Finding Common Patterns in Heterogeneous Perturbation Data

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

KHALID ABNAOF

aus Atbara

Bonn, 2015

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Dr. Holger Fröhlich
2. Gutachter: Prof. Dr. Andreas Weber
3. Gutachter: Prof. Dr. Martin Hofmann-Apitius
4. Gutachter: Prof. Dr. Frank Bertoldi

Tag der Promotion: 6. April 2016 Erscheinungsjahr: 2016

To Souad & Ali,

Acknowledgment

The cooperation projects that led to this thesis were carried out at the Algorithmic Bioinformatics group of the Bonn-Aachen International Center for Information Technology (B-IT) at Bonn University, and the whole work was performed under the invaluable guidance and direction of Univ-Prof. Dr. Holger Fröhlich.

I am very thankful to Holger for introducing me to the exciting field of Bioinformatics, for his strive to avail funding for my position and for his continued training and encouragement through many discussions, suggestions and challenges coupled with unfailing patience.

My collegial group made the working atmosphere at the institute very pleasant for me. The former as well as the current members of the group helped me in many ways. My thanks go also to the IT team and the secretary department.

I would like also to express my gratitude to the Neuroallianz consortium, the Open Phacts initiative and the Alma-in-Silico team who funded different phases of my work. I thank all the people from the different cooperation institutes with whom I had the opportunity to work.
In this regard, I'm particularly thankful to Dr. Marc Zimmerman.

I would also like to thank Prof. Dr. Andreas Weber of the Computer Science Institute for taking over as referent and for his valuable help and suggestions.

My appreciation is extended to the other members of the doctoral committee,
Prof. Dr. Martin Hofmann-Apitius of Fraunhofer SCAI
and Prof. Dr. Frank Bertoldi of the Argelander-Institute for Astronomy
for their willingness to act as co-examiner and for their time spent to read and evaluate this thesis.

Last but not least, my family deserves endless thanks for their patience and support.

Abstract

This work investigates and proposes statistical analysis methods for pattern detection in high-throughput data from perturbation experiments in biology and medicine. This is demonstrated in three examples.

The first part of this thesis investigates the transcriptional responses of TGF-β stimulation in different human and mouse cell types based on time-course microarray data from extensive experiments. The used statistical and bioinformatics methods enabled to identify commonly affected biological subsystems across different cell types. In particular the analysis suggests an important role of transcription factors, which appear to have a conserved influence across cell-types and species. Validation via an independent dataset confirms the findings and network analyses suggest explanations, how TGF-β perturbation could lead to the observed effects

The second part investigates pro epileptic markers in microRNA expression profiling data from perturbation-induced pathogenic animal models. Experimental implications resulting in incomplete and censored high-throughput qPCR data impairs the performance of analysis methods. A designated test procedure, which showed higher detection power at lower false positive rates base on simulated data, is proposed to resolve this issue. The method enabled the identification of novel pathogenic relevant miRNAs in epilepsy models.

In the last part of this work a new method for drug-drug similarity assessment based on drug-proteins interaction network and drug pharmacological effects on disease related targets is proposed. The similarity measure, which does not require chemical structure information, is applied within a consensus clustering algorithm to detect useful patterns in a large compound dataset from different diseases. The method produced separated and stable clusters that could not be found using chemical structure-based approaches. Target proteins of compounds falling into one cluster suggested several new compound-target combinations, which could in several cases be confirmed by independent data.

Altogether this thesis demonstrates that advanced analysis methods could help to extract common patterns from complex and seemingly heterogeneous data.

Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

Advances in biotechnology made targeted perturbations and genome-scale measurements possible. Perturbation experiments help understand and characterize biological systems by revealing causal relationships among biomolecular entities (Lin et al., 2005; Nelander et al., 2008; Azmi, 2012) and high-throughput genomics, transcriptomics, proteomics and metabolomics have the potential to identify the functional consequences of induced and natural genetic variations (Jansen, 2003). A single run of such experiments has only limited informative value. Therefore, evidences are gathered from several experiments and statistical methods are employed to facilitate inference and prediction of perturbation responses.

As early as high-throughput technologies for gene expression began to be used routinely, Tilstone (2003) documented the alarming warnings raised by many parties that lack of proper statistical analysis methods could undermine the revolution promised by genomics and biotechnology. Few years later many approaches for mining perturbation high-throughput expression and drug affinity data have been established (Beaumont and Rannala, 2004; VanGuilder et al., 2008). However, the advancing and changing technologies and the vast amounts of data generated still represent great challenges for handling, computational integrating the data and devising appropriate analysis methods.

In molecular biology, a targeted perturbation typically inhibits or activates functions of molecules in the cell (Nelander et al., 2008), e.g. by indel mutation of target DNA, or gain of function e.g. by recruiting transcriptional activation domains (Shalem et al., 2015). Perturbations can be permanent or temporary, global or local (Kravchik et al., 2014), it can also be natural or investigational (Nelander et al., 2008). It is easier to build concise hypotheses and test them under parsimonious assumptions by single-perturbations. However, combinatorial-perturbations experiments are now

increasingly carried out, for example in drug discovery libraries of compounds are tested against a panel of different cancer cell lines (Lamb et al., 2006). Perturbation experiments are used to generate and investigate akin cellular states of medical relevance (e.g. pathological models). Such an approach allows for upfront establishment of causal relationship between genotypes and phenotypes and has potential for e.g. discovering functionally relevant genes that are not mutated or directly affected by the perturbation (Ginsburg and Willard, 2012).

An overview of common perturbation types and their point of action in the cell is given in Figure 1.1. Gene expression can be changed through competition triggered by excess amount of small RNA (Loinger et al., 2012), e.g. by fusion of virus particles into the cell leading to depletion & degradation of active target mRNA and translation inhibition during cell division (Kim et al., 2005; Loinger et al., 2012; Shalem et al., 2015).

Drugs (chemical compounds) and natural protein ligands influence the cell by acting on receptors, transporters or enzymes of cell membrane. Drug perturbation involves changing proteins or compounds when adding ligands in order to determine location of binding, affinity of the ligand or the structure of the resulting complex (Brader et al., 1997; Williamson, 2013).

Direct manipulation of the genetic makeup via recombinant DNA technology, e.g. CRISPR/Cas-System, can be used to create transgenic organisms that differ from their wild type only in the affected gene (Baltimore et al., 2015; Liang et al., 2015). This involves indel mutation of target DNA leading to complete (knock-out) or partial (knock-down) deletion of the target (Shalem et al., 2015). Thus, the transcription of the affected genes is either reduced or completely stopped so that they are translated into non-functional proteins or not translated at all. Altered targets are typically genes with known sequence whose functions are not fully determined. Genetic modification, specially in human, is tangled with ethical issues (O'Connell et al., 2014; Lanphier et al., 2015).

The result of perturbation ultimately is a discernible phenotype that is different from the wild (naive) type. The effects of perturbation can be assessed through phenotype observations at different levels: (I) by observing features of the discernible phenotype (e.g. growth and differentiation), or by observing indirect genotype measurements (II) by readouts at transcriptic gene regulation level (e.g. from high-throughput microarray and next generation sequencing data), (III) or by assessing down-stream measurement at protein level (Figure 1.1). Perturbation (pathological/treatment) system is compared to the normal (wild type/naive) system by investigating significant changes in defined read-outs (e.g. microscopic phenotype, transcriptome, proteome, metabolome and DNA methylation).

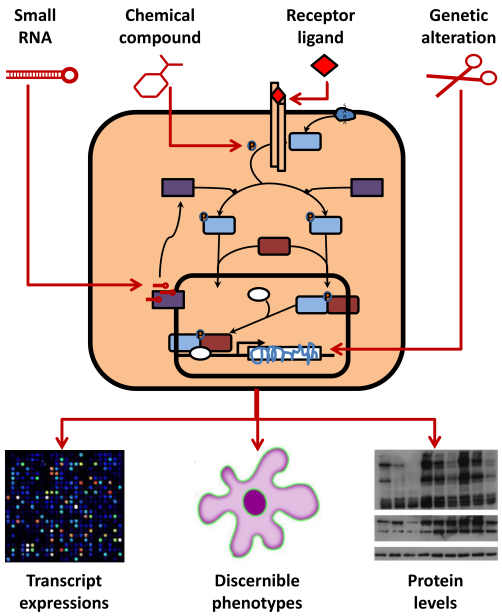

Figure 1.1: An overview of common perturbation types, their point of action in the cell and the observation possibilities for assessment. Small inhibitory RNA change gene expression. Drugs (chemical compounds) and natural protein ligands act on receptors, transporters or enzymes of cell membrane. Direct genetic alteration, e.g. by knockdown genes have diverse functional effects and is widely applied to generate disease models. Perturbation effect can be assessed by: (I) observing the discernible phenotype features (e.g. growth and differentiation), or by observing indirect genotype measurements (II) by readouts at transcriptic level (e.g. from high-throughput data), (III) or by examining down-stream measurement at protein level (modified after Nelander et al. (2008)).

Depending on the experimental setting, phenotypic observations of the different levels are taken at useful and practicable points that represent snapshot of the perturbed biological system. Investigating transcription and protein levels involves observing high numbers of entities (Oliver and Leblanc, 2003). Ever evolving technologies can now provide unbiased quantification of multiple molecular and phenotypic changes across tens of thousands of individual features from perturbed cells simultaneously (Liberali et al., 2014). High-throughput technologies made this feasible on a large scale by performing high numbers of individual experiments in parallel using automation without compromising the quality. These methods can measure cellular responses and regulations by quantitative measures of activity of the elements of the cell (Bork and Koonin, 1996; Quackenbush, 2001; Lonnstedt and Speed, 2002; Yang and Speed,

2002; Maeda et al., 2003). The transcriptom (all from DNA transcribed RNA) is typically measured by microarray, quantitative polymerase chain reaction (qPCR) and high-throughput RNA-sequencing, the genome through DNA-sequencing, mass spectrometry & gel electrophoesis techniques are usually used to determine proteom & metabolom (Parmigiani et al., 2003; Stingele et al., 2012).

However, these technologies introduce variation into the acquired data that is not purely due to the biological impact. variations have many sources and make the information value of a single perturbation experiment limited and unreliable. Therefore, evidences are gathered from adequate repetitions of the experiment (Kuo et al., 2000; Liang et al., 2003; Lin et al., 2005). Biological replicates, for example, are raised to accounts for biological variation and technical replicate to compensate technical variation.

The sheer volume and high dimensionality of high-throughput data make advance computational statistical methods indispensable to extract meaningful patterns that help elucidate the real biological systems. Replicates enable statistical inference through evaluation of variability within and between experimental conditions and advanced normalization techniques allow for accurate estimation of absolute and relative expressions.

Advancing technologies and innovations permit for intricate experimental settings and provide for generating data of different types and formats. This makes analyzing the data more challenging. For instance, a lack of consensus still exists on how to perform, analyze and interpret qPCR experiments (Bustin et al., 2009). This problem is exacerbated e.g. by introducing multiple chips for high-throughput qPCR. It is also often the case that multiple and heterogeneous data sets need to be evaluated and integrated in the context of the same research questions. Thus, the process of choosing proper analysis methods for the data and accurately implementing them is still challenging. In many cases novel approaches need to be devised or generic existing methods need to be adapted to cope with the data and research assumptions. Altogether, this emphasize the essence of coordinating experimental and analytical methods. Experimental validation of analysis findings is not always possible, therefore, analytical validation methods are required. Literature & text mining methods and functional analysis approaches can help to structure and validate information gleaned from the data.

Secondary predictive methods can use perturbation effects to computationally reverse engineer individual biological processes (Froehlich et al., 2007; Zacher and Abnaof et al., 2012). These methods are outside the scope of this thesis, however, they greatly depend on the outcome of the primary methods described here.

From the above we conclude that perturbation experiments allow gaining insights about the functions of biological systems. However, major challenges lie in consolidating the complex read outs of various such experiments and in identifying meaningful patterns. The aim of this work is to explore, devise and develop methodologies for efficient extraction of biologically relevant patterns from diverse perturbation experiments. The stress lies in the coordination and adaptation of statistical methods to varieties of high-throughput data types and formats.

This thesis is organized as follows:

Chapter 2 gives the biological and technological background of the relevant categories of high-throughput data which are analyzed in the context of this thesis. These include microarray technology, qPCR arrays and compound biological activity data. Post experiment advanced normalization techniques and pre-processing methods for these data types are discussed.

Chapter 3 covers the most relevant aspects of established statistical methods and bioinformatics techniques that have been used throughout this work. Many methods and applications have been discussed in this chapter and illustrated in variety of applications in the subsequent chapters. This include differential expression analysis methods, clustering approaches, functional analyses approaches for gene sets & cluster groups in e.g. gene ontologies & biochemical pathways and prediction methods for transcription factors, sequence patterns and protein domains.

Chapter 4 shows, how statistical and bioinformatics methods can be used to obtain biologically relevant insights from a collection of complex perturbation experiments. More specifically, the transcriptional response of TGF-beta stimulation in different human and murine cell types is investigated using extensive microarray experiments (Gudrun and Abnaof et al., 2013; Abnaof et al., 2014). I demonstrate that differential time-course analysis together with biological sequence and network analysis methods can unravel commonly affected biological sub-systems across different cell types. A devised visualization tool demonstrates efficiency in integrating comprehensive results of various functional analyses and helped discovering patterns at tissue and organism levels. In particular the analysis suggests an important role of transcription factors, which appear to have a conserved influence across cell-types and species. Validation via an independent dataset confirms the findings and network analyses suggest explanations, how TGF-β perturbation could lead to the observed effects.

Chapter 5 provides another example for pattern mining from perturbation data. microRNAs (miRNA) and relevant target genes are investigated as potential biomarkers in epilepsy (Kretschmann and Abnaof et al., 2015a). Epileptic seizures are induced in rat and mouse models and longitudinal microarray and high-throughput qPCR data are generated from blood and hippocapus samples. MicroRNAs naturally have

a low abundance in the blood, and thus often fall below the detection limit of the qPCR technique. This effectively results in right censored measurements, which complicate follow-up statistical analysis. In order to address this issue I suggest a specific work-flow, which is also tested in simulations. The devised procedure enabled the identification of novel pathogenic relevant miRNAs. Relevant target mRNAs could be identified and indirect characterization of epilepsy types is achieved through comparisons of miRNA activities in the animal models at different time points. A number of disease relevant biological processes and pathways were significantly associated to gene target sets of deregulated miRNAs.

Pattern mining in a different context is demonstrated in chapter 6. a novel *biological effects similarity measure (BES)* for chemical compounds is proposed. In contrast to existing similarity measures the BES integrates compound-target affinities and captures compound perturbation induced effects by integrating large set of diverse features of target proteins. The BES was validated with gene expression data and then used for pattern discovery in a dataset of over 4,500 chemicals. BES based consensus clustering detected several separable and statistically stable clusters, of which targets could be related to specific pathways, biological processes, protein domains and sequence motifs. The identified clusters could not be found in a traditional chemical structure based clustering using fingerprints and the Tanimoto-Jaccard similarity. Targets of compounds falling into one cluster suggested several new compound-target combinations, which could in several cases be confirmed by independent data. BES based clustering may thus help to explore compound libraries and identify interesting novel compound-target indications for follow-up experiments.

Final discussion, concluding remarks and prospectives for research are presented in chapter 7.

CHAPTER 2

HIGH-THROUGHPUT TECHNOLOGY IN BIOLOGY AND MEDICINE

The development of novel and advancing high-throughput technologies in the course of human genome research enables the comprehensive analysis of biological information. These technologies have been increasingly used for the profiling of molecular biology entities and chemical substances in the research fields of biology, medicine and pharmacology.

Examples of high-throughput technologies are mass spectrometry based on measurement of proteom and metabolome, transcriptome assessment by DNA microarrays or by next generation sequencing. Importantly, these technologies enable to get a comprehensive picture of the biological processes that are executed in a living cell by quantifying biological activities of thousands of molecules. For example, from the measurements of gene expressions and the resulting proteome, insights can be derived about metabolic processes, cellular differentiation and intra- and extracellular signaling pathways.

Different types of high-throughput technologies produce data of different types and format. Therefore, a thorough understanding of theses technologies is essential for the active role of the analysis of the produced data. This incorporates the correct data acquisition, handling and integration. Three categories of high-throughput technologies are addressed in the context of this thesis. These are the *microarray technology*, *high-throughput quantitative polymerase chain reaction (qPCR) assays*, and *measuring compound's biological activities*. Data sets involved in the subsequent chapters of this work are majorly gained using these methods. Normalization techniques for these data types are briefly discussed.

2.1 The Microarray Technology

The genetic information of an organisms is decoded in the *Deoxyribonucleicacid (DNA)*. Genes are short segments of the DNA. The genetic information in these segments can be read via transcription process and transformed into RNA molecules. Certain types of these RNA molecules, the *messenger RNAs (mRNAs)*, can be synthesized via translation process into proteins (Chang, 1983; Schena et al., 1995).

The type of RNAs molecules and proteins expressed in the cell reflect the activities of the corresponding genes in it. Thus, the magnitude of gene activity can be indirectly quantified by measuring the abundance of the proteins and RNA molecules in the cell. Microarray is a technology by which such types of indirect regulation of gene expressions can be investigated (Heller, 2002; Ehrenreich, 2006; Dufva, 2009). For instance, differences in the amounts of RNA between cells/organisms which have been exposed to various controlled perturbation conditions can be assessed. This way, genes showing different expressions between different cell conditions can be specified. Such genes which are assumed to cause the cell response due to perturbation are usually identified via statistical methods for differential expression analyses of microarray data (section 3.2).

A microarray is a chip, usually made of glass or silicon, that contains thousands (25,000-60,000) of microscopical small spots. There are many types of microarrays; e.g. DNA, protein & tissue microarrays, or antibody microarrays. In DNA microarrays multiple copies of a uniquely specific DNA segment (e.g. parts of a gene) of interest are placed in each one of the spots on the microarray chip. The DNA segments are olignucleotides or cDNA segments of moderate length (500-500bp). The DNA is amplified and purified by *polymerase chain reaction (PCR)* technology (section 2.3) before applying it to the chip. The biological material is printed on the chip by a robot arm that has several hundreds needles thus allowing for application of large number of molecules per printing step (Goldmann and Gonzalez, 2000). After doing this the points in the chip are ready to act as docking spots e.g. for mRNA molecules. The printed (loaded) chips can then be used to measure abundance of biological material in samples applied to them (Figure 2.1).

Thousands of unique mRNA molecules extracted from the cell are labeled with certain fluorescent solution and applied to a microarray chip. Hydrogen bonds are built, under adequate application of heat, between the *Amine* and *carbonyl* groups on the complementary base-pairs of mRNA and the DAN material on the chip (Adenine (A) & Thymine (T) and Cytosine (C) & Guanine (G) are complementary to each other).

Guided by the property of complementary base-paring each mRNA molecule docks (binds) onto a certain point in the microarray chip, namely onto the spot where the gene is located to which the mRNA molecule corresponds. This docking process is termed hybridization. The rest of the biological material that did not bind is washed away after hybridization step (Maskos and Southern, 1992; Guo et al., 1994; Lockhart et al., 1996).

Sometimes binding takes place between some strand of mRNA and the DNA in the spot although these are only partially complementary to each other. This process called cross-hybridization negatively affects the accuracy of the binding by producing unspecific spots (none perfect-match). Cross-hybridization can be avoided by carefully adjusting the temperature at which the hybridization is done. Higher temperature reduces the chance for cross-hybridization. However, very high temperature might block the hybridization even in the perfect-match spots (Zhang et al., 2005; Pozhitkov et al., 2006; Deng et al., 2008).

The intensity of the light emissions of the fluorescent solution and thus the amount of hybridized mRNA for each point of the microarray is then measured with a scanner. This is done by exposing the chip to laser light which activates the fluorescence mixture in the spots. The spots glow with colored light emissions that has different intensities depending on the amount of biological material in each spot. These light signals are then scanned by a designated scanner that gives a pixel-image for each chip. The intensities of color pixels that correspond to the spots on the chip are filtered and transformed by an image-processing software (e.g. GenePix® Pro) to digital intensities. The measured intensities are calibrated to produce gene expression which reflect the activity of corresponding genes. Figure 2.1 gives an over view of manufacturing and utilization steps of microarray chips.

There are a number of different main types of microarrays. One type is the patented and commercially available the *oligonucleotide-array* like the *GeneChip®*, *Affymetrix®* and *Agilent®* microarrays. The other one is a freely accessible technology such as the *cDNA-Array* which is developed at Stanford University. The cDNA-Array uses cDNA guidance for spot binding and is usually utilized for double probe hybridization, thus, producing two-color (tow-channels) arrays (containing r-color fluorescent probes and g-color fluorescent probes). The oligonucleotide-array uses synthetic oligonucleotides as docking mark for the probes. This type is usually utilized to produce one-color (one-channel) arrays where each probe is hybridized only once in the chip.

