4128*, 2011.

Tarazona, S., et al., Differential expression in RNA-seq: a matter of depth, *Genome Research*, 2011.

Tenenbaum, J. D., et al., Expression-based pathway signature analysis (EPSA): mining publicly available microarray data for insight into human disease, *BMC Medical Genomics*, *1*, 51, 2008.

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

Tian, L., et al., Discovering statistically significant pathways in expression profiling studies, *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 13,544–13,549, 2005.

Trapnell, C., et al., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks, *Nature Protocols*, *7*, 562–578, 2012.

Veerla, S., M. Ringnér, and M. Höglund, Genome-wide transcription factor binding site/promoter databases for the analysis of gene sets and co-occurrence of transcription factor binding motifs, *BMC Genomics*, *11*, 145, 2010.

Venter, J. C., et al., The sequence of the human genome, *Science*, *291*, 1304–1351, 2001.

Verhaak, R. G. W., et al., Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1, *Cancer Cell*, *17*, 98–110, 2010.

Walker, W. L., et al., Empirical bayes accomodation of batch-effects in microarray data using identical replicate reference samples: application to RNA expression profiling of blood from duchenne muscular dystrophy patients, *BMC Genomics*, *9*, 494, 2008.

Wang, L., et al., DEGseq: an r package for identifying differentially expressed genes from RNA-seq data, *Bioinformatics (Oxford, England)*, *26*, 136–138, 2010, PMID: 19855105.

Wang, Z., M. Gerstein, and M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, *Nature Reviews. Genetics*, *10*, 57–63, 2009, PMID: 19015660.

Wasserman, W. W., and A. Sandelin, Applied bioinformatics for the identification of regulatory elements, *Nature Reviews Genetics*, *5*, 276–287, 2004.

Watson, J. D., and F. H. C. Crick, The structure of dna, *Cold Spring Harbor Symposia on Quantitative Biology*, *18*, 123–131, 1953.

Wilke, R. A., et al., Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges, *Nature Reviews Drug Discovery*, *6*, 904–916, 2007.

Wingender, E., The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation, *Briefings in Bioinformatics*, *9*, 326–332, 2008, PMID: 18436575.

Workman, C. T., and G. D. Stormo, ANN-Spec: a method for discovering transcription factor binding sites with improved specificity, *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, p. 467–478, 2000, PMID: 10902194.

Wu, D., and G. K. Smyth, Camera: A competitive gene set test accounting for Inter-Gene correlation, *Nucleic Acids Research*, 2012.

Wu, D., et al., ROAST: rotation gene set tests for complex microarray experiments, *Bioinformatics (Oxford, England)*, *26*, 2176–2182, 2010, PMID: 20610611.

Young, M. D., et al., Gene ontology analysis for RNA-seq: accounting for selection bias, *Genome Biology*, *11*, R14, 2010.

Yu, H., et al., The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics, *PLoS Comput Biol*, *3*, e59, 2007.

Zhang, B., and S. Horvath, A general framework for weighted gene co-expression network analysis, *Statistical Applications in Genetics and Molecular Biology*, *4*, Article17, 2005, PMID: 16646834.

Zhu, J., B. Zhang, and E. E. Schadt, A systems biology approach to drug discovery, *Advances in Genetics*, *60*, 603–635, 2008, PMID: 18358334.

Zilliox, M. J., and R. A. Irizarry, A gene expression bar code for microarray data, *Nature Methods*, *4*, 911–913, 2007, PMID: 17906632.