In order to ensure that microarray data can be easily interpreted and that results derived from its analysis can be independently verified certain standards and guidelines have been formulated. The standards are summarized in the so-called “minimum information about microarray experiment” (MIAME) guide (Brazma et al., 2001).

These guidlines involve the use of designated software tools, standardized data annotation and unified file-exchange formats.

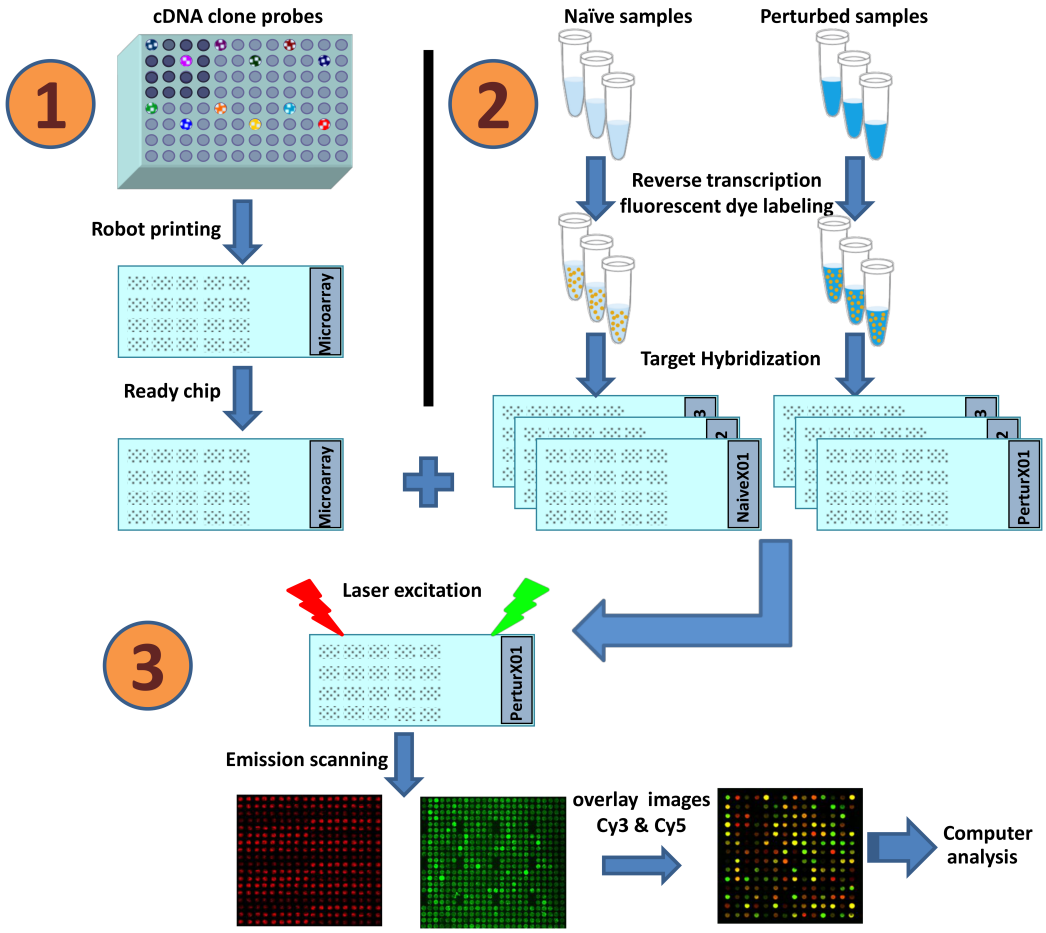

Figure 2.1: cDNA Microarray Technology. The DNA segments are amplified and purified by PCR technology and then printed onto microaray's chip. Target cDNA or oligonucleotides material (e.g. mRNA of two cell types; the naive control type and the stimulated 'perturbed' type) are marked with fluorescent marker and hybridized under heat to the printed chips. Chips are exposed to laser and signal emissions are scanned to images. The images are transformed into numerical intensities by an image-processing software (e.g. GenePix® Pro). Intensity values are calibrated to gene expression values by normalization and pre-processing method and analyzed by further statistical methods.

The expression values are usually assessed by the *log-ratio* of the red-channel and green-channel probe intensities in the two-channels array (subsection 3.2.1). In the one-channel arrays the calibrated absolute expression values are compared between the different perturbation sample sets. Therefore, it is very important to know the type of array before applying data pre-processing and further analysis steps.

Biological molecules (e.g. mRNAs) are not exclusively hybridized to their corresponding features (e.g. genes) from which they were transcribed. They usually refer to parts of these features. This leads to random error in the measurements of high-throughput microarray data. Statistical methods are therefore required for the evaluation of this data. High-throughput data have three major problems: (i) sample size is typically small, (ii) samples are of high dimension (iii) explanatory variables are usually highly correlated.

The small sample size problem is due to the fact that the technology is still expensive and biological material required for experiments are usually either limited or costly. The high dimensionality of the data is due to the typically very high number of features measured in each sample. Because of these problems, the application of classical statistics approaches for the analysis of this type of data is often ineffective. Therefore, new methods have been proposed that allow for accurate analysis and deeper insights into the predictive structure of high-throughput data (Rao et al., 2008).

2.2 Normalization of Microarray Data

The signal intensities in oligonucleotide chips and other microarray types contain variations that are not solely due to biological influences under investigation. The sources of these variations are many. Examples of these sources are; fluctuation in the amount of RNA in the biopsy, the efficiency of the RNA extraction and reverse transcription procedure and fluorescent detection. But also technical artifacts such as dye bias, different incorporation of dyes and stray signals resulting from unspecific binding and cross-hybridization and different scanning parameters can cause such undesired variations. However, the intensities should act as unbiased estimator of feature's abundance. Therefore, removing these variations is necessary before the values can be used in further analyses.

The undesired variation in signal intensities are of two types; stochastic effects and systematic effects. The former, are difficult to estimate and thus remain as noise and model error, the later are similar effect of many measurements which can be estimated from the data and therefore can be removed by calibration.

Calibration or normalization is the transformation by reducing intensity-dependent variations and technical artifacts. Part of the systematic effects influence all the arrays in a similar way, another part differs within the array. Therefore, two different types of normalization are required to eliminate the systematic variations (in two-channel cDNA arrays): (i) within-array normalization to compensate for the systematic variations in the individual arrays, (ii) between-array normalization to calibrate the different arrays to each other. In some microarray types a background correction of the intensity values is usually performed. Background correction can be seen as part of the within-array normalization procedure. The normalization is done under the assumptions that expressions have the same distribution and only small proportion of the total features are differentially expressed (Yang et al., 2001; Quackenbush, 2002; Irizarry et al., 2003; Wilson et al., 2003; Smyth and Speed, 2003; Chen et al., 2003; Do and Choi, 2006; Rao et al., 2008; Hua et al., 2008).

The aspects of within-array normalization, between array normalization and background correction are explained briefly in the following and the used methods and procedures are discussed.

2.2.1 Background Correction

Microarray chip scans provides foreground and background intensities for each spot. The background intensities arise the unspecific binding to array chips or caused by the optical noise during chip scanning. These background intensities vary considerably in the different parts of the chip. Therefore, they are locally calculated for each individual spot, usually using the median value. For each spot the background intensity is computed as the median intensity of the intensities in the area surrounding the spot. This is done because different areas of the chip have different background intensity levels. There are many technical reasons that different areas in the chips share similar spot intensity levels which are different from each other. Systematic differences like this might be caused by the 'Print-Tip' groups in the chip (by the move of robot arm printing head in each printing step) or due to marker-specific effects.

The foreground intensities are the actual spot raw intensities. Thus the foreground intensities are composed of specific and unspecific signals. It is important to perform background-correction for signal intensities before normalizing the data.

Various methods for background correction have been proposed in the literature (Edwards, 2003; Silver et al., 2009; Schützenmeister and Piepho, 2010). In a simple approach the specific true signal intensities are computed by subtracting the background intensities from the foreground intensities. This is done assuming that the

unspecific intensities of a certain spot can be derived from the spot surrounding-area intensities (in this case by the median value). However, this method often produces negative signal intensity values or values that are close to zero. The negative intensities are difficult to interpret and the methods for further analysis usually do not allow for negative expression values.

Therefore, other methods beside this subtraction method have been proposed (Ritchie et al., 2007). For this thesis the 'Normexp' approach is of relevance. The 'Normexp' method is based on mixture model of normal and exponential distributions. The background noise is modeled as normally distributed and the signal is assumed to follow an exponential distribution (Ritchie et al., 2007; Silver et al., 2009). The method produces only positive signal intensity values. If these intensity values happen to be close to zero then an 'off-set' values of 50 can be added to all spot intensities. We used the 'Normexp' method for background correction of our two-channel microarray data in chapter 5.

2.2.2 Within-Array Normalization

Systematic variation within an array is partially due to the intensity dependent variation. This cause the signal differences (e.g. between two conditions as in 'Log Fold Change', or between red and green signal intensities in two-color arrays as in 'LogRatio', for details see subsection 3.2.1) for high intensities to be higher than signal differences for low intensities although the difference magnitude might be the same (Yoon et al., 2004).

To remove within array systematic variation of this sort, methods such as the LOWESS normalization method is used. This normalization is based on the LOWESS (locally weighted scatterplot smoothing) regression method (Workman et al., 2002; Berger et al., 2004). This is a non-parametric regression where a smoothed curve is estimated for the M-values in relationship to the A-value (see subsection 3.2.1). The link function type and the number of data points involved in fitting the smoothing curve can be flexibly specified. Furthermore, a separate LOWESS curve might be fitted for each 'Print-Tip' spot group in the chip in order to balance the systematic differences in them. However, the smoothing of the curves might not be optimal if the number of the spots in the print-tip group is low.

The resulting global regression curve is subtracted from the M-values so that the normalized values are distributed closer to the null horizontal line in an MA-plot. This true under the assumption that the signal intensities and hence the differentially expressed features are symmetrically distributed around the line $M = 0$. However, this assumption might not always be reasonable (Yang and Thome, 2003).

2.2.3 Between-Array Normalization

The within-array normalization removes systematic variations within the individual arrays. However, differences between the chips due to various scales still exist. Such differences can be reflected by the differences in variance values of the different chips (typically visualized by Box-plots of background corrected and within-array normalized spot intensities in each chip). Moreover, the variability of the intensities depends on the mean intensity values. Therefore, a further between-array normalization step, involving the scale transformation of the array, is required. Such normalization is done, for instance, by adjusting the variances of the chips so that they are comparable.

A number of scale transformation normalization methods have been proposed (Quackenbush, 2002; Smyth and Speed, 2003; Yang and Thome, 2003; Wilson et al., 2003). A frequently used normalization method is the quantile normalization (Bolstad et al., 2003; Boes and Neuhäuser, 2005; Rao et al., 2008). The quantile normalization method adjusts the quantiles of intensity values of the different arrays to each other so that the expression values of the aligned arrays are comparable. This method is strict in assimilating the arrays to each other. Therefore, it is suitable for oligonucleotide arrays which naturally contain tens of thousands of features. For arrays with few features it is rather not suitable.

The variance stabilizing normalization (VSN) method and another related method, the log-transformation normalization, are frequently utilized (Huber et al., 2002). Both methods involve the estimation of calibration parameters; the shift and scale parameters. The VSN normalization method is based on additive-multiplicative error model (Rocke and Durbin, 2001; Huber et al., 2002, 2003). Here, intensities in chip i and probe k are fitted into a two terms model $y_{ik}^{caliberate} = \alpha_{ik} + b_{ik} \times y_{ik}$. According to Huber et al. (2002) these terms are; an offset term α_{ik} which is a chip-specific constant term representing background signal; and a signal term which is a multiplicative term of the signal value y_{ik} and a scaling factor b_{ik}. The shift parameter is estimated as: $\alpha_{ik} = \alpha_i + \varepsilon_{ik}$, and the scaling factor as: $b_{ik} = b_i \times b_k \times \varrho_{ik}^{\eta}$, where b_i is a chip-specific normalization factor, b_k affinity factor per probeset and ϱ_{ik}^{η} is a multiplicative error term.

These normalization methods assume that only a mall proportion of the features in the chips are truly differentially expressed. These normalization methods focus on removing variance-mean dependency in signal intensities. However, other sources of variations such as variations in tightness of gene transcriptional control between the experimental conditions are ignored.

2.3 High-Throughput qPCR Technology

The polymerase chain reaction (PCR) is an enzyme-dependent processes for the reproduction of certain (gene) sequences in the DNA. The principles of this natural process that take place during replication in the cell can be imitated for *in-vitro* amplification of gene sequences. Done this way, PCR can be used for measuring DNA molecules abundance in the cell by exponentially copying it. The principles of this method, by which even very small quantities of DNA can be proliferated and measured in short time, have been invented by the chemist Kary Mullis in 1983. Few years later PCR was one of the most utilized technologies in molecular biology (Higuchi et al., 1992, 1993; Pabinger et al., 2014).

Single-stranded or double-stranded DNA chains with known sequences (of length 100-600 bases) can be replicated using PCR (VanGuilder et al., 2008). To carry out amplification, primers, individual nucleotide-triphosphate molecules as well as a heat resistant DNA ploymerase and a thermal-cycler are required. The primers are complementary to the sequences in the DNA so that they can bind to certain sites of sequence strand. PCR involves heating the mixture material (up to 94 °C). Therefore, a polymerase that is not easily destroyed with heat is essential. A frequently used heat resistant DNA ploymerase is the Tag-ploymerase which is isolated from the Thermus-Aquaticus bacteria.

The DNA polymerase extends the primers so that copies of the starting sequences are generated. Through heat modification of the biological material the double stranded chains can be denatured (separated). Now new copies of the starting sequences and the copied sequences are generated using the same primers in the second PCR cycles. Thus, the DNA material is replicated exponentially to the number of amplification cycle. In cases of single-stranded RNA molecules (mRNA, miRNA) reverse transcription is used to transform the molecules into complementary DNA (cDNA) before running PCR. Figure 2.2 gives an over view of one cycle in DNA PCR process.

In order to measure the abundance of DNA molecules so many PCR runs are performed until the DNA quantity accumulated (amplicon) is more than a certain threshold. The cycle number at which this occurs is called the *Cycle threshold (Ct)* value. Thus, the data acquired in PCR technology are the Ct values.

In the end-point-PCR the detection and quantification of amplified sequence is achieved at the end of the procedure. In real-time-PCR (RT-PCR) the accumulation of amplicon is detected and quantified during the reaction progression in real time.

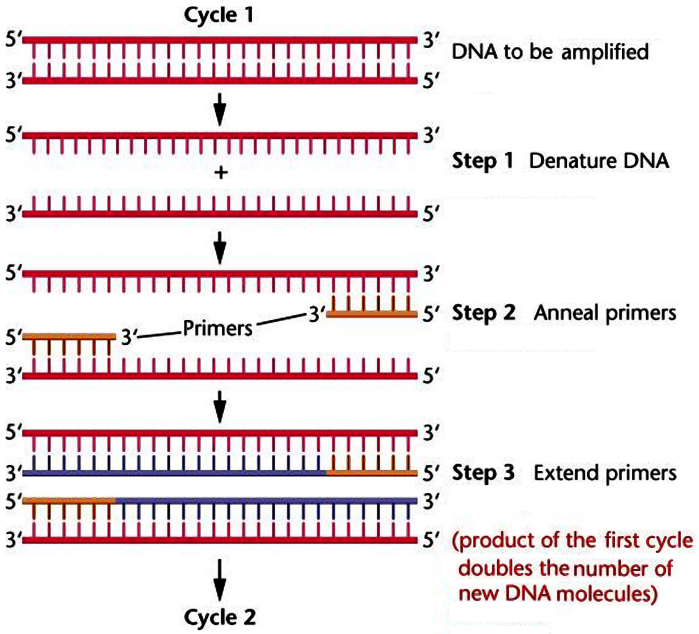

Figure 2.2: Ploymerase Chain Reaction (PCR). Short DNA sequences (100-600 bp) generated (from e.g. mRNA) by reverse transcription is denatured at $\sim$ 94 °C and the strands separate. The sample is cooled to below 60 °C and two sequence specific primers (orange-colored) are annealed to the separated strands, each one to its complementary site of the respective strand. By heating the samples ($\sim$ 70 °C) Tag polymerase extends the two single strands from the primers into double stranded DNA (blue-colored) so that the DNA sample is now duplicated. The process is repeated in the subsequent cycles[1].

The amount of the amplified product in both PCR types is measured by a fluorescent dye that produce detectable signals. Fluorescent dyes that are widely used in PCR technology are the SYBER® Green (Ponchel et al., 2003) and the fluorogeneic probes like the Tagman probes (McGoldrick et al., 1999).

The fluorescence signals have starting phase at early cycles, exponential phase in the middle and saturated phase at the end of PCR reactions. The amount of the accumulated amplified product remains low at early cycles, so that the fluorescence signal remains as background noise. This amount is duplicated in each PCR cycle so that the amount of amplified product accumulates enough to generate a detectable signal. This usually happens at the exponential phase of the signal curve. The cycle number at which the signal cross a designated threshold is called the

[1]Adapted from source: (accessed: 24th July 2015) `http://www.microbiologybook.org/pcr/pcr-home.htm`

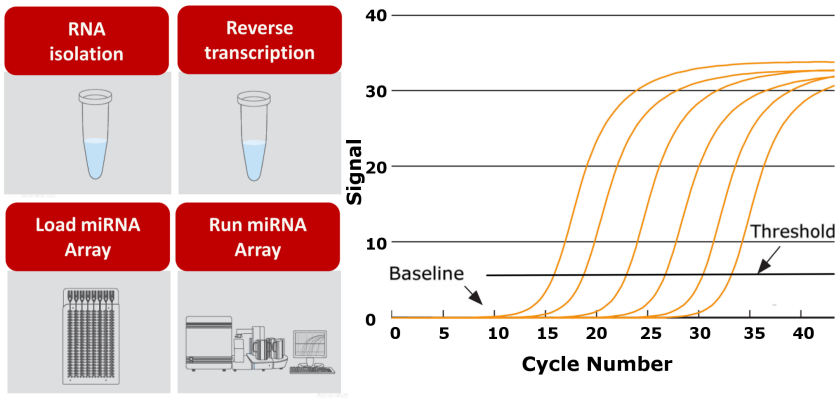

Figure 2.3: Quantitative Ploymerase Chain Reaction (qPCR) procedure. RNA are isolated from samples and transformed to cDNA by reverse transcription. In high-throughput PCR all DNA molecules are loaded in plate before running PCR. After several amplification cycles (most of the thermal-cycler machines usually allow upto 40 cycles; cycle number plotted in the X-axis) Ct values are detected when fluorescent signals curve (fluorescence signal are plotted in the Y-axis) cross the horizontal threshold line (preferably in the exponential part of the signal curve). The plot shows different amplification curves which are generated by different-fold diluted starting materials. The most diluted samples produce the curves towards the right-hand side and their signals (Ct values) are detected at higher cycle numbers.

threshold cycle Ct. Most of PCR machines typically allow for up to 40 amplification cycles. Features which are not detected until maximum cycle number are labeled as "undetermined" and typically assigned Ct value of 40. This practice raise problems for the normalization and analysis of this type of data (see chapter 5).

In high-throughput qPCR high number of different DNA sequences are amplified in parallel. After RNA isolation and reverse transcription of the DNA sequences the purified samples are loaded in chips that typically contains 384 wells to accommodate the samples (Figure 2.3 left). Instead of individual sequences the amplification is done for all samples in the chip (Higuchi et al., 1992; Schmittgen and Livak, 2008). Figure 2.3 right shows PCR fluorescence signal curves of 6 amplified samples with different dilutions.

The starting material (template) at the beginning of amplification reactions affect the resulting Ct values at the end of the cycles. Therefore, dilution of starting material is utilized as strategy to control the resulting Ct values. Variation errors caused in samples extraction and PCR procedures make the raw Ct values less optimal for the

quantification of amplified product. Therefore, these values need to be calibrated and relative PCR expression values are required.

2.3.1 Normalization & Analysis of qPCR Array Data

The Ct values are determined by the amount of the biological material at the beginning of the amplification reaction. In addition, there are other sources of systematic errors and noise. For example, the efficiency of RNA extraction and the reverse transcription step before PCR process adds to the experimental complexity and variation errors in the acquired data. In order to obtain useful quantification values that are also suitable to compare expression levels of features across different conditions, normalizing the raw Ct values is required.

A number of methods for normalization of RT-PCR data have been proposed (Vandesompele et al., 2002; Huggett et al., 2005; Peltier and Latham, 2008). A frequently used normalization and analysis method for relative gene expression RT-PCR data is the $\triangle\triangle$ Ct method suggested by Livak and Schmittgen (2001). This method assumes doubling of target DNA in each cycle and requires one or more endogenous reference genes (Schmittgen and Zakrajsek, 2000; Hamalainen et al., 2001; Eisenberg and Levanon, 2003; Pfaffl et al., 2004; Silver et al., 2006; de Jonge et al., 2007). The reference genes should fulfill certain criteria. For instance, they should be expressed (i.e. $Ct \leq 40$), however, they should also be non-regulating towards the experimental setting so that their expression is stable and none-differential across the perturbing conditions. The Ct values of the reference gene are subtracted from the Ct values of the target gene producing $\triangle$ Ct values. Expression ratios of target gene between control and perturbation are then computed by the differences of their $\triangle$ Ct:

$$\begin{aligned} \triangle\ Ct &= Ct_{target} - Ct_{reference} \\ \triangle\triangle\ Ct &= \triangle\ Ct_{treatment} - \triangle\ Ct_{control} \\ FC &= 2^{-\triangle\triangle Ct} \end{aligned} \tag{2.1}$$

This normalization reduce the variations in in target gene expression because the sources of errors in expression values affect the reference and target genes in the same manner.

High-throughput qPCR data are often normalized and analyzed by typical microarray normalization and analysis methods (Schmittgen et al., 2008). However, methods of large-scale mRNA microarray data are not appropriate for high-throughput-qPCR which typically comprise few hundreds of features (Pabinger et al., 2014). In case

of microRNA qPCR data there is not even assessment of the performance of these methods. There are many other issues, for instance, the majority of miRNA are either expressed or not expressed at very low levels with many "undetermined" Ct values. That can bias the normalization and the analysis (Pradervand et al., 2009; Git et al., 2010). Therefore, a number of normalization methods have been suggested for high throughput qPCR data, these include reference gene (house-keeping gene) normalization (Livak and Schmittgen, 2001; Mar et al., 2009) and geometirc mean, rank-invariant, scale-rank and quantile normalization methods (Dvinge and Bertone, 2009; Mestdagh et al., 2009).

A number of tests for differential-expression analyses approaches have been proposed for PCR data. These range from non-parametric tests such as Wilcoxon-Mann-Whitney test to usual t-test and linear-models aided t-test. However, most of these methods which are traditionally used for gene expression data from microarray adapt assumptions that may not be appropriate for high-throughput PCR data. Performance of some of these methods have been investigated based on simulated and real data sets in chapter 5. Furthermore, we showed that aiding one of the existing methods with a variance estimation technique for censored data improved its performance.

2.4 Assessing Biological Activity of Chemical Compounds

Chemical substances can be classified descriptively according to their qualitative effects on the biological targets into classes such as inhibitors, suppressors or activators. However, such classes are highly overlapping on targets levels and the classification does not explain how the activity takes place and lacks quantitative aspects. Thus, a more refined and quantitative measure of compound effect is needed. Drug effects can be derived from their activity mechanism at the molecular and cellular levels.

The quantification of compound effect is decided by the dose-response relationship (Crump et al., 1976; Altshuler, 1981; Sühnel, 1998; Shoichet, 2006; Gaydos et al., 2006). Compound combines with a receptor at a rate that is dependent on the concentration of the ligand and the concentration of the receptor. This relationship can be described in its kinetic and dynamic equilibrium by the law of mass action (Waage and Gulberg, 1986). Using this concept, performance of a bioactive substance on targets can be characterized by the required concentration of the substance to invoke certain level of effect on a target.

Biological activity of a compound with respect to a biological target can be assessed by the dose-response principle. The dose-response principle describe the change in effect in a biological target that is caused by exposing the target to different dose

concentrations of the drug. Dose-response relationship is experimentally determined by incrementally increasing the concentration of the chemical substance to reach maximum effect in a solution of individual isolated tissues, cells or membranes. The maximum effect is reached if all solution molecules have responded to the chemical substance. So now instead of the administered dose, the drug concentration in the solution can be used for the quantification of the compound activity.

A mathematical relationship can be set between concentration and effect e.g. by fitting a curve (Altshuler, 1981). Most drugs exhibit non-linear relationship between concentration and response. Plotting the logarithm of the dose concentrations against the percent levels of the effect ideally gives a sigmoid curve. The position of the curve along the x-axis reflect the potency of the drug (concentration that is necessary for e.g. half-maximal effect), the peak of the curve stands for drug efficacy (intrinsic Activity; extent of the maximum effect), and the slope of the curve reflects how wide is the concentration rage between minimal and maximal effects. Using the dose-response relationship comparable measurements for compound biological activity can be defined. For example the *EC50* measures the dose concentration (e.g. in mole) of an agonistic compound required to induce 50% of the maximum possible effect, and the *IC50* measures the concentration of antagonist drug needed to inhibit 50% of a given biological function. EC50 and IC50 are opposite to each other, i.e. lower IC50 and higher EC50 values mean higher biological activity of the compound.

Another measure is the *inhibition constant (Ki)* which assesses the rate of competitive inhibitory binding affinity of antagonistic compound. The Ki measures the concentration of the inhibitor that causes 50% inhibition of enzymatic reaction. Lower Ki value means higher binding affinity of the drug. While EC50 and IC50 depend on how the experiments are done, Ki is an absolute inhibition constant of a compound and does not depends on the experiment. IC50 values for inhibitor of enzyme activity and ligand binding can be converted into Ki values by the Cheng-Prusoff equation (Cheng and Prusoff, 1973; Munson and Rodbard, 1988).

The advancing technology in the computational design and automated screening help to experimentally detect certain perturbation effects. Such perturbations can be caused, among others, by chemical agents. Special assays can be used to test hundreds of thousands of chemical compounds (usually set in libraries of similar compounds) for their biological activities towards numerous biological perturbed targets in short time. This process called *High throughput screening (HTS)* is done using small and efficient *in-vitro* test systems and *microtiter plates*.

The assay type used depends on the target and whether it is cell based (functional assays) or cell free non functional functional assays. Biological activity detection techniques used in most of the assays are based on radioactivity, or light emissions in form of fluorescence/luminescence or based on other target-supported methods

(Skehan et al., 1990; Johnson et al., 2007; Martinez Molina et al., 2013). An assay type that is frequently used to measure binding affinity is the class of binding assays. These are microplates with radioligand markers that are susceptible to biological targets. The advantages of this assay type are that they are quick, and through their multiple-well setting they can process many samples in parallel (Rishton, 1997; Macarron, 2006).

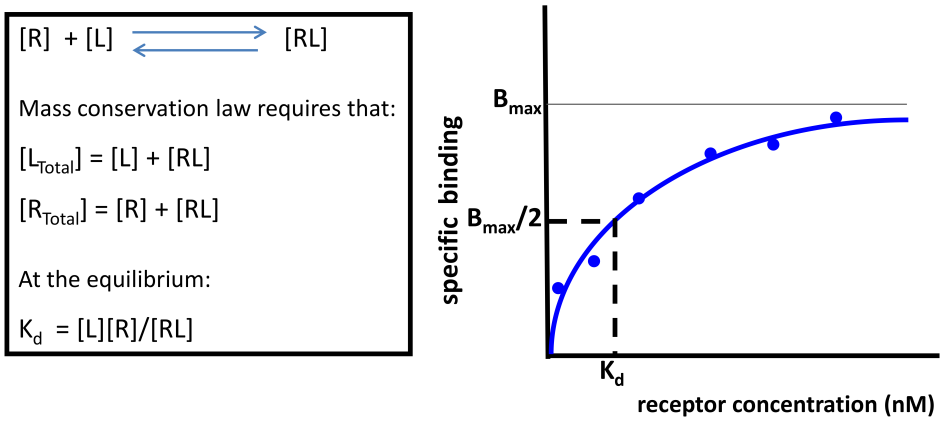

Figure 2.4: Saturation binding curve for radioligand assays.In saturation radioligand binding assay a labeled ligand [L] binds to a receptor [R] (x-axis). In competitive binding assays an unlabeled ligand [L] 'the competitor' binds to a receptor [R] in competition with a labeled ligand (x-axis). The equilibrium of the dissociation constant (Kd) which is the opposite of the inhibition constant (Ki) is given by the equation in the left-hand side box that describe a saturation experiment. Kd is figured out by non linear curve fit, so that Kd is the free ligand/receptor concentration that corresponds to 50% ($B_{max}/2$) of the maximum concentration of binding (B_{max}). If Kd for the labeled ligand is known then the inhibition constant (Ki) can also be computed in competitive binding assays (Hulme and Trevethick, 2010).

The inhibition constant (Ki) and the dissociation constant of the inhibitor (Kd) are measured in saturation or competitive binding inhibition (e.g. generic CYP) assays (see Figure 2.4). by varying the inhibitor concentrations e.g. for different spots in a microtiter plate that contains constant enzyme concentration. Fluorescent signals measure the inhibition kinetic in the enzyme activity (Bisswanger, 2002; De Jong et al., 2005; Findlay and Dillard, 2007; Hulme and Trevethick, 2010).

CHAPTER 3

PATTERN MINING & KNOWLEDGE DISCOVERY METHODS IN HIGH-THROUGHPUT DATA

3.1 Introduction

The vast amounts of data generated by high-throughput technologies need to be adequately and thoroughly analyzed. This requires multivariate statistical methods for the extraction of meaningful patterns.

Particularly relevant to mining of high-throughput data are statistical tests for the identification of relevant features (e.g. to perturbation experiment). Clustering and classification techniques are essential for determining functionally related groups of samples and features.

In the following sections the major methods used in the subsequent chapters are explained comprehensively including brief literature reviews. The categories of methods explained here include statistical tests for differential expression analysis, clustering approaches and enrichment analysis methods for gene sets.

3.2 Differential Expression Analysis Methods

One of the central tasks in biological high-throughput data analysis is the detection of relevant features. For example, in gene expression data analysis one objective is to find genes that exhibit differential expression between two conditions. The assumption behind differential expression analysis is that the true differences between perturbation conditions can be attributed to a small number of deregulated features.

In the following, differential expression analysis methods for stationary and time-course expression data are briefly discussed and the methods used in the subsequent chapters are explained in more detail. It is assumed that the data have passed the quality control check and are properly normalized. The null hypothesis for each feature under investigation is that there is no difference in feature's mean expression between the conditions. Each condition is usually represented by multiple sample replicates. Usually it is important to distinguish technical replicates from biological replicates.

3.2.1 Assessing Differential Expression by Log-Ratio

The objective of biological expression data analysis is to detect differential expressed features in the data. In this context the difference or relative expression values of the feature (e.g. between different cell types) is of interest rather than merely the expression values themselves. Because the difference of expressions is not scale-independent the ratio between the average expressions of two conditions is usually used instead (Huang et al., 2004). However, the relative values lack the symmetry which make them difficult to interpret. This problem is circumvented by taking the logarithm of the ratios (Log-Ratio; to the base 2) to express differential expression (Quackenbush, 2001). For instance, in a two-color arrays the Log-Ratio (M-value) is defined as:

$$M(g_i) = \log 2 \left(\frac{R_i}{G_i} \right) = \log 2(R_i) - \log 2(G_i) \tag{3.1}$$

where, R_i is expression of the red-marked spot of the feature probe g_i, and G_i the the other spot of the same probe which is green marked. The logarithm to the base of 2 make the interpretation of the Log-Ratio values easier. Instead of two-color arrays nowadays microarray experiments are often done in one-channel arrays with replicated samples in each condition. In that case the averages of ratios and their logarithm are used:

$$\log_2(FC) = \log_2(\bar{y}) - \log_2(\bar{x}) = \log 2 \left(\frac{\bar{y}}{\bar{x}} \right) \tag{3.2}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_i$ are the average expression values for the feature in the samples of the two conditions (e.g. control condition x and treatment condition y).

Another value which is often used in the context of within-array normalization is the average of expression logarithms denoted A-value and defined as:

$$A(g_i) = \log_2(\sqrt{R_i \times G_i}) = \frac{1}{2}(\log 2(R_i) + \log 2(G_i)) \tag{3.3}$$

The Log-Ratio (M-Value) and logFC are not good measures to assess differential expression. Averages upon which the logFC is computed are highly influenced by outliers. It has been shown that these ratios are not stable at low intensity values, highly sensitive to rounding errors and do not allow for negative expression values. Furthermore, their range is unbounded so that they can produce extremely high values that affects their variation thus making them unreliable measures (Ultsch, 2003). Other measures are proposed to avoid these limitations. An example is the relative-difference of R & G defined as: $RelDiff(g_i) = 2\frac{(R_i - G_i)}{R_i + G_i}$. However, this measure does not eliminate all the above limitations. In addition, both the Log-Ratio and the RelDiff measures are not suitable in the case that different features have different variations. Moreover, these measure are useless in case the feature is off (below detection limit) in one condition (Newton et al., 2001; Yang et al., 2002; Ultsch, 2003).

3.2.2 Assessing Differential Expression by Statistical Tests

Simply ranking genes according to their logFC does not take into account the variance of the fold changes. Therefore, a number of sophisticated statistical tests have been suggested for the identification of differentially expressed features. These tests have been enhanced in different ways so that they can handle particular study designs and data specifications.

Hypotheses can be formulated for each test, in a null-hypothesis (H_0) and an alternative (e.g. here two-sided) hypothesis (H_1), as follows:

$$\begin{aligned} H_0 &: \text{feature}_i \textbf{ is not} \text{ differentially expressed} \; (M_i = 0) \\ H_1 &: \text{feature}_i \textbf{ is} \text{ differentially expressed} \qquad (M_i \neq 0) \end{aligned} \tag{3.4}$$

The test statistics value is then used to decide whether the a null-hypothesis (H_0) can be rejected and thus the subject feature can be considered differentially expressed, i.e. its log-ratio between the conditions M_i is none-zero, or not.

The test is performed individually for each feature so that the variance is estimated in the standard form from the observations of that feature only. Variances estimated this way get affected by small sample sizes and outlying observations. In addition these local estimates ignore variance heterogeneity among all features. All that affects test statistics stability and comparability for the individual features and accordingly its detection power.

The most commonly used method to determine differential expression of genes in high-throughput data is the Student's t-test method. In replicated two-conditions study designs a standard t-test can be done for each single feature and t-statistics can be used to determine which feature is differentially expressed. This test is valid under the assumption that the log-ratios between the conditions follow a Student's t-distribution.

Many modifications have been introduced to the t-test. For instance, a global variance that is pooled across all the features can be estimated in order to mitigate variance heterogeneity effect. However, this might not completely eliminate the effect. Baldi and Long (2001a) suggested regularizing the variance by combining feature-specific and a global average variance estimates by a weighted average (Baldi and Long, 2001a,b). The *significance analysis of microarrays (SAM)* tool uses a modified t-test (called S-test) where a small constant S_0 is added to the local variances in order to minimize the coefficient variations of the test statistics (Tusher et al., 2001; Chu et al., 2011). Various other t-test modifications have been suggested (Pan, 2002; Dudoit et al., 2002; Yu et al., 2011).

Other statistics have also been suggested for differential expression analysis, the *B-statistic* is one example. The B-statistic is the logarithm of posterior odds of differential expression i.e. the logarithm of the ratio of two probabilities; the probability that the feature is differentially expressed and the probability that it is not (Lonnstedt and Speed, 2002).

The B-statistic and most of the suggested modifications for the t-test statistic have serious limitations. For instance, in the adjustment of the denominator of the t-test the methods do not distinguish between differential and none-differential feature's

classes. Moreover, the estimated prior variance is treated equally to the local or global variances. The major problem however, is that most of these methods are limited to two-condition (treatment versus control) designs.

3.2.3 Linear models for Microarray Data (limma)

The experimental design of high-throughput expression studies is usually complicated and involves more than barely a comparison of two perturbed conditions. Experiments with multiple perturbations, different time-points responses with biological & technical replications and associated covariates are frequently encountered. Planning the analysis of the data of such experiments based on ratios is not sufficient as these ratios do not consider data from other conditions which are not being compared.

A method that avoids the drawbacks of the above methods and has the advantages of handling very complicated experimental designs (e.g. direct designs, factorial designs and time course experiments) is the *linear models for microarray data (limma)* method (Smyth, 2004a, 2005). The limma method allows for tow-conditions as well as experiment designs with multiple conditions through the use of moderated t-test and F-tests. It can handle one-channel oligonucleotides arrays as well as two-channel cDNA arrays with and without dye swapping. Furthermore, it can incorporate mixture models and thus allows for incorporation of spot & array quality weights and other covariates in the analysis. The method mitigate heterogeneity of variations in the different features by adapting the hierarchical models of Lonnstedt and Speed (2002) to high-throughput expression data in an empirical Bayes setting (Robbins, 1956; Casella, 1985; Efron et al., 2001; Efron, 2003).

The analysis of complicated experiment designs in limma is refined and organized by the help of two matrices. The first is the *design matrix* which is used to store coefficient's pointer of the different RNA targets which have been hybridized to the different arrays. The design matrix contains a row for each array (sample) and a column for each comparison condition (phenotype). This way, the design matrix defines the statistical dependencies of samples between different conditions and replicates using biological factors underlying the experimental layout. The second matrix is the *contrast matrix* which allows coefficients of array sets defined in the design matrix to be combined into contrasts of interest that correspond to the planned experimental comparisons.

Limma starts by expressing each gene response in a linear models setting. Let $Y_g^T = y_{g1}, y_{g2}, \ldots, y_{gn}$ be the normalized expression vector of gene g in n microarays (log-ratios or log-intensities, see subsection 3.2.1). A single observation form the above

vector can be expressed by mean expressions μ_{gj} and an error term $\varepsilon_{gj} \sim N(0, \sigma_g^2)$ as:

$$y_{gj} = \mu_{gj} + \varepsilon_{gj} \tag{3.5}$$

In experiment designs with multiple conditions the mean expressions can be expressed as:

$$\mu_{gj} = \boldsymbol{X}_j^T \boldsymbol{\beta}_g \,, \tag{3.6}$$

where $\boldsymbol{X}_j^T$ is the row in a design matrix of full column rank that corresponds to array j. Testing the null hypothesis $H_0 : \mu_{treatment} - \mu_{control} = 0$ is now equivalent to testing the corresponding null hypothesis that the individual contrast coefficients are equal to zero; $H_0 : \beta_{gj} = 0$. The ordinary empirical t-statistic for regression coefficient under certain assumption can be formulated as follows:

$$t_{gj} = \frac{\hat{\beta}_{gj}}{\sqrt{s_g^2}\sqrt{v_{gj}}} \,, \tag{3.7}$$

where v_{gj} is the variance of the coefficient $\hat{\beta}_{gj}$, and the residual variance s_g^2 is assumed to follows a χ^2 distribution. However, as discussed before the denominator in Equation 3.7 dose not consider variance heterogeneity among the different features. This problem is solved by estimating a moderated variance $\grave{s}_g^2$.

The empirical Bayes method is used to estimate a prior variance s_0^2 for each feature by borrowing information from the other features under the assumption that the variance σ_g^2 for feature g follows an inverse gamma distribution. The method adapts the hierarchical models of Lonnstedt and Speed (2002) to high-throughput expression data in order to estimate σ_g^2. A posterior feature-specific estimate $\grave{s}_g^2$ of the variance σ_g^2 is then given by a weighted combination of the s_g^2 estimate and the prior s_0^2 as follows:

$$\grave{s}_g^2 = \frac{d_o s_o^2 + d_g s_g^2}{d_o + d_g} \,, \tag{3.8}$$

where d_0 and d_g are weights that depends on the relative size of the prior degrees of freedom and the observed degrees of freedom respectively. The estimate $\grave{s}_g$ reduces to the classical standard error of mean difference s_g if $d_0 = 0$. So now, the limma moderated t-statistic $\grave{t}_{gj}$ for regression coefficients is:

$$\grave{t}_{gj} = \frac{\hat{\beta}_{gj}}{\grave{s}_g \sqrt{v_{gj}}} = \frac{\hat{\beta}_{gj}}{\sqrt{\frac{d_o S_o^2 + d_g s_g^2}{d_o + d_g}} \times \sqrt{v_{gj}}}, \tag{3.9}$$

This statistic $\grave{t}_{gj}$ is independently distributed from s_g^2, and follows a Student t-distribution under the null hypothesis: $H_0 : \beta_{gj} = 0$ with $d_g + d_0$ degrees of freedom.

Executing the analysis in experiments with multiple comparisons simultaneously can be done by performing *analysis of variance (ANOVA)* which uses a (moderated) F-test (Stuhle and Wold, 1989; de Gruyter, 1996; Kerr et al., 2000; Lee et al., 2002; Wu et al., 2003).

3.2.4 Multiple Testing Correction

Differential expression analysis involves performing thousands tests simultaneously. This leads to underestimation of the false positive rate in the tests which means type I error α (rejecting the null hypothesis although it is correct) for the individual tests is no longer valid for all tests. This multiple testing problem is addressed either by correcting the significance level α or the nominal p-value of the individual tests. A number of multiple test correction methods have been suggested.

One approach for multiple testing correction is based on controlling *family-wise error rate (FWER)* (probability of making at least one type I error in the family of tests given that the null hypothesis is true). Bonferroni (1935) suggested a method for strong control of FWER (Bonferroni, 1935, 1936). The type I error for the individual test is controlled if $FWER \leq \alpha$ regardless of which and how many null hypothesis are true. This can be translated to a new and more stringent significance level $\grave{\alpha}$ defined as:

$$\grave{\alpha} = \frac{\alpha}{n} \tag{3.10}$$

A number of other multiple test correction methods have been suggested that are also controlling FWER like the step-down correction method (Dudoit et al., 2002) and the permutation based one-step adjustment method (Wu et al., 2003). The Bonferroni multiple test correction is very conservative and sets a high threshold for significance which might drastically affects the detection power of the statistical tests. This also applies to other FWER adjustment methods, in addition some are computationally expensive.

Another suggested procedure that is less conservative and frequently used is based on the correction of *false discovery rate (FDR)*. The method balances between the true discovery of statistically significant genes, the true-positive (TP), and possibility of false-positive classifications (FP; rejection of the null hypothesis although it is correct). The FDR is computed as the ratio of the FP to the sum of FP and TP (true-positive):

$$FDR = \frac{FP}{FP + TP} \tag{3.11}$$

The FDR thus gives the expected proportion of genes which are falsely classified as differentially expressed within the total number of genes which are declared as differentially expressed. This way, it balances between specificity and sensitivity of the test classification.

A number of methods have been suggested to estimate the FDR. In case that the p-values can be easily derived from the test statistics distribution, then the methods suggested by Benjamini and Hochberg (1995) can be used to estimate FDR_{BH}. In case the test statistic distribution can not be estimated then the FDR can be established by permutation plug-in estimates (Tusher et al., 2001; Storey, 2002; Storey and Tibshirani, 2003b; Storey, 2003; Storey and Tibshirani, 2003a; Parmigiani et al., 2003; Storey, 2010). Storey (2002) suggested a direct approach to false discovery rates. Story's measure can be approximated by multiplying the above FDR_{BH} by π, where π estimates the probability that a gene is not differentially expressed. Efron and Tibshirani (2002) suggested a modified version of local FDR using empirical Bayes.

Benjamini and Yekutieli (2001) extended the FDR estimation so that it is also valid for multiple testing under dependency. The modification proved to be useful for multiple testing under dependence structures such as in the case of gene set enrichment analysis of gene ontology.

3.3 Differential Expression Analysis Methods for Time-Course Data

Gene activities and regulation processes in the cell are intrinsically dynamic. A stationary study design is not enough to capture these processes in their full scope. Therefore, it is often the case that observations are acquired at different successive time points. In such study designs the objective usually is to uncover the underlying dynamic regulatory and transcriptional processes in the living cell by the means of

observable features (genes) of these processes. This is typically the case in time-course microarray data.

As in classical static study designs, such data have the high-dimensionality problem of microarray data and have an inherent complexity due to hidden relatedness (correlations) caused by co-expression of genes. These co-expressed genes are generally functionally related, belong to the same pathway or protein family or are controlled by same transcriptional regulatory programme (Dehmer et al., 2011). Samples of same donor in different time points are naturally related. This introduces additional relatedness (correlations) between samples at different time points. In this sense, time-course data are repeated measurement data where the sampling time-points are ordered. This ordering entails a non-uniform relationships between the different time points. It is expected that the relationship between two consecutive time-points will usually be higher than that between other time points. Because the sampling time-points are usually sparse and irregular distributed over the course of time, replicates are very essential for the analysis of such data. Relatedness within replicates of a donor further adds to the complexity of time-course expression data. Approaches that take all the complexity of time-course expression data are therefore essential for appropriate differential expression analysis.

Linear models approaches prove utility in evaluating differential expression in stationary experiments (section 3.2) (Smyth, 2004b; Chu et al., 2002; Wolfinger et al., 2001; Kerr and Churchill, 2001; Park et al., 2003). Some of these approaches are further adapted to time-course expression data (Smyth, 2004a; Aryee et al., 2009). However, linear modeling approaches treat time points as unordered thus ignoring the dynamic structure of this type of expression data. This results in weaker detection power of genes, which exhibit differential time courses.

An approach that allows for integrating the dynamic structure of time-course expression data is represented in a number of methods that involve fitting smoothed curves to the individual features and applying test statistics on these curves (Storey et al., 2005; Bar-Joseph et al., 2003). Such methods exploit the relationships between samples of same donors in the different time points. An example of such methods is the application *Extraction and Analysis of Differential Gene Expression (EDGE)* which is introduced by Leek et al. (2006) and extends the method by Storey et al. (2005). However, due to experimental constrains, sampling time points are usually sparse and irregularly distributed over the course of time. That might lead to high bias in assuming too simple model and result in uderfitting in the smoothed curves.

Within the empirical Bayes framework a number of approaches have been suggested to test differentially expressed time-courses of features. The empirical Bayes procedure is used to circumvent the problem of few sampling time-point (small sample size) by stabilizing estimates (Robbins, 1956; Efron, 2003). A number of methods have been

suggested in this context. For example a non-parametric empirical Bayes methods which permit for pre-defined relationships between the different time-points have been proposed in (Efron et al., 2001), Eckel et al. (2004) and Tai and Speed (2006). The relationships between the samples of different time points are limited to uniform correlations between the time points in this method. However, the correlation between measurements of the ordered time points is nonuniform, e.g. correlations between expression of consecutive time points are expected to be higher than correlations between non-consecutive time points. Guo et al. (2003) suggested an improved version that allows for none-uniform serial correlations between the time points. However, their method can be applied only to one-condition time-course experiments. Tai and Speed (2006) introduced a method that allows for the analysis of data from single-condition time-course experiments and for time-course experiments with multiple experimental conditions.

Aryee et al. (2009) improved the empirical Bayes approach to circumvent the drawbacks of the previous methods. Their Bayesian estimation of temporal regulation method estimates the probabilities of differentiall expression of individual features. It uses time-dependent structure in the data at different magnitude for the different time-points, and thus incorporates the dynamic and intrinsic non-uniform relationships between the different time-points. The method also allows for single- and multi-condition expression data by detecting the differences between two conditions or by comparing to a baseline in one condition. Furthermore, this method can be used for one- and two-color microarray data (see section 2.1). This method has been used in the analyses of microarray data of TGF-β stimulation study in different cell types in human and mouse (see chapter 4). In the following paragraphs the method of Aryee et al. (2009) is briefly explained (the reader is referred to the original paper for more details).

3.3.1 A Bayesian Approach for Time-Course Differential Expression Estimation

This method suggested by Aryee et al. (2009) fits two different models for each feature (gene) in the data. The first one is a simpler model in case the feature is non-differentially expressed and the other model in the case the feature is differentially expressed. In this model the differences of mean expressions of features at each time point between perturbed groups are taken as random effects. In this way it allows for none-uniform relationship structures which is typical for time-course expression data with many sources of variations. By determining which of these models better fit the time course expression data probability of differential expression is estimated for each feature in the data.

Modeling Setup

The idea behind this method is that a feature which is in reality not differentially expressed will have zero log-ratio between the two conditions in all time points (see subsection 3.2.1). Any deviation from a flat zero-based-line across all time point will be due to random noise. On the other hand, a truly differentially expressed feature will show a substantial effect δ which is empirically measured by the expression changes between the experimental groups. Let the differential expression state of a feature be represented by a Bernoulli indicator random variable $I_g \in \{0, 1\}$, where $I_g = 0$ if the feature is not differentially expressed and $I_g = 1$ if it is differeitially expressed. The aim is to estimate the probability of $I_g = 1$.

Assuming a two-condition experiment with at least two replicates in each time-point, let $X_{gi} = (X_{gi1}, X_{gi2}, \ldots, X_{giT})$ denote the log-transformed expression value for replicate i of gene g at time points $1, 2, \ldots, T$

$$X_{gi}^{(Tr)} = \mu_g^{(Tr)} + \varepsilon_{gi}^{(Tr)} \tag{3.12}$$

$$X_{gi}^{(Co)} = \mu_g^{(Co)} + \varepsilon_{gi}^{(Co)}, \tag{3.13}$$

where $\mu_g^{(Tr)}$ and $\mu_g^{(Co)}$ are mean expression of the gene g in the control (Co) and perturbed (treatment) (Tr) groups, $\varepsilon_{gi}^{(Tr)}$ & $\varepsilon_{gi}^{(Co)} \sim MVN(0, \Sigma_{Eg})$. The covariance structure Σ_{Eg} represents correlations of error terms between time-points and within the different replicates at each time point.

For one condition and equal number of replicates N define $\bar{X}_g = \sum_i X_{gi}/N$ and $\bar{\varepsilon}_g = \sum_i \varepsilon_{gi}/N$ means expression and mean error across replicates.

The objective is to examine whether the difference between the man expressions of the control and perturbed groups for each feature g is larger than zero. This fold change $Y_g = \bar{X}_g^{(Tr)} - \bar{X}_g^{(Co)}$ can then be modeled as follows:

$$Y_g = (\mu_g^{(Tr)} - \mu_g^{(Co)}) + (\bar{\varepsilon}_{gi}^{(Tr)} - \bar{\varepsilon}_{gi}^{(Co)}),\ with \tag{3.14}$$

$$Y_g \sim MVN_T\Big(\delta_g, \frac{2}{N}\Sigma_{Eg}\Big), \tag{3.15}$$

with $\delta_g = (\delta_{g1}, \delta_{g2}, \ldots, \delta_{gT})$ a vector of log ratios for all time points. These are modeled as random effects in order to capture non-uniform correlation.

The distribution of gene's expression values Y_g takes the following forms when it is differentially expressed $\delta g|(I_g = 0) = 0$ and when it is not $\delta g|(I_g = 1) \sim MVN(0, \Sigma_{Dg})$. We note that Σ_{Dg} is non-zero only for the true differentially expressed features. The distribution of expression fold change of features given it is differentially expressed or non-differentially expressed will take the following two forms:

$$Y_g|(I_g = 0) \sim f_0 \equiv MVN_T\Big(0, \frac{2}{n}\Sigma_{Eg}\Big) \tag{3.16}$$

$$Y_g|(I_g = 1) \sim f_1 \equiv MVN_T\Big(0, \Sigma_{Dg} + \frac{2}{n}\Sigma_{Eg}\Big), \tag{3.17}$$

The Bayes rule is used to infer the probability of differential expression of a feature g as follows:

$$P(I_g = 1|Y_g) = \frac{\rho f_1(Y_g)}{(1-\rho)f_0(Y_g) + pf_1(Y_g)}, \tag{3.18}$$

with ρ being the proportion of the differentially expressed features, defined as those which have $P(I_g = 1|Y_g)$ greater than a threshold $1 - \alpha$.

Parameter Estimation

The unknown parameters Σ_{Eg} and Σ_{Dg} are estimated from the data. The variance of the treatment group Σ_{Eg} is estimated by the empirical pooled sample covariance as follows:

$$S_{Eg} = \frac{(N^{(Tr)} - 1)S^{(Tr)} + (N^{(Co)} - 1)S^{(Co)}}{N^{(Tr)} + N^{(Co)} - 2} \tag{3.19}$$

with $S^{(Tr)} = \sum_i (X_{gi}^{(Tr)} - \bar{X}_g)(X_{gi}^{(Tr)} - \bar{X}_g)^T/(N^{(Tr)} - 1)$ and $S^{(Co)} = \sum_i (X_{gi}^{(Co)} - \bar{X}_g)(X_{gi}^{(Co)} - \bar{X}_g)^T/(N^{(Co)} - 1)$. The variance in the case of differential expression Σ_{Dg} is estimated by a weighted average of the sample covariance matrix $S_{Dg} = Y_g Y_g^T$ and a target matrix represented by the a covariance matrix for differentially expressed features only i.e. where $\hat{I}_g = P(I_g = 1|Y_g)$ is above a threshold $1 - \alpha$. ρ is estimated by the proportion of the features that fulfills $\hat{I}_g \geq 1 - \alpha$.

An iterative procedure is used to update parameter estimations of Σ_{Dg} and I_g successively until convergence. This procedure starts with a default estimate for ρ and initial gene ranking. This iterative algorithm is found to be fast converging

(3-6 iterations). The final $1 - \hat{I}_g$ are corrected for multiple testing using Storey's positive false discovery rate method (Storey, 2002) and the probabilities of differential expression of features are driven from them.

Method Evaluation

The performance of the method has been compared by the authors to the Linear Models for Microarray Analysis (limma) method (Smyth, 2004a), Extraction and analysis of Differential Gene Expression (EDGE) (Storey et al., 2005; Leek et al., 2006; Storey et al., 2007) and the Multivariate empirical Bayes MB-statistics (Tai and Speed, 2006) through application on simulated data where differentially expressed features are known. The authors showed that their method has higher detection power at a lower false positive rates.

3.4 Clustering Methods

Clustering is an unsupervised exploratory descriptive learning approach that has the goal of discovering new categories or groups in a data set. The aim of the various clustering algorithms is to achieve maximum homogeneity of the data elements within each group while providing a maximum separation between them. The assessment by which groups are separated is usually intrinsic i.e. driven solely from the data itself. Therefore, the groups built this way arrange the element in an efficient representation that characterize the population of which the data is driven. Formally, the clustering groups are arranged in non-overlapping k subsets (clusters) $P =: P_1, P_2, \ldots, P_k$ of the data set D such that; $D = \cup_{i=1}^{k} P_i$ and $P_i \cap P_j = \phi$ for $i \neq j$ (Kaufman and Rousseeuw, 1990; Hastie et al., 2001; Berry and Castellanos, 2007; Everitt et al., 2011).

Cluster analyses is a complementary step in the analysis of microarray data. It helps to subset features (e.g. genes or samples) in groups that show similar behavior in the experiments. Features or samples are grouped together if they are similar to each other. The degree of similarity between two genes can e.g. be assessed by comparing their expression levels. For each gene one expression vector in the n-dimensional space is defined, where n is the number of experiments. Thus, each experiment represents an axis in space and the measured expression value represents the corresponding coordinate. Gene expression data are usually represented in a matrix format. This representation helps to visualize and interpret the data. The rows of such a matrix usually contain the expression vectors of the genes and columns represent the individual experiments.

In distance-based clustering methods dissimilarity between two gene expression vectors is expressed as the distance between the vectors (where, $dissimilarity = 1 - similarity$). Depending on the type of data there are several distance metrics. For instance, there are distance metrics for nominal, ordinal and binary attributes. Most common dissimilarity/similarity measures in gene expression data include distance measures like the Euclidean (Cartesian), Mahalanobis, Manhattan, Chebyshev, Minkovsky and χ^2 distances. Other measures like the correlation coefficient and relative entropy are also frequently used.

The distance $d(X, Y)$ between two expression vectors X and Y can be assessed e.g. by the Euclidean distance measure defined as:

$$d(X, Y) = \sqrt{\sum_{i}^{n} (x_i - y_i)^2}, \tag{3.20}$$

where x_i and y_i are the expression values for the genes X and Y in the experiment $i = 1, 2, \ldots, n$.

The Euclidean distance and Pearson correlation coefficient are very often used in the *hierarchical clustering* e.g. for visualization of gene expression data in *heatmaps*. The hierarchical clustering is explained in the following.

3.4.1 Hierarchical Clustering Methods

The hierarchical clustering approach is a distance-based clustering approach that builds hierarchy of clusters based on two major types of procedures; (I) agglomerative (bottom-up) algorithm, and (II) divisive (top-down) algorithm (Ward, 1963; Johnson, 1967; Kaufman and Rousseeuw, 1990; Johnson and Wichern, 2007; Moore, 2001). In case of agglomerative hierarchical clustering the algorithm begins by forming singleton clusters, so each gene expression vector has its own cluster. Pairwise distances between expression vectors are computed (e.g. using the Euclidean distance) and usually arranged in a distance matrix. The two elements with the lowest entry, that is, with the greatest similarity in the distance matrix are merged to form a cluster. The distances between this new cluster and the remaining elements are calculated in the following step. The algorithm continues until all elements are contained in one cluster. The divisive procedure follows an opposite logic where all elements initially put into one cluster that is recursively split into smaller clusters based on the recomputed distances until each element hat its own cluster. Merging and splitting clusters is a greedy algorithm problem. Complexity of agglomerative clustering ($O(n^3)$) is

reduced in single linkage and complete linkage applications. For the calculation of the distance between two clusters there are different linkage criteria:

- Single Linkage; the distance $d(I, J)$ between two clusters I and J is taken as the *minimum* distance between any element from cluster I and any other element from cluster J:

$$d(I, J) = min(d(i, j) \mid i \in I, j \in J) \tag{3.21}$$

- Complete Linkage; the distance $d(I, J)$ between two clusters I and J is taken as the *maximum* distance between any element from cluster I and any other element from cluster J:

$$d(I, J) = max(d(i, j) \mid i \in I, j \in J) \tag{3.22}$$

- Average Linkage; the distance $d(I, J)$ between two clusters I and J is taken as the *average* of distances between all element from cluster I and all other element from cluster J:

$$d(I, J) = Arg_{Average}(d(i, j) \mid i \in I, j \in J) \tag{3.23}$$

- Ward's minimum variance criterion; the cost $d(I, J)$ of merging two clusters I and J is taken to optimize (minimize) an objective function e.g. the error sum of squares. The initial distance is considered a weighted squared Euclidean distance proportional to:

$$d(I, J) = \frac{n_I \times n_J}{n_I + n_J} \mid\mid m_I - m_J \mid\mid^2, \tag{3.24}$$

  where m is the center of the cluster and n its size.

The result of a hierarchical clustering can be visualized in a *Dendrogram*. A Dendrogram is a binary tree, in which each leaf represents one gene (Kaufman and Rousseeuw, 1990). An agglomorative hierarchical clustering algorithm has been used in a context of consensus clustering procedure (subsection 3.4.2) for grouping chemical compounds based on a novel compound similarity measure in chapter 6.

3.4.2 The Consensus Clustering Approach

Generic clustering methods try to find optimal clustering based on method-specific clustering criteria and from a single run. Thus, applying different clustering algorithms on the same data often produce different clustering results. The algorithms used in these methods are often heuristic so that different runs of the same clustering algorithm produce different results. Many clustering methods are not equipped with internal system to determine the adequate number of clusters and do not provide for assigning confidence to the resulting clusters. Clustering stability is a major problem particularly in high dimensional gene expression data with typically small sample size (Berry and Castellanos, 2007; Everitt et al., 2011).

Few methods have been suggested for the assessment of clustering stability from simulated data perturbations gained by re-sampling from the original data (Jain and Moreau, 1987; Levine and Domany, 2001; Ben-Hur et al., 2002; Dudoit and Fridlyand, 2002; Tibshirani and Walther, 2005). The consensus clustering (also called aggregation or ensemble clustering) is such an elaboration of the classical clustering problem that also provide for the assessment of clustering stability with respect to sampling variability (Strehl and Ghosh, 2003; Monti et al., 2003; Senbabaoglu et al., 2014b,a). The consensus clustering involves sub-sampling from the data points and an optimization problem of reconciling clustering ensemble to consensus formulated e.g. as an objective function. The criteria that two data points belong to same consensus cluster is represented here by the relative frequency of these two points falling in same clusters in an ensemble of clustering runs. The optimization problem of the consensus clustering has been shown to be NP-complete (Filkov and Skiena, 2003). Therefore, iterative algorithms such as expectation maximization (EM) algorithm are used to estimate objective function optimization. The consensus clustering procedure can often be combined with many clustering algorithms.

There are many versions of consensus clustering that majorly differ from each other in the way their consensus optimization problem is formulated and solved. For example, Strehl and Ghosh (2003) suggested graph and hyper-graphs representations of the pairwise consensus similarity matrix and implemented three different algorithms to partition the graph/hypergraph into clusters. Topchy et al. (2003) proposed two approaches; an approach that uses estimation maximization (EM) algorithm to estimate maximum likelihood of a mutual information objective function (Topchy et al., 2003), the other approach uses EM algorithm to estimate maximum likelihood of a mixture model (Topchy and Jain, 2004). Abu-Jamous et al. (2013a,b) demonstrated the utility of a consensus clustering of certain data points base on multiple clustering methods as well as based on multiple information sources. Nguyen and Caruana (2007) proposed three algorithms for finding the consensus clustering in clustering

ensemble. Their approach allows for different numbers of clusters in the clustering runs in an ensemble, however in the final consensus clustering they left the question of determining appropriate number of clusters unanswered.

Monti et al. (2003) proposed a consensus clustering approach that provides a clever solution for most of the draw-backs of the consensus clustering approaches mentioned above. This method is based on multiple runs of a base clustering algorithm where each base clustering algorithm is allowed to have a different number of clusters. The method provides an internal criterion to determine the appropriate number of clusters in the provided range. Furthermore, the approach allows for the use of many clustering algorithm and uses resambling to asses cluster stability. In our application in chapter 6 we used this consensus clustering approach proposed by Monti et al. (2003). This is explained below in short (for further details refer to the original paper).

Given a dataset $D = e_1, e_2, ..., e_n$ the aim of clustering is to divide the data into exhaustive and not overlapping k-clusters. Formally; $P \equiv P_1, P_2, ..., P_k$ such that $\cup_{k=1}^{k} P_k = D$, and $P_i \cap P_j = \phi, \forall_{i,j}; i \neq j$. In consensus clustering the dataset D is resampled into H datasets, where an Indicator $(n \times n)$ matrix I^h controls whether items i and j are present in dataset D^h by a corresponding entry of 1 if so, and 0 otherwise. By applying the clustering algorithm on D^h, a connectivity $(n \times n)$ matrix M^h can be defined as follows:

$$M^h(i,j) = \begin{cases} 1 & \text{if item i and j belog to the same cluster,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.25}$$

For cluster number k a consensus matrix M^k can be defined by the normalized sum of all connectivity matrices for all resampled data sets as follows:

$$M^k(i,j) = \frac{\sum_h M^h(i,j)}{\sum_h I^h(i,j)} \tag{3.26}$$

Choosing appropriate number of clusters

A consensus matrix with values either 1 or 0 amounts to a perfect consensus. However, generally these values are positive fractions. By running the clustering for a series of cluster numbers $(K = 2, 3, ..., K_{max})$, one can chose the number of clusters such that the dichotomized consensus matrix is as close as possible to the perfect one. For this purpose a measure of area under the empirical cumulative distribution function (CDF) for each M^k is computed as follows:

$$A(k) = \sum_{i=2}^{m} [x_i - x_{i-1}] CDF(x_i) \tag{3.27}$$

where $[x_1, x_2, ..., x_m]$ is the sorted set of entries of the consensus matrix and $m = n(n-1)/2$ is the number of its possible sorting. The CDF for a histogram of a consensus matrix M^k over the range [0, 1] is defined as:

$$CDF(c) = \frac{\sum_{i<j} I\{M(i,j) \leq c\}}{n(n-1)/2} \tag{3.28}$$

where $I\{...\}$ is an indicator function, The area under CDF (AUC) supposedly increases by increasing number of clusters. The maximal relative increase $\frac{A(k+1)-A(k)}{A(k)}$ delineate the appropriate number of clusters in the data.

Silhouette Analyses of Consensus Clusters

The silhouette analysis is a method to assess the quality of the resulting sets produced by a clustering algorithm. The idea is to compare within cluster distances and between cluster distances for each point in the data. A silhouette value is computed for each data point assessing its relatedness to all other point in its appointed cluster compared to data points in other clusters. Formally, the silhouette value for a data point i is defined as follows:

$$S(i) = \frac{(b(i) - a(i))}{max(a(i), b(i))}, \tag{3.29}$$

where $a(i)$ is the average distance of the point i to all other points in its appointed cluster. $b(i)$ is the distance of the data point i to the nearest neighboring cluster, measured by the minimum of the by-cluster-averaged distance of the data point i to all the points in each one of the other clusters. The value of $S(i)$ is between [0, 1] and averaging these value in a cluster give cluster's silhouette width and averaging the values for all data points give silhouette width for the clustering algorithm. Silhouette width values closer to 1 indicate compact cluster which are distinct from each other and a good performance of the clustering algorithm.

3.4.3 Clustering Methods For Time-Course Data

Clustering time-course data serves as a tool for identifying co-regulated features and discovering sets of features which have similar temporal or spatial expression patterns. This can thereafter be used to simplify the analysis of gene regulatory networks in order to detect cellular processes underlying these networks and assign functionality to genes (Sturn et al., 2002; Shannon et al., 2003; Gollub and Sherlock, 2006; Song et al., 2007).

Although probabilistic-model-based approaches predominate in the analysis of time-course expression data, generic distance-based clustering methods like k-means, hierarchical clustering and self-organizing maps (SOM) are still quite popular (Costa et al., 2004; Jonnalagadda and Srinivasan, 2008; Chen, 2009). For instance, a modified correlation coefficient is suggested which takes into account the concordance between two temporal expression profiles and order of time points at which maximum and minimum expression levels are measures in the two profiles (Son and Baek, 2008). Other proposed distance metrics include the Edwardian, maximum and Manhattan distance metrics (Miller, 1974; Efron, 1982; Heyer et al., 1999; Scharl and Leisch, 2006).

The distance measure used for clustering greatly influence the resulting clusters. However, there is no methodological guidance for selecting appropriate distance measure in the clustering process, so that this task remains difficult. Furthermore, not all distance measures are distance metrics. This introduces many problems. For example, when assessing clustering performance on multiple data sets, distance measure normalization is necessary to match their ranges in order to asses the different results. Conversely, the effective comparison of different clustering results based on distinct distance measures on a given data set becomes highly challenging. Many of the distance measures can not handle outliers and most of them can not even be computed if parts of the data are missing. All these issue make distance-based clustering unfavorable for time-course gene expression data.

A number of probabilitsic model approaches have been proposed for clustering time-course gene expression data. Fraley and Raftery (2002, 2003) sugested a general multivariate Gaussian model for clustering time-course expression data that takes into account the relatedness structure between time point. This model considers the time points only as un-ordered observations. However, the time order in the data is very important for feature expression and cluster interpretation.

Yi et al. (2009) proposed a clustering method that ranks feature's temporal expression values after discretizing them. This method then uses a bootstrap significance test to classify features into predefined candidate profiles. Apart from the fact that

data discretization will inevitably results in information loss, the method uses the Euclidean distance to measure dissimilarity between rank vectors which might lead to sporious correlation problem (Pearson, 1897). In addition, no suggested mechanism by which the number of candidate profiles and consequently number of clusters could be internally determined. Another class of clustering methods for time-course gene expression data are based on smoothing splines models. Chen (2009), for instance, used self-organizing maps (SOM) for clustering longitudinal expression data utilizing cubic smoothing splines. First, Each feature is represented by a cubic smoothing spline thus adhering for time-course data requirements. Then the SOM clustering is applied on these smoothed splines.

Further spline methods have been introduced for high-dimensional longitudinal expression data (Luan and Li, 2003, 2004; Coffey et al., 2014). In these method the mean expression of a supposed cluster is modeled as a linear combination of spline bases of all features falling in that cluster. James and Sugar (2003) proposed a curve-based approach for sparsely sampled time-course data. Their method can be used to handle missing data or to predict parts of missing portions of the features fitted curves with confidence intervals.

Effectively, the spline features (number of knots and spline bases) are treated as fixed and has to be pre-defined. In particular the knots dose not necessarily correspond to the measurement time-points in these methods. Setting pre-specified basis spline function is not done only to find best shapes that tally that data, but also in order to avoid over-fitting. However, while different choices renders different shapes of spline curves, the pre-selection of spline bases and knots is still presumptuous and lacks methodological guidance (Ruppert et al., 2003). In Addition, the estimation maximization (EM) algorithm used in these methods is not feasible for high dimensional expression data with tens of thousands of features. Furthermore, the problem of defining the number of clusters is not solved in most of these methods.

Ma et al. (2006) proposed a data-driven method that adheres to requirements of time-course expression data and overcomes some disadvantages of previous methods. A mixed-effect smoothing spline model is used for curves in the data. A rejection-controlled estimation maximization (RCEM) algorithm is used to fit the model. This algorithm is less computationally expensive than the traditional EM algorithm. The method provides cluster assignment of the features in the data, predicts the mean curve for each cluster accompanied with confidence interval bands and R^2 values. Mean curve estimates for the clusters and for the individual features are made simultaneously. The number of clusters is determined internally and automatically using Bayesian information criterion (BIC). The method by Ma et al. (2006) is described briefly in the following.

Model Formulation

The function $y_j = f(t_j) + \varepsilon_j$ can be used to represent expression values with respect to time t_j $(j = 1, 2, \ldots, T)$, where the error terms $\varepsilon_j \sim N(0, \sigma^2)$. Such curves are fitted by minimizing the residual sum of squares (RSS): $RSS = \sum_{j=1}^{T}(y_j - f(t_j))^2$.

In order to enforce smoothness of f the second derivative at each time point should remain sufficiently small. That means for a specific positive value η, the following constraint over a twice-differentiable f is assumed:

$$\int (f''(t))^2 \, dt < \eta, \tag{3.30}$$

Formulation of the optimization problem of residual sum of squares (RSS) subject to the above constraint in Lagrange multiplier method is then equivalent to minimizing the following:

$$\sum_{j=1}^{T}(y_j - f(t_j))^2 + \lambda T \int (f''(t))^2 \, dt, \tag{3.31}$$

This function f is represented via a cubic smoothing spline:

$$\hat{f}(t) = d_0(\lambda) + d_1(\lambda)t + \sum_{j=1}^{T} c_j(\lambda) \int (t_j - \mu)_+ (t - \mu)_+ \, du, \tag{3.32}$$

where $(.)_+$ refer to the positive part of the numbers.

Mixed-effect Model

Each gene's time course may differ from the mean curve by a certain shift. In the model this is represented by a random effect i.e. a zero mean centered random variable. This way, the observed expression of a feature i at time-point j belonging to cluster k can be expressed as:

$$y_{ij} = \mu_k(t_{ij}) + b_i + \varepsilon_{ij}, \tag{3.33}$$

where μ_k is a cluster-specific shape function of time represented by the mean curve of cluster k, $b_i \overset{ind}{\sim} N(0, \sigma_{bk}^2)$ and $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ is a measurement random error, with

$i = 1, 2, \ldots, N$ feature labels, $j = 1, 2, \ldots, T$ time labels and $k = 1, 2, \ldots, K$ cluster labels.

For a vector of observations over time-course y_i and cluster mean vector over time-course μ_k the equation 3.33 above is equivalent to:

$$y_i \sim (\mu_k, \Sigma_k) \tag{3.34}$$

if $\Sigma_k = \sigma_{bk}^2 E_{T\times T} + \sigma^2 I_{T\times T}$, where $E_{T\times T}$ is a square matrix of ones of dimensions $T \times T$, and $I_{T\times T}$ is the identity matrix of dimension $T \times T$. The variation in the clusters are combinations of covariance structure of feature-specific random effect b which depends on cluster k and variance term $\sigma^2 I_{T\times T}$ which depends only on σ_{bk}^2.

The cluster membership of genes and the number of clusters are not known a priori. Therefore, the time-course expression of a feature y_i can be modeled by a mixture of Gaussian distributions:

$$y_i \sim p_1 N(\mu_1, \Sigma_1) + p_2 N(\mu_2, \Sigma_2) + \cdots + p_K N(\mu_K, \Sigma_K) \tag{3.35}$$

where the cluster mean curve μ_k and cluster curve variance Σ_k are defined above. $p_1, p_2, \ldots, p_K$ are the size proportions of the clusters and represent the probabilities that a certain feature i belongs to a cluster k.

Parameter Estimation via Rejection-Controlled Expectation Maximization Algorithm (RCEM)

Estimates of the parameters p_k and μ_k can be obtained by maximizing a penalized log-likelihood function. Such penalized log-likelihood function can be expressed in terms of a smoothness constraint, feature-specific shifts and a random measurement error for each cluster. However, such log-likelihood functions are not traceable for multiple clusters with simultaneous assignment of features to clusters.

A possible alternative that circumvent this problem is using the expectation maximization (EM) algorithm to estimate the p_k and μ_k parameters. However, the EM algorithm is not stable and computationally very expensive specially in case of time-course gene expression data with thousands of features. Ma et al. (2006) suggested a modified version of EM algorithm called rejection-controlled expectation maximization (RCEM) to overcome the problems of conventional EM algorithm.

The difference to conventional EM algorithm is that, very low probabilities of gene-to-cluster memberships are set to zero thus reducing the computation cost in the

M-step. The expectation step estimates the probabilities of a gene belonging to all clusters given models parameters as follows:

$$P(gene_i \in k) = \frac{p_k N(\mu_k, \Sigma_k)}{p_1 N(\mu_1, \Sigma_1), ... + p_k N(\mu_k, \Sigma_k)} \tag{3.36}$$

A weighted penalized log-likelihood is computed in the maximization step as follows:

$$-\sum_{k=1}^{K} \left\{ \sum_{i=1}^{n} P(gene_i \in k) \left(\sum_{j=1}^{T} \frac{(y_{ij} - \mu_k(t_{ij}) - b_i)^2}{2\sigma^2} + \frac{b_i^2}{2\sigma_{bk}^2} \right) - \lambda_k T \int [\mu_k^{''}(t)]^2 dt + C \right\} \tag{3.37}$$

Optimal values of smoothing paramters σ_{bk}^2 and λ_k are chosen via a generalized cross validation procedure (Gu and Ma, 2005; Craven and Wahba, 1978). The parameters P_k are updated in this step as follows:

$$p_k = \frac{\sum_{i=1}^{n} P(gene_i \in k) + a_k}{\left(n + \sum_{k=1}^{K} a_k\right)} \tag{3.38}$$

The expectation and maximization steps are repeated until convergence of gene-to-cluster assignment is reached.

Determining Number of Clusters

Bayesian Information Criterion (BIC) is used to restrict the number of clusters to an optimal trade-off between model complexity and goodness of fitting (Gu, 2004; Gu and Ma, 2005). For a number of free parameters v_k in cluster k BIC is defined as follows:

$$BIC = -2 \sum_{i=1}^{n} log \sum_{k=1}^{K} p_k N(\mu_k, \Sigma_k) + \sum_{k=1}^{K} v_k log(nT) \tag{3.39}$$

The quality of the clusters are measured in this method by R^2 which estimates the fraction of variations in the clusters that can be explained by the modeling. 95% confidence intervals of mean curve in each cluster are also computed within RCEM algorithm (Gu, 2002; Gu and Ma, 2005).

3.5 Functional & Enrichement Analysis Methods

Transcriptomic experiments investigate gene expressions in different conditions. Differentially regulated genes or genes falling into a specific cluster can further be investigated for their membership to predefined groups of functionally related gene products. These predefined groups of functionally related genes are represented by controlled vocabularies (ontologies) which are standard terminologies that describe biological concepts. They are defined by capturing experimental information from published literature. These functional sets help in structuring the analysis, formulation and interpretation of biological information in genomic studies. Gene products are assigned to such groups if they have similar biological function or belong to the same biochemical reaction chain. The motivation for this categorization of genes and gene products is that investigating individual genes is a tedious task, in addition, Many diseases are believed to be associated with modest regulation in a set of related genes rather than to a strong regulation in individual genes (Subramanian et al., 2005).

Prominent annotation databases for functional gene sets are the Gene Ontology (GO®)[1] (Ashburner et al., 2000; Camon et al., 2004) and Kyoto Encyclopedia for Gene and Genome database (KEGG®)[2] (Ogata et al., 1999; Kanehisa and Goto, 2000). These databases are conceptually different from each other. The latter database groups genes based on their participation in biochemical pathways and the former categorizes genes based on molecular functions, cellular compartments and biological processes. Furthermore, these databases differ in their underlying structure.

The elucidation of biological concepts associated to differentially expressed genes is very important for the analysis and interpretation of genomic data. The number of these concepts and functional gene sets is growing due to the availability of annotation. All that drives the need for systematic analyses and visualization methods. A number of methods have been proposed which utilize different test statistics that reply on various statistical hypotheses. They all assess the association or enrichment of a functional gene set mostly by assigning a p-value to it that reflects the significance of the association.

The *overrepresentation analysis* (ORA) tests the overlap of the differentially regulated genes and the predefined functional gene sets (Breitling et al., 2004; Geistlinger et al., 2011). The hypergeometric test is widely used and accepted in ORA (Khatri and Draghici, 2005).

[1] The Gene Ontology Consortium Database: `http://www.geneontology.org/`
[2] KEGG® Database: `http://www.genome.jp/kegg/`

A limitation of the ORA is the usual discrimination of genes into a set of interest (e.g. the differentially expressed ones) and the large rest and the need for a predefined threshold for this discrimination. This discrimination is highly sensitive to the defined threshold so that different threshold choices may lead to dramatically different enriched categories, and thus different biological conclusions (Pan et al., 2005). Moreover, any ranking (e.g. by p-value of fold change) is completely ignored (Goeman and Bühlmann, 2007). Nevertheless, ORA remains quite popular due to the propitious features of the hypergeometric test which only requires that the *set of interest* is clearly defined without the need of any other additional information. This is especially useful when no such additional information is available, for example when investigating genes within a given cluster.

The *Gene Set Enrichment Analyses* (GSEA) method avoids the shortcomings of ORA (Subramanian et al., 2005; Dinu et al., 2009). In this method the ranking of the genes is compared to a uniform distribution using a Kolmogorov-Smirnov test. Various versions of the GSEA have been proposed (Dinu et al., 2009). Rather than considering only the subset of differentially expressed genes, GSEA takes the whole list of features and use their P-Values for ranking. Therefore, it is not limited to significant features. GSEA also considers features with moderate significance. This is beneficial, because biological relevance of features is not always reflected only by their significance (Geistlinger et al., 2011).

Most gene set enrichment and association analysis methods assume independence and do not consider correlations between genes resulting from co-regulation and co-expression mechanisms (Tamayo et al., 2012). Moreover, these methods ignore the dynamical behavior of the trascriptional process as well as the direction of regulatory interactions. Some recently established methods try to avoid some of these disadvantages (Shojaie and Michailidis, 2010; Geistlinger et al., 2011; Poirel et al., 2011; Glaab et al., 2012; Massanet-Vila et al., 2012). These are, predominantly, graph based approaches that exploit the network structure and ontology of the functional annotation in the databases to enhance the enrichment analysis. These methods are, however, not of further interest for this thesis.

In gene set enrichment analysis the tests are usually repeated several times (for hundreds of GO® terms of KEGG® pathways). Therefore, multiple test correction should be performed (see subsection 3.2.4). The method suggested by Benjamini and Yekutieli (2001) for multiple testing under dependency is used for this purpose. This method is useful for multiple testing under dependence structures such as in the case of gene set enrichment analysis of gene ontology (GO®).

In the following paragraphs, two gene set enrichment and association methods used in the subsequent chapters are briefly explained, namely the ORA approach and a test that is based on univariate logistic regression.

3.5.1 Over-representation Analysis & The Hypergeometric Test

ORA approach tests whether certain *functional gene set* (pathway, GO-term) within the gene *set of interest* (differentially expressed genes) are more often contained than expected by chance. For doing so, a two-dimensional contingency table is constructed for the overlap groups of the tow sets considering the gene universe, i.e. the exhaustive set of all genes (Table 3.1). The entries of this table, as shown exemplary here below, are therefore non-negative integers.

Table 3.1: Two-dimensional contingency table. N represents the population size (gene universe), n the sample size (size of functional set) and k represents the number of successes (or the probability of success).

	in *set of interest*	¬ in *set of interest*	Total
in *functional gene set*	k	n - k	n
¬ in *functional gene set*	M - k	N + k - n - M	N - n
Total	M	N - M	N

The probabilities of observing all possible sets of frequencies as they appear in the contingency table given row and column totals are computed using a hypergeometric distribution (Agresti, 1996, 2002). The hypergeometric distribution, like the binomial distribution, is a discrete probability distribution that models the number of successes k in a sequence of n trials using an *urn model*. However, unlike the binomial distribution the hypergeometric distribution is based on the notion of sampling without replacement. Considering the following null hypothesis: H_0: there is no association between the *set of interest* and the *functional gene set*, the exact probability mass function of observing k in the contingency table given its row and column sums is calculated in a hypergeometric distribution by the formula:

$$P(X = k) := f(k|N; M; n) = \frac{\begin{pmatrix} M \\ k \end{pmatrix}\begin{pmatrix} N-M \\ n-k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}, \tag{3.40}$$

where $\begin{pmatrix} a \\ b \end{pmatrix}$ is a binomial coefficient.

In the case of over-representation (enrichment) the p-value is calculated by summing up the probabilities of all the frequency tables that are more extreme ($P(X \geq k)$) to

the observed frequency table. Reciprocally, $P(X \leq k)$ define the probabilities of the not-so-often investigated case of under-representation (depletion). If the p-value is small enough ($p - value \leq \alpha$; usually a significance level $\alpha = 0.05$ is considered) then the null hypothesis can be rejected and it is concluded that the gene *set of interest* is statistical-significantly enriched (or depleted) in the *functional gene set* and the overlap between the two sets is not due to random chance.

Falcon and Gentleman (2007) proposed a conditional hypergeometric test procedure, primarily for GO-terms, to avoid the problem of highly overlapping ontologies. Their method (available in the R-package 'GOstats') considers the hierarchical graph structure of gene ontology (Alexa et al., 2006; Falcon and Gentleman, 2007). It starts the testing from the bottom of the graph and remove features of the significant tested children.

3.5.2 Univariate Logistic Regression-based Association Analysis

A logistic regression-based method has been proposed for functional association of gene set categories to the differential expression level of genes (Sartor et al., 2009; Montaner and Dopazo, 2010). This method overcomes the drawbacks of ORA methods as it does not requires significance threshold, considers all genes in the chip and it uses the differential expression analysis ranking criteria. Unlike the GSEA, this method does not depend on which or how many genes have been measured in the chips and/or ranked in the differential expression analysis.

Let π represents the relative size of the category, i.e. the probability that a gene falls in this category c at a certain significance level. The odds corresponds to this probability can be modeled by the logistic regression:

$$log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x, \tag{3.41}$$

where the explanatory variable x represents significance of differential expression. Here x is defined as $-log_{10}(p\text{-}value)$. The slope parameter β corresponds to the change in log-odds values when changing x. If changing the values of x results in changes in log-odds values then it is concluded that the category c is associated with the differential expression. Thus, a null hypothesis for association of a specific category c can be formulated as: $H_0 : \beta_c 0$ (Enrichment: $\beta \neq 0$, depletion $\beta < 0$). This can be assessed using the Wald test which, in this case, can be formulated as:

$$W = \left(\frac{\hat{\beta}}{s_{\hat{\beta}}}\right)^2 \tag{3.42}$$

where $\hat{\beta}$ is a maximum likelihood estimate of β and $s_{\hat{\beta}}$ is the estimated standard deviation of $\hat{\beta}$. Under the null hypothesis this test follows a χ^2 distribution (Buse, 1982; Phillips, 1986).

The authors showed, based on experimental and simulated data, that this method outperforms other relevant methods in the identification enriched GO® terms and the reproducibility of the results. This method is extended to time-course expression data (Sartor et al., 2009) and a multivariate version is proposed by Montaner and Dopazo (2010) for additional covariates.

CHAPTER 4

TRANSFORMING GRWOTH FACTOR BETA (TGF-β) STIMULATION EFFECTS IN DIFFERENT TISSUE TYPES OF HUMAN AND MOUSE

4.1 Introduction

The focus of this thesis is the study of pattern discovery in perturbation experiments. Detecting phenotypic responses patterns to cellular perturbation from array data is, therefore, an important part. This chapter shows how different cell systems which have been exposed to the same perturbation, namely TGF-β1 stimulation, exhibit common response patterns on the level of biological functions, pathways and protein networks. This allows for conclusions about the treatment and the different cell systems and helped to enhance the knowledge about the interactions of the hormone-regulating transforming growth factors signaling pathway.

The transforming growth factor-beta1 (TGF-β signaling pathway is a fundamental pathway in the living cell, which plays a role in many central cellular processes. The TGF-β superfamily contains over 30 different proteins, such as BMPs, Activins, Inhibins, and the TGF-β isoforms (Giacomini et al., 2006; Hinck, 2012; Renner et al., 2004). The pathway contributes to regulation of various cellular processes, such as apoptosis, cell differentiation, cell growth as well as tumor suppression and immune regulation processes (Oomizu et al., 2004).

There are three TGF-β isoforms (TGF-β1, TGF-β2, TGF-β3) which have different physiological and pathological effects on epithelial, endothelial, lymphatic, myeloid and mesenchymal tissues (Hentges and Sarkar, 2001). The TGF-β pathway is one of the most studied pathways (Alexandrow and Moses, 1995; Massagué, 1990, 1998; Roberts AB, Heine UI, Flanders KC, 1990; Roberts AB, 1993). However, the complex and sometimes contradicting mechanisms by which TGF-β yields phenotypic effects is not yet completely understood (Hentges and Sarkar, 2001). The classical TGF-β pathway is already well established since several years (Massagué, 1998). However, the identification of alternative signaling pathways that contain different receptors and Smad proteins has increased the overall complexity of the TGF-β signaling pathway (Orlova et al., 2011). Figure 4.1 shows a simplified cartoon sketch comprising mainly Smads cascades in the TGF-β signaling pathway.

In this study the downstream effects of TGF-β perturbation on the dynamical response of gene expression in mouse and human in different cell and tissue types are investigated and compared. Two types of mouse hematopoietic progenitor cells were used: multipotent progenitor (MPP) and dendritic cell (DC) committed progenitors, referred to as common dendritic progenitor (CDP) cells. CDP differentiate from MPP and give rise to two types of DC: plasmacytoid DC (pDC) and conventional DC (cDC). MPP and CDP were obtained from bone marrow by *in-vitro* culture with a specific cytokine cocktail and FACS sorting (Felker et al., 2010; Seré et al., 2012). Further cell type that is also employed is the human mesenchymal stromal cell type (MSC), which differentiate into osteocytes, chondrocytes or adipocytes (Dominici et al., 2006; Wagner and Ho, 2007; Walenda et al., 2012). Finally, primary murine hepatocytes (HPC) and immortalized human hepatocytes (human HPC, HepG2) cells were used. These different cell types are taken for three reasons: (i) All these cells are highly responsive to TGF-β. (ii) The different cell types reflect different degrees of differentiation. (iii) The different cells show a variable response to TGF-β. While in hepatocytes TGF-β induces apoptosis, multipotent progenitors initiate a differentiation programme in response to TGF-β.

Very little and vague information is known about the detailed influence of TGF-β in these different cell systems. For example, TGF-β is known to be necessary for MSC proliferation. It is essential for chondrogenic differentiation. On the other hand, TGF-β participates in inhibition of adipogenic and osteogenic differentiation. Furthermore, there are evidences, that TGF-β contributes to supporting myogenic differentiation of MSC (Gao et al., 2010; Post et al., 2008; Wang et al., 2002). There are also evidences that the TGF-β pathway play a role in the induction of cellular senescence in MSC (Ito et al., 2007). Although TGF-β1 triggers primary early responses (e.g. Smad activation) and EMT in human HPC (HepG2) cells, cell cycle arrest and apoptosis are generally not promoted by TGF-β1 (Buenemann et al., 2001; Xu XM, Yuan GJ, Li QW, Shan SL, 2012). Furthermore, TGF-β1 is known to be

crucial for development of Langerhans cells, the cutaneous contingent of migratory dendritic cells, both *in-vivo* and *in-vitro* and it evidently contributes in accelerating their differentiation and directing their subsets specification toward cDCs (Borkowski et al., 1996; Felker et al., 2010; Strobl et al., 1996, 1997).

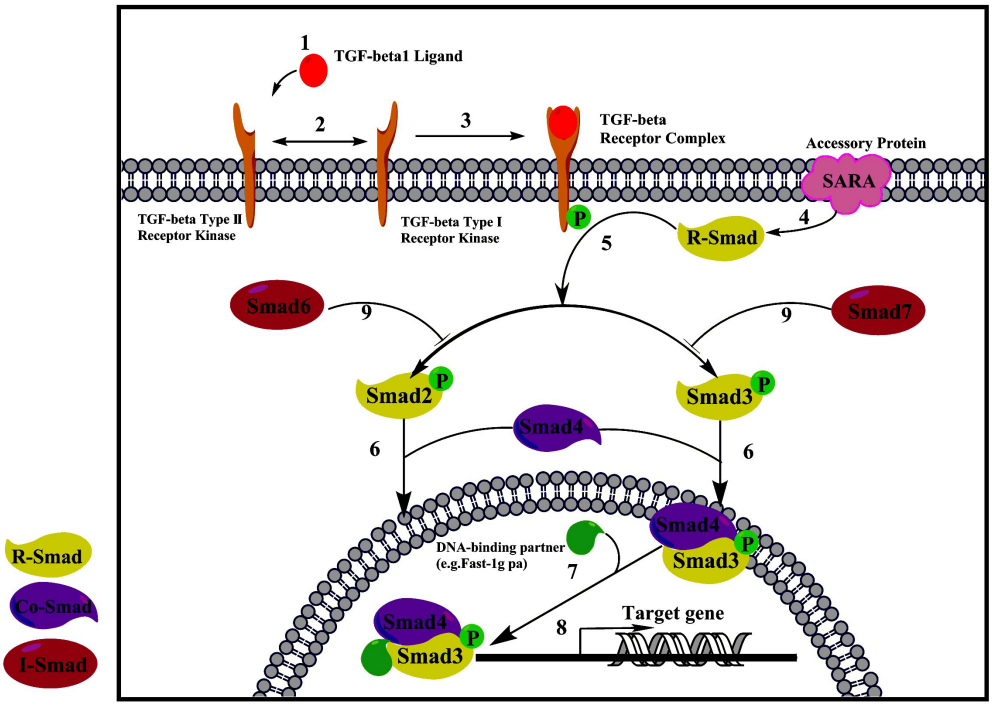

Figure 4.1: Transforming growth factor-β1 (TGF-β1) docks on a type II (1) and type I TGF-β receptors (TGF-βRI and TGF-β RII) (2). The two receptors then form a receptor complex where TGF-βRI get phosphorylated (3). Subsequently, TGF-βRI phosphorylates the receptor-regulated cytoplasmic proteins (R-Smads) Smad2 and Smad3 (5). This happens with the help of accessory proteins e.g. SARA which is located in the extracellular matrix (ECM) (4). The R-Smads form a complex and bind with the phosphorylated common mediator (co-Smad) Smad4 and transduce into the nucleus (6). There they interact with different DNA proteins, co-activators and co-repressors (7) to induce or suppress the transcription of numerous target genes (8). The inhibitory Smads (I-Smads) Smad6 and Smad7) form a negative feedback and mark the receptors for degradation (9) while R-Smads become inactive by the Smurf effect (modified after (Massagué, 1998)).

A panel of bioinformatics methods is used, ranging from statistical testing over functional and promoter sequence analysis to clustering for pattern discovery in the gene expression time series data. Only one gene, the SKI-like oncogene (Skil), was commonly found to be differentially expressed (DE) in all cell types. Skil is a component of the SMAD-pathway, which regulates cell growth and differentiation. Moreover, Smad7 that blocks TGF-β receptor activity seems to play a major common role, because it was identified as DE in most cell types. Despite of the differences on the level of individual genes a conserved effect of TGF-β perturbation on a number of biological processes and pathways is observed. Moreover, a number of overrepresented Transcription Factor Binding-Sites (TFBS) could be identified. These are commonly found in several cell types. Specifically EGR1 seems to have major relevance for the transcriptional perturbation response in mouse and human.

By analysis of an independent dataset on human A549 lung adenocarcinoma cells (CRL) from GEO (access No. GSE17708: first published in Sartor et al. (2010)) we were able to reproduce a highly significant proportion of the commonly identified biological processes, pathways and transcriptional factors in the datasets. Network analysis suggests explanations, how TGF-β perturbation in different organisms could lead to the observed effects.

4.2 Material and Methods

4.2.1 Normalization and Preprocessing

Raw probe intensities were normalized and summarized to expression levels using the FARMS algorithm which utilizes a factor analysis approach (Hochreiter et al., 2006). A rigorous quality assessment confirmed a fairly good quality of the chips with exception of mouse HPC chips where Initial chip quality assessment revealed a strong batch effect and one bad chip (replicate no. 1 at time 1 hour). The bad chip was excluded and batch adjustment was performed to alleviate that effect on those chips via the "ComBat" method (Johnson et al., 2007). Affymetrix probe IDs were mapped to Entrez gene IDs using the Bioconductor annotation packages "mogene10sttranscriptcluster.db" in mouse chips and "hugene10sttranscriptcluster.db" in human chips (Arthur, 2010b,a). Details about microarray chips technology and normalization are in section 2.1.

4.2.2 Differential Gene Expression

Time Point-Specific Analyses

Differential gene expression analyses via "limma" Linear Models for Microarray Data (Smyth, 2005) using empirical Bayes method (Casella, 1985) was performed by comparing samples at each time point after TGF-β stimulation to the unstimulated cells at time point 0. Statistical dependencies of samples between time points and replicates were considered via a factorial design matrix in "limma" using a "time" and a "replicate" factor, and contrasts are considered for interaction effects. Corrections for multiple testing was done using the Benjamini & Hochberg's method (Benjamini and Hochberg, 1995). Significant differentially expressed genes are considered those with $FDR_{BH} \leq 0.01$ and absolute logarithm of fold change value $|\log_2(FC)| \geq 1$. Details about differential expression analysis methods and the "limma" method can be found in section 3.2 and subsection 3.2.3.

Analysis of Whole Time-Courses

The small number of replicates in the experiments limits the power of statistical testing procedures for assessing differential gene expression at individual time points. Furthermore, the number of measured time points is not the same for each cell type, which complicates any further meta-analysis. Therefore, the "betr" method to analyze whole time series at once is employed (Aryee et al., 2009). The algorithm of this method uses a random-effects model together with the empirical Bayes method to estimate probabilities for differential expression of whole time courses. Genes were considered to be significant at a probability cutoff of $P \geq 0.99$ for the whole time-course analysis and absolute logarithm of fold change value $|\log_2(FC)| \geq 1$. Since "betr" requires the same number of replicates per time point and one chip in mouse HPC had to be omitted due to low quality (see Normalization & Preprocessing) unfortunately in this particular cell line the time point 1 h had to completely be excluded from the time-course analysis. The "betr" method is explained in subsection 3.3.1.

4.2.3 Cluster Analyses

Clustering of gene expression time series was done via the MFDA method proposed in (Ma et al., 2006).It is worth mentioning that the MFDA is not applied on raw gene expression data here, but on log fold-changes relative to the reference time point 0 hours. The reason was that the genes should not be grouped merely on the

basis of their absolute expression values, they should be rather grouped on the basis of similar responses to the perturbation stimulus. RCEM with 5 Markov chains, rejection threshold of 0.5 and iteration maximum limit of 100 are used. Details about MFDA clustering method can be found in subsection 3.4.3.

4.2.4 Pathways and Gene Ontology Analyses

Analyses of pathways in KEGG® (Kanehisa and Goto, 2000) and biological processes in Gene Ontology project (GO®) (Ashburner et al., 2000) were performed as follows: The $-\log 2P\text{-}value$ of all genes in the individual time point analysis and $-\log 21 - probabilities$ of all genes in the whole time-course analysis, respectively, were taken as a ranking score for each transcript. Gene sets of KEGG® pathways and GO® terms were then tested for their association with these ranking scores via a univariate logistic regression based test (see subsection 3.5.2 and Sartor et al. (2009); Montaner and Dopazo (2010)). Unlike enrichment analysis, this kind of association analysis considers all genes in the chip. Thus, overcoming enrichment analysis drawbacks, beside it is more suitable for comparisons were the numbers of DE genes in a cell type are different than in the others which was the case here. A hypergeometric test based over-representation analysis have been used for the gene cluster groups (subsection 3.5.1). Details of the gene set enrichment analysis methods are explained in section 3.5.

Resulting p-values of KEGG® pathways and GO® terms were adjusted according to Benjamini & Yekutieli's false discovery rate control under dependency (Benjamini and Yekutieli, 2001), and significant KEGG® pathways and GO® terms reported at a cutoff value of $FDR_{BY} \leq 0.05$. Details about multiple test correction can be found in subsection 3.2.4

4.2.5 Transcription Factor Binding-Sites Analyses

Analyses of transcription factor binding-sites (TFBSs) are performed using the *de novo* sequence motif detection method XXmotifs (Luehr et al., 2012). Identified sequence motifs were then aligned to known TRANSFAC TFBS via STAMP (Mahony and Benos, 2007) and the top match is considered. The XXmotif method uses BLAST (Altschul et al., 1990) all-against-all comparisons to mask regions of local homology in order to avoid false positives. The method then performs an enrichment analysis after transforming the found patterns to position weight matrices (PWMs). The STAMP method utilizes a global or un-gapped local alignment to detect DNA motifs similarities to defined PWMs. Furthermore, it considers familial binding profiles, thus improving transcription factors (TF) classification accuracy. TFBSs

analysis was done using these methods in each cell type for those genes, which according to the time-course analysis showed a probability of $p \geq 0.99$ for differential expression. Promoter sequences of the genes under consideration (2Kbp upstream of transcription start site[1]) were obtained from the Ensembl database (Flicek et al., 2011) via "biomaRt" (Durinck et al., 2005, 2009). Only the top matching motif for each TRANSFAC TFBS was considered and significant TFBSs were reported at threshold of $E\text{-}value \leq 0.001$.

Mapping of TFBS to individual transcription factors was performed via manual inspection of TRANSFAC PWMs. Proteins which had been used to construct each of the PWMs are obtained, and their names are mapped to Entrez gene IDs with the help of the commercial software GeneGo Metacore® [2].

As a consequence, the differentially expressed genes Foxp2 and FoxP1 are found for the transcription factors TFBS FOXP1 (in MPP, CDP mouse cells). For the TFBS FOX the human gene FAU was identified (human HPC). Egr1/EGR1 (mouse HPC, human MSC) and EGR2 (human MSC) are found For KROX. For TEF Klf3/KLF3 are identified (MPPs, CDPs, human HPC), TRIM37 and USP7 (human HPC, MSC).

4.2.6 Identification of Homologous Genes

Human homologs of mouse genes were identified via the KEGG® Sequence Similarity Data Base (SSDB), which contains local alignments of amino acid sequences for protein coding genes from different species. Two genes are considered to be homologs, if the alignment E-value was below $1e-30$ and $bit\text{-}score \geq 112$. In case of more than one homologous gene, all are considered.

4.2.7 Network Analyses

Information about protein-protein interactions was collected separately for human and mouse from the BioGRID database version 3.2.109 (Stark et al., 2006). Correspondingly, a network comprising 16,011 nodes and 140,471 physical interactions was constructed for human. For mouse the network consisted of 6,233 nodes and 16,100 physical interactions. Nodes in these networks were weighted by the average probability (mean over all cell types from the same organism) for differential time course expression according to the "betr" model (subsection 3.3.1). A "distance" for

[1] Promoter and enhancer regions which act as target sequence for DNA-binding proteins (TFBS) are up-stream sequences of transcription start site in a gene.

[2] `https://portal.genego.com`

each edge was then calculated as 2 minus the sum of its incident nodes' weights. Hence, the smaller the distance the higher the weight of its incident nodes. The Dijkstra's algorithm is used to search for minimum distance (i.e. maximum node weight) path connecting TGFB1 with each of SKIL, SMAD7, EGR1, PPARG and all genes annotated to *glutathione metabolism*, *purine metabolism*, *oxidation-reduction process*, *innate immune response*, *negative regulation of apoptotic process*, *angiogenesis*, *positive regulation of cell proliferation and positive regulation of cell migration*. For each of the last mentioned terms those genes are kept as representatives which showed the minimum distance to *TGFB1*. If there were several paths of the same minimum distance, all of them were considered. In the network for mouse *Tgfb1* was not identified and hence the analysis is started with *Tgfbr1* instead.

4.2.8 Functional Similarity Maps

We developed a technique for computing and visualizing similarities of different cell types with respect to different levels of functional annotation, such as GO®, KEGG® and predicted TFBS. The idea is to associate each cell type with a vector in which each position corresponds to a GO® term, KEGG® pathway or TFBS being significant in at least one cell type. In case of GO® and KEGG® annotations the vectors contain $-\log_2(FDR)$ values (logarithm to base 2), and in case of TFBS they contain $-\log_2(E\text{-}Values)$. In order to compare whole cell types with respect to these vectors their cosine similarities using a dot-product and magnitude are computed:

$$Similarity(i,j) = \frac{(x^i \cdot x^j)}{\parallel x^i \parallel \parallel x^j \parallel}, \tag{4.1}$$

where x^i and x^j are the vectors associated to cell types i and j. Calculation of this similarity for each pair i, j hence yields a similarity matrix (one for GO®, one for KEGG® and one for TFBS). Plotting these functional similarity matrices yields what we call *functional similarity maps*. These visualize proximities of cell types with respect to associated GO®, KEGG® and TFBS annotation. Here, functional similarity maps are plotted separately for GO®, KEGG® and TFBS, but potentially also weighted combinations would be possible (Figure 4.12).

4.3 Results

Time Series Transcriptome Measurements

All cell types were treated with TGF-β in three biological replicates. TGF-β treatment concentrations were optimized in each cell type to show a maximal effect. Extracted RNA samples were hybridized to microarrays (Affymetrix Gene 1.0 ST) for genome-wide transcriptome analysis. Mouse progenitor cells and HepG2 cells were measured at 6 successive time points, mouse primary HPC cells at 5, and human MSCs at 4 different time points. Table A.1 gives an overview of the experiments and the measured time-points. Details about the cell types, cell cultures, stimulation, RNA-isolation and array hybridization in our experiments can be found in the appendices section A.1.

4.3.1 Differential Gene Expression

Transcriptional Response is Highly Tissue-Specific on Gene Level

The "betr" method (Aryee et al., 2009) is employed to quantify the probability of differential expression of genes in whole time-courses (see section 3.3). Using this approach it is possible to assess differential gene expression for each gene in each cell type in a comparable manner. A gene is considered to have differential time-course expression (DE), if it had a probability of $p \geq 0.99$ and is at least two-fold up- or down-regulated at one time point minimum (Figure 4.2 a & b, Table Table 4.1, details in appendices A Excel file 8). The strongest stimulatory effect of TGF-β was observed in CDP cells (614 genes). Eight out of these genes in CDP are already known to play a role in the TGF-β pathway (Tgfb3, Smad7, Thbs1, Tgfbr1, Smurf1, Smad3, Smad6, Tgfbr2). In mouse HPC a significantly lower number of DE genes were found compared to other cell types.

Set comparisons of DE genes across cell types are conducted. It is worth mentioning in this context that comparisons between mouse and human genes were done on the basis of homologous genes (see Material and Methods). Not surprisingly, the found overlap was particularly high among mouse hematopoietic progenitor cells (MPP and CDP). These were 173 genes, which equals a harmonic mean of above 41% of DE genes in both cell types (Figure 4.2 a). Only two of these genes, namely Smad7 and Tgfbr1 are known to play a role in the TGF-β pathway. Three genes (Lox, Pmepa1, Skil) are found to be DE in all mouse cells (CDP, MPP and HPC). Pmepa1 (Prostate Transmembrane Protein) is known to interact with Smad and suppress the TGF-β pathway (Xu et al., 2000, 2003).

Table 4.1: Numbers of differentially expressed genes. Genes with probability $p \geq 0.99$ and $|logFC| \geq 1$ in each cell type and condition according to the time-course analysis. Comparisons between mouse and human are based on homologous genes (see Material and Methods). The diagonal in the table indicates the number of DE genes in each cell type. The other numbers are pair-wise overlaps.

Organism		Mouse			Human	
		MPP	CDP	HPC	HPC	MSC
Mouse	MPP	228	173	2	9	10
	CDP	173	614	4	16	19
	HPC	2	4	15	1	3
Human	HPC	9	16	1	226	18
	MSC	10	19	3	18	208

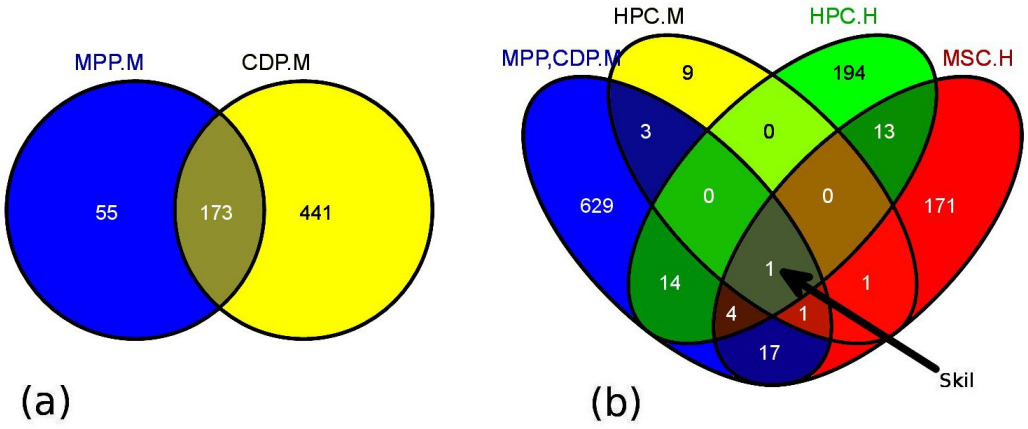

Figure 4.2: (a) Venn diagram of DE genes (probability ≥ 0.99 and $|\log_2(FC)| \geq 1$) in mouse MPP and mouse CDP. (b) Union of DE genes in MPP and CDP compared to all other cell types (homolog genes, see subsection 4.2.6).

Only the protein-coding gene Skil (Ski-like-oncogene) that encodes a protein in the SMAD-pathway (Cohen et al., 1999; Nomura et al., 1989) was found to have a DE time-course in all cell types. In addition, the gene Smad7 was commonly found in all cell types except mouse HPC cells. 18 genes including ROR1, C10orf10, SMAD7, FSTL3, GADD45B, JUNB, ZFP36, OLFM2, SPTLC3, ID1, LMCD1, SLC38A3, GXYLT2, SKIL, HES1, RASGEF1B, CITED2 and PDGFA were DE in all human cells (MSC, HepG2). The heatmaps in Figure 4.3 visualize patterns of temporal behavior for particular groups of genes. Here again, similarity in gene expressions between mouse dendritic cells is evidenced.

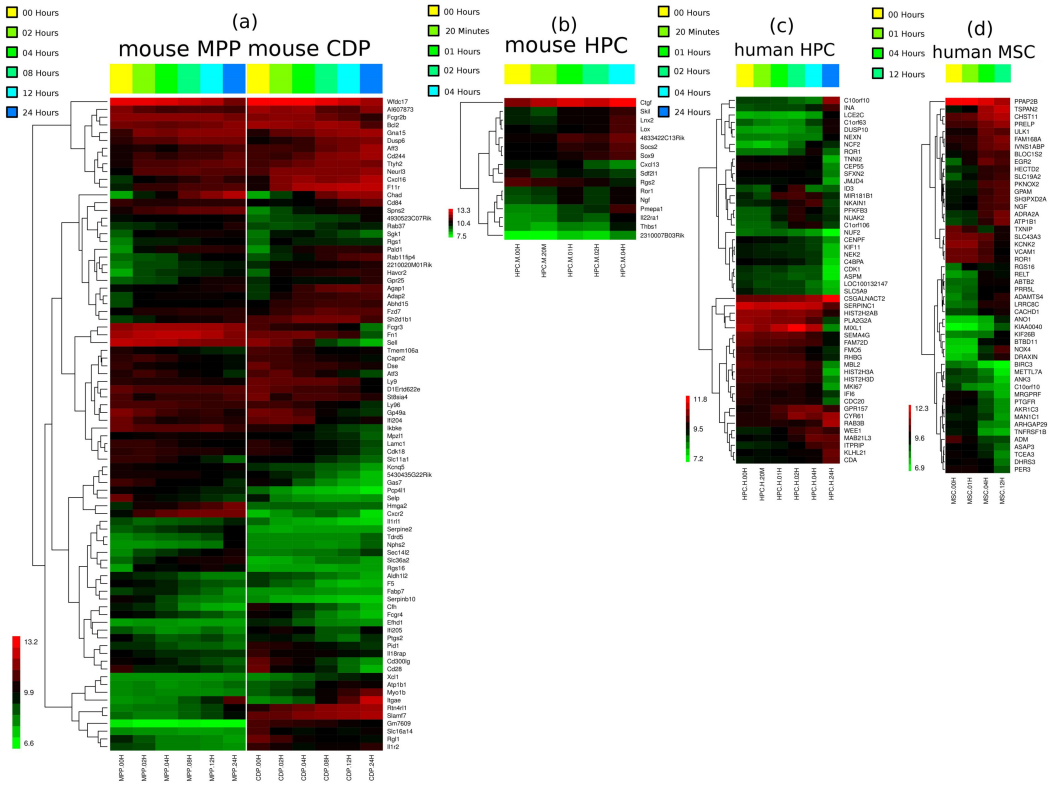

Figure 4.3: Heatmaps depicting mean logarithm of fold changes of top DE genes at different time points. (a) Mouse MPP and CDP (together 84 genes), (b) mouse HPC (16 DE genes), (c), (d) top 50 DE genes in HepG2 (HPC) and human MSCs, respectively

These findings on one hand stress the similarity of the transcriptional response in MPP and CDP, which is not very surprising given the fact that these cells were both derived from bone marrow. On the other hand they highlight that TGF-β treatment affects by far not only genes within the canonical TGF-β pathway, but leads to a large number of diverse secondary downstream effects, which are only partially overlapping across different cell types. In other words there is a high tissue specificity of the transcriptional TGF-β perturbation response on the level of individual genes.

4.3.2 TGF-β1 Pathway Genes React Time-Dependant and Tissue-Specific

Genes which are known to play a role in the TGF-β pathway, such as Bmp(s), Smad(s) and Id(s), are closely investigated. In Figure 4.4 the log fold changes (logarithm to the base 2) of 17 genes involved in the TGF-β pathway, which are DE in at least one cell type, are depicted. It can be noticed that almost all genes show time-dependant transcriptional effects. These effects are distinct between early and later time points, with moderate activities until 4h and mostly higher activities at late times.

It can also be noticed that cells of similar origin are more alike. For example, Bmp2, Bmp4, Bmp6, Cdkn2b and Comp are dys-regulated (i.e. significantly differ from 0 level according to "betr") only in human and not in mouse tissues. Fs1 is similar to these genes, but also shows activity in mouse HPC. Id1 in human cells is up-regulated at earlier time points and a down-regulated after 4h. Inhba shows activity only in MSC cells where its expression after 1 hour constantly increases. Smad3, Smad6 and Smad7 reveal similar time courses in mouse MPP and CDP cells and in human MSCs. Smad3 is increasingly down-regulated over time and the other two genes are always up-regulated. Smurf1 is always over-expressed and shows a curve that is opposite to Smad3, Smad6 and Smad7. Tgfb3 is over-expressed at later time points in MPP and CDP cells and shows almost no activity in the other cell types. Thbs1 is highly active in all cell types. However, while it is underrepresented in MPP and CDP, it shows elevated expression in mouse and human HPC. Tgfbr1 and Tgfbr2 behave similar, in particular in mouse progenitor cells, where Tgfbr2 is less up-regulated than Tgfbr1.

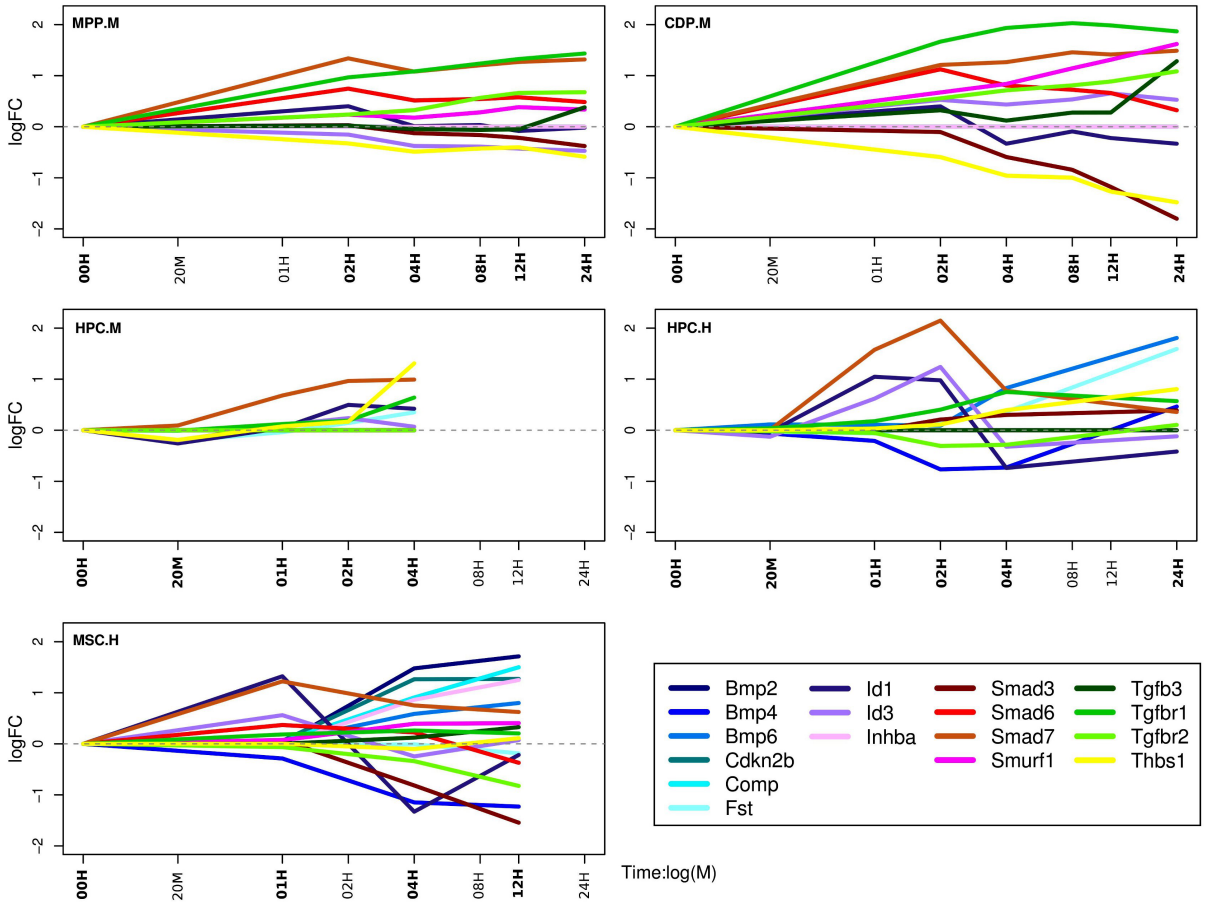

Figure 4.4: Time-course expressions patterns of differentially expressed TGF-β genes. The plots depict the logarithm of fold-changes of 17 genes, which are DE in at least one cell type and are known to play a role in the TGF-β pathway (according to KEGG® annotation).

4.3.3 Time-Point Specific Analyses Confirms Highly Tissue-Specific Expression Changes on Gene Expression Level

In order to cross-validate the previous analysis, which considers time series as a whole, also time-point specific analyses of differential gene expression using linear models for microarray data (limma) are conducted. For this purpose gene expressions at 4 hours after stimulation are compared to the initial expression at time point 0 hours. The time period of 4 hours was chosen because at least short-time relevant effects are expected in all cell types after this period.

In the context of time point analysis of transcriptional effects a gene is considered to be differentially expressed (DE) if $FDR_{BH} \leq 0.01$ and the absolute fold changes was $\log_2(FC) \geq 1$. The overlap analysis of DE genes at 4h agrees with the time-course analysis. There are no or very few genes in common between the different cell types except in the case of mouse dendritic cells (Figure 4.5 A). Moreover, the direction of regulation (up or down) differs between cell types (details in appendices A Excel file 9).

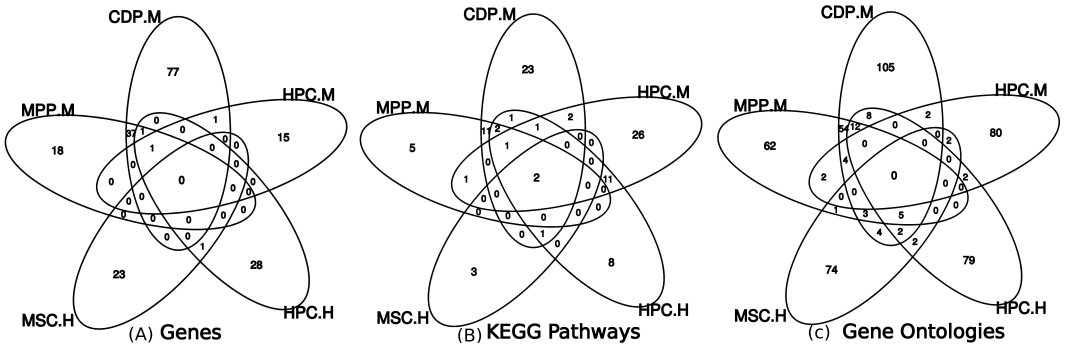

Figure 4.5: Venn Diagrams of differentially expressed genes ($FDR_{BH} \leq 0.01$ and $|\log_2(FC)| \geq 1$) and associated KEGG® pathways and GO® terms ($FDR_{BY} \leq 0.05$) in mouse MPP & CDP and in human HPC & HPC cell types at 4 hours (taking homologous genes between human and mouse into account, see subsection 4.2.6, details in appendices A excel files).

The heatmap in Figure 4.6 depicts the log fold changes of all genes, which are DE in at least one cell type. The plot indicates two gene sets, which clearly show a similar behavior in mouse MPP and CDP cell types. The first set contains 36 genes that are over-expressed. The other set (42 genes) is under-expressed. Interestingly, the 36 genes being up-regulated in MPP and CDP cells are not regulated by TGF-β1 in other cell types. Although not DE genes in every cell type, the genes Smad7, Pmepa1 (beside the gene Skil) seem to be up-regulated in all the cells. The rest of the genes are regulated in a rather cell-type specific manner.

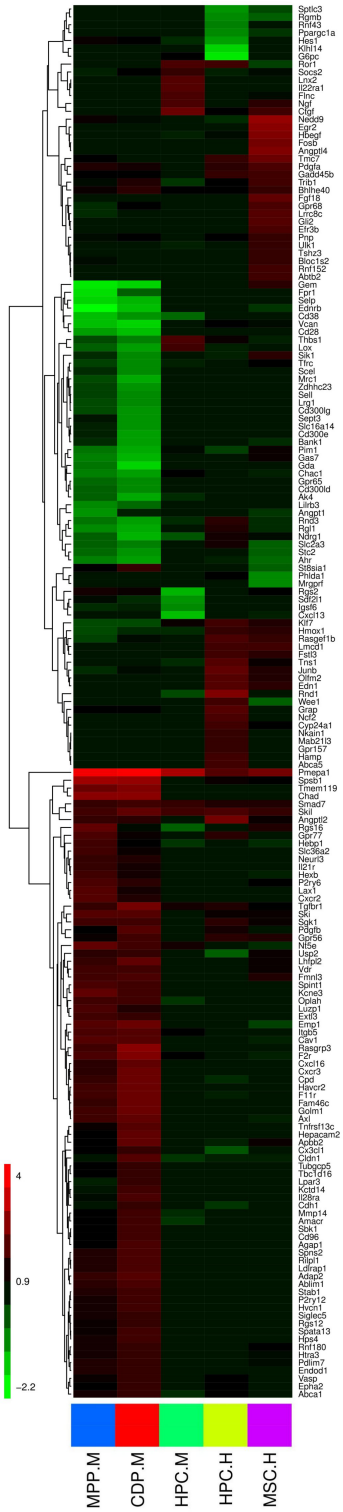

Figure 4.6: Heatmap depicting logarithm of fold changes of all genes that are differentially expressed in at least one cell type 4 hours after stimulation.

4.3.4 Cluster Analyses Revealed Functionally Similar Gene Groups in Different Cell Types

Time series cluster analyses are conducted in order to find groups of DE genes showing similar expression changes over time (observed as within cell-type temporal behavior shown in Figure 4.3, between cell-types similarities shown in Figure 4.6). The cluster analyses yielded 12 clusters in MPP and mouse HPC, 20 in CDP and 11 in human MSC (Table 4.2). Genes contained in individual clusters can be found in details in appendices A Excel file 14). Figure 4.7 depicts the mean curves for each of these clusters in each cell type. Functional similarity of genes across different clusters is investigated. For this purpose the R-Package "GOSemSim" (Yu et al., 2010) utilizing the semantic similarity measure proposed by Wang et al. (Wang et al., 2007) was employed. Semantic similarities are means to compare GO® annotations of gene pairs in a quantitative manner, for example on the basis of the information content of GO® terms. Pesquita et al. (2009) provide an overview in this subject.

Table 4.2: Clusters overview. Differentially expressed genes in each cell type and number of resulting clusters (details in appendices A Excel file 14).

Organism	Mouse			Human	
	MPP	CDP	HPC	HPC	MSC
Differentially expressed genes	230	631	15	232	208
Number of Clusters	12	20	1	12	11

A heatmap depicting these GO® semantic similarities suggested a high functional similarity of genes in several clusters from different cell types (Figure 4.8 details in appendices A Excel file 15 & Excel file 16). In particular cluster B (MPP), and cluster B (CDP) are highly similar to each other ($semanticsimilarity \geq 0.7$). Time-course logarithm of fold changes of the corresponding genes are shown in (Figure 4.9 top). As can be noticed the clusters are of different size, but have several genes in common (13 genes). Functional analysis revealed that genes in these clusters are enriched for *cell adhesion molecules (CAMs)*, *valine, leucine and isoleucine biosynthesis*, *Pantothenate and CoA biosynthesis and regulation of cellular extravasation* biological processes gene ontology terms. Enrichment analysis was conducted here via the R-package GOstats (Falcon and Gentleman, 2007), which employs a hypergeometric test taking into account the dependency structure among GO® terms.

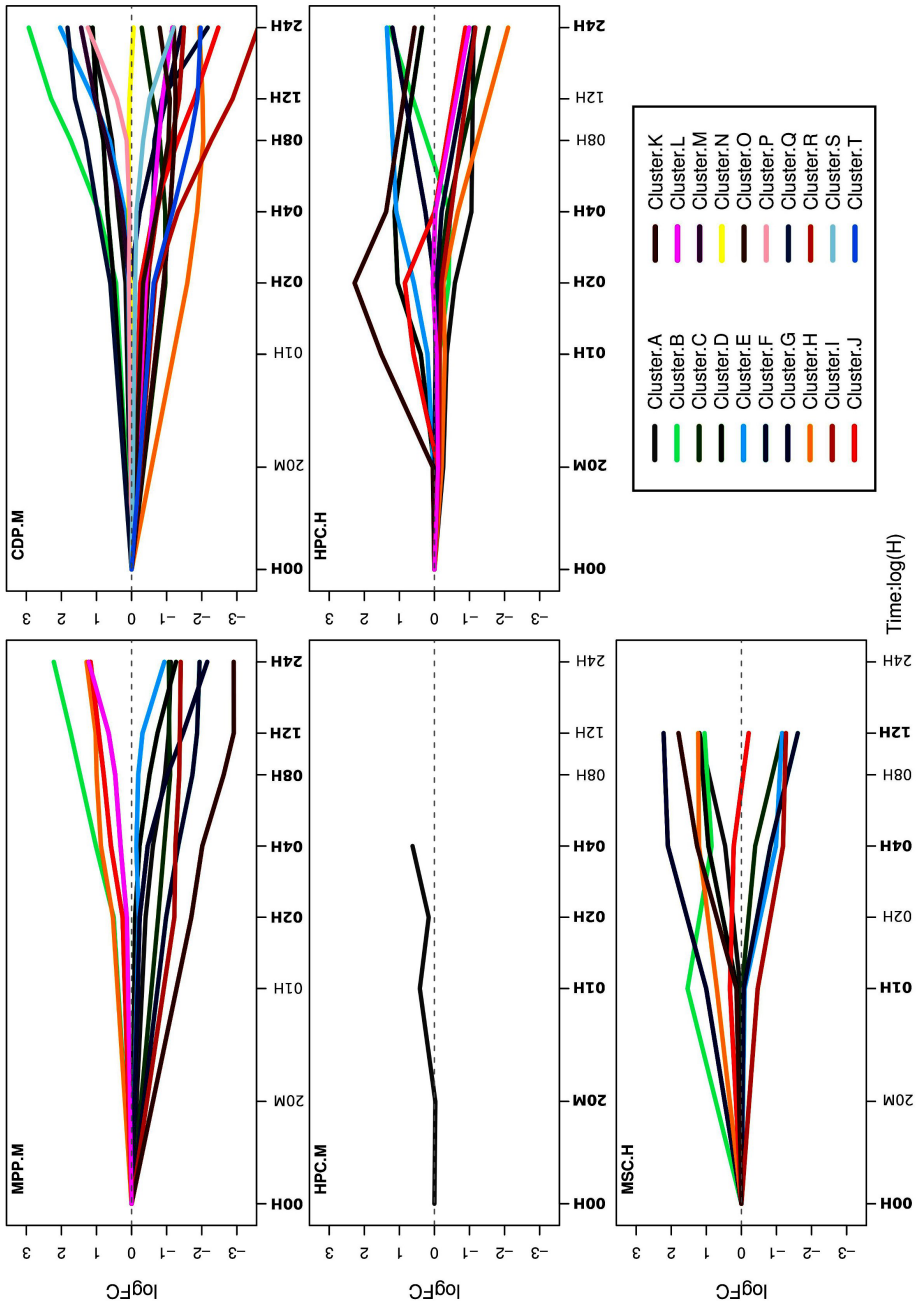

Figure 4.7: Mean-curves of logarithm of fold changes of gene groups in the clusters detected in the different cell types. For mouse HPC no clusters could be identified and hence all DE genes treated as one group.

The second group of functionally similar clusters (Figure 4.9, bottom) contains cluster K (human HPC) and cluster B (MSCs). Genes in these clusters play (among others) a role in TGF-β and Notch signaling pathways (details in appendices A; Excel file 15 & Excel file 16). Taken together the cluster analyses showed that despite evident differences on the level of individual genes, functionally similar clusters of genes can be identified across cell types.

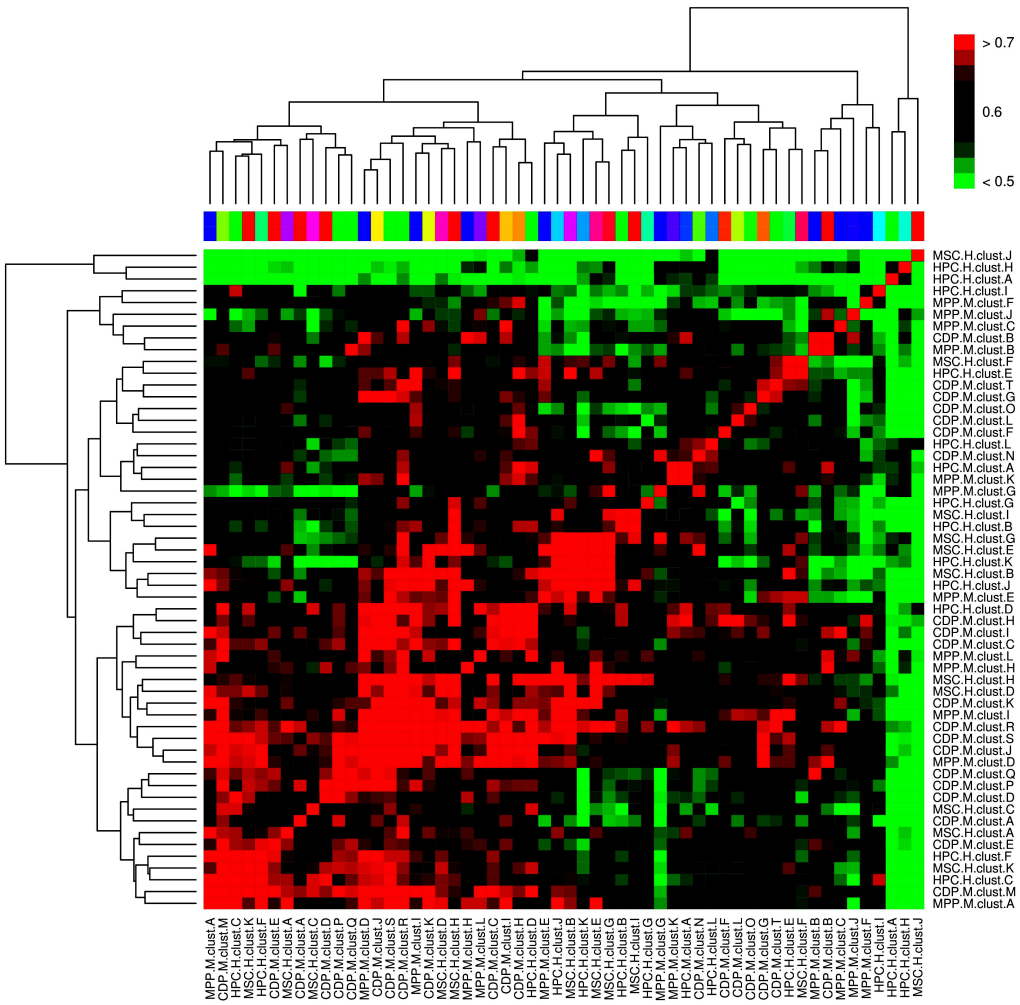

Figure 4.8: GO® semantic similarity heatmap for all the resulting clusters in all cell types. The color code indicates the degree of functional similarity between clusters according to their GO® annotation. GO® semantic similarities were computed via the Bioconductor R-package "GOSemSim" using the similarity measure by Wang et al. (2007).

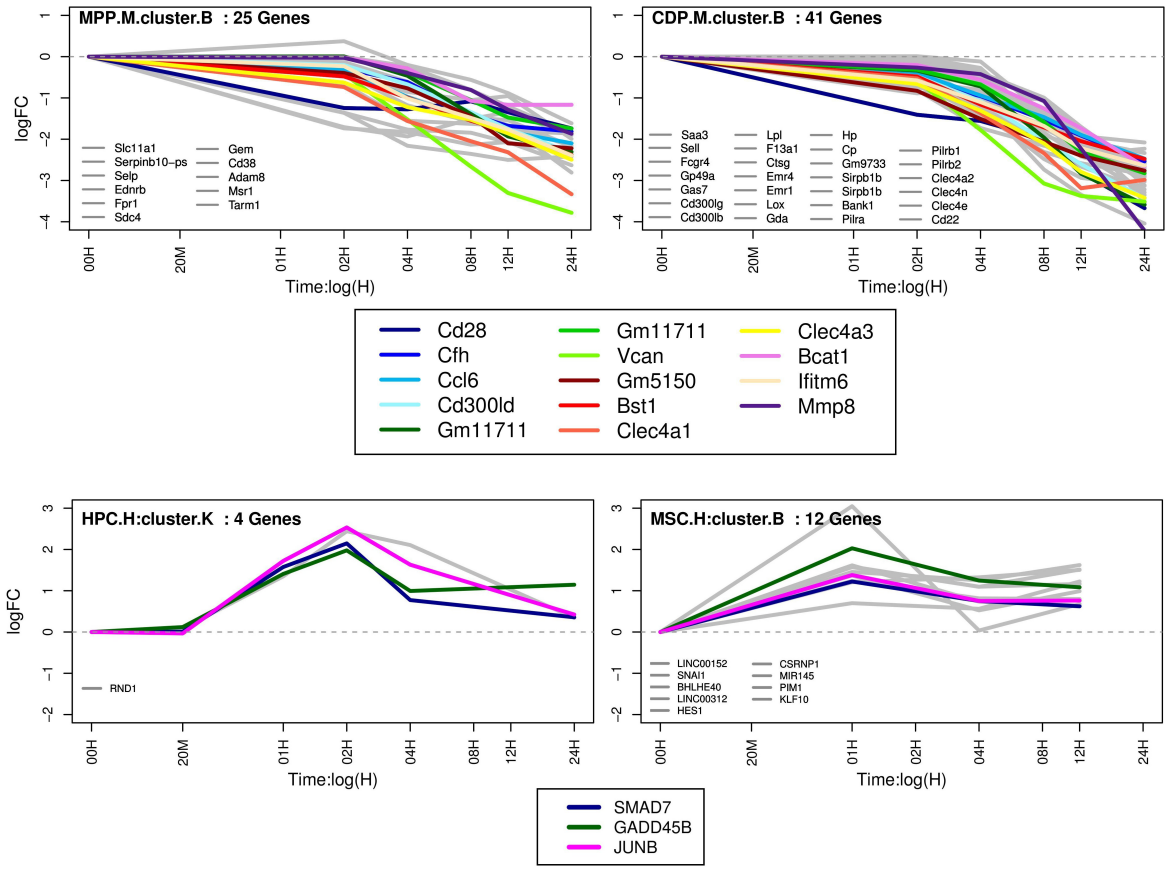

Figure 4.9: Logarithm of fold changes of two groups of functionally similar clusters detected in different cell types. Genes appearing in more than one cluster depicted in colors, gray curves are cluster-specific genes. Upper group: two similar clusters MPP and CDP. Lower group: two similar clusters in human HPC and MSC.

4.3.5 Enrichment Analyses Reveal Commonly Affected Biological Processes, Pathways & TFBS

Commonly Found Pathways and Gene Ontology Terms

The previous findings motivates investigating whether there are common functional patterns across all cell types. For this purpose GO® terms and KEGG® pathways are scanned for significant association with differential time course gene expression in each cell type (overview in Table 4.3 & Table 4.4, details in appendices A Excel file 10 & Excel file 12). The analysis brought up 6 KEGG® pathways and 11 GO terms, which were significantly associated to all cell types ($FDR_{BY} < 0.05$, Figure 4.10). The 6 KEGG® pathways associated to all cell types were: *Metabolic pathways*, *Glutathione metabolism*, *Lysosome*, *Purine metabolism*, *Peroxisome* and *PPAR signaling pathway*. The 11 GO terms associated to all cells were: *oxidation-reduction process*, *innate immune response*, *positive regulation of transcription from RNA polymerase II promoter*, *negative regulation of apoptotic process*, *angiogenesis, lipid metabolic process*, *positive regulation of cell proliferation*, *positive regulation of cell migration*, *proteolysis* and *positive regulation of transcription DNA-dependent and response to drug*. The role of TGF-β in *apoptosis*, *cell proliferation* as well as *immune response* is well known. Moreover, an effect of TGF-β perturbation on *PPAR signaling* has been described in skin fibroblasts (Ghosh et al., 2004).

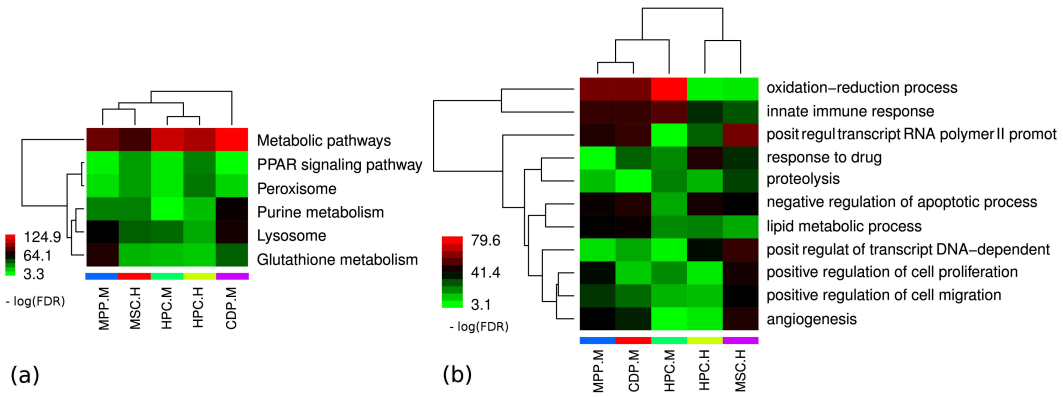

Figure 4.10: Clustered heatmaps of (a) the 6 common KEGG® pathways and (b) 11 GO® terms in different cell types. The color code indicates the degree of association ($-\log_2(FDR)$) of a KEGG® pathway and GO® term to each cell type, respectively.

In (Liu and Gaston Pravia, 2010) the authors describe TGF-β mediated oxidative stress and decreased glutathione concentration in fibrosis models. Finally, there is evidence that TGF-β has an effect on angiogensis and cell migration (Yang and Moses, 1990). Hence, our findings largely fit to the current biological knowledge about TGF-β

Table 4.3: KEGG® pathway Enrichment overview. Numbers of enriched KEGG® pathways in each cell type and condition at $FDR_{BY} \leq 0.05$ (diagonal) according to time-course analysis. The other numbers are pair-wise overlaps, these are all significant overlaps between the corresponding two cell types, according to a hyper-geometric test with $P\text{-}value \leq 0.05$ (details in appendices A Excel file 10).

Organism		Mouse			Human		
		MPP	CDP	HPC	HPC	MSC	CRL
Mouse	MPP	98	85	22	31	54	57
	CDP	85	116	24	36	58	68
	HPC	22	24	47	16	18	26
Human	HPC	31	36	16	58	32	37
	MSC	54	58	18	32	84	64
	CRL	57	68	26	37	64	106

Table 4.4: GO® terms Enrichment overview. Numbers of enriched GO® terms in each cell type and condition at $FDR_{BY} \leq 0.05$ (diagonal) according to time-course analysis. The other numbers are pair-wise overlaps. **black** numbers are significant overlaps between the corresponding two cell types and red are insignificant overlaps, according to a hyper-geometric test with $P\text{-}value \leq 0.05$ (details in appendices A Excel file 12).

Organism		Mouse			Human		
		MPP	CDP	HPC	HPC	MSC	CRL
Mouse	MPP	255	166	36	48	68	87
	CDP	166	238	33	55	66	94
	HPC	36	33	139	30	22	35
Human	HPC	48	55	30	191	73	92
	MSC	68	66	22	73	252	127
	CRL	87	94	35	92	127	504

Conserved Role of EGR1/2 Transcription Factors

DE genes are analyzed with respect to overrepresented sequence motifs in their promoter regions with the XXmotif tool (Luehr et al., 2012). Significant motifs were then compared to known position weight matrices (TRANSFAC) of transcription factors (TFs) via STAMP (Mahony and Benos, 2007). The analysis in each cell type predicted between 11 and 21 regulating Transcription Factor Binding-Sites (TFBS) in the time-course analysis (Table 4.5, details in appendices A Excel file 17 & Excel file 18), except for mouse HPC, where no overrepresented TFBS could be detected. This may be attributed to the small number of 16 DE genes in this cell type. Overlaps were particularly high within mouse MPP and CDP and within human cells.

FOXP1, KROX, TEF, POU6F1, FOX and PITX binding-sites were commonly identified in mouse MPP and CDP. KROX, HFH4 and PAX4 were found in all human cells. FOX, FOXP1, KROX and TEF were found to be themselves representatives of DE genes. Figure 4.11 shows a network representation of all eight TFBS together with the set of DE genes containing respective binding-sites. The plot reveals a relative clear difference between mouse and human cells with the exception of the KROX TFBS, which appears in all four cell types. KROX represents EGR1 and EGR2.

Table 4.5: TFBS analysis overview. Total number of predicted transcription factor binding-sites (TFBS) in each cell type and condition (best match according to STAMP and $E\text{-}value \leq 1e-3$) according to time-course analysis. The other numbers are pair-wise overlaps (details in appendices A Excel file 17).

Organism		Mouse			Human	
		MPP	CDP	HPC	HPC	MSC
Mouse	MPP	11	6	0	5	2
	CDP	6	18	0	3	2
	HPC	0	0	0	0	0
Human	HPC	5	3	0	21	3
	MSC	2	2	0	3	13

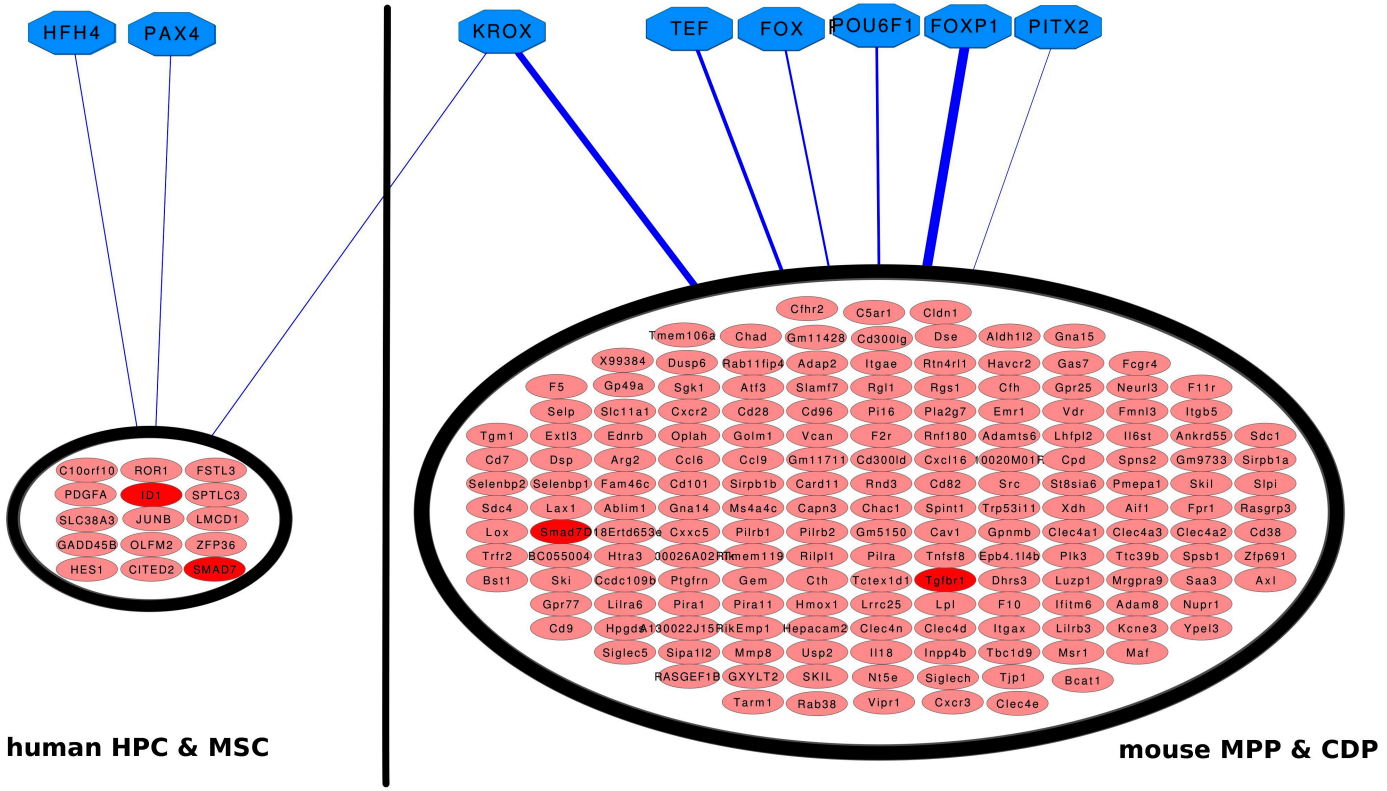

Figure 4.11: Network of eight overrepresented Transcription Factor Binding-Sites (TFBS) and differentially expressed genes containing these binding-sites. For the sake of better visualization only the set of genes being DE in both, HPC and MSC as well as both, MPP and CDP, are shown. Red genes are known to play role in the TGF-β pathway. The width of the blue lines is chosen to be proportional to the average $-\log_2(\textit{E-value})$, which resulted from the transcription factors analyses.

Functional Similarity Maps Display Conservation of Biological Processes and Pathways

In addition to the previously shown heatmaps, a technique called *functional similarity maps* was developed (see Material and Methods) that enables visualizing of overall similarities of cell types with respect to their GO® and KEGG® annotation as well as predicted TFBS from a more global perspective. Functional similarity map in Figure 4.12 demonstrates a comparably high similarity of all cell types with respect to their GO® annotation, a more cell type specific reaction with respect to KEGG® pathways and relatively high tissue specificity with respect to over-represented TFBS. An interesting observation is that mouse HPC are more similar to human HPC with respect to affected biological processes, but more different with respect to associated KEGG® pathways and TFBS. Moreover, on the level of GO® annotation all HPC are more similar to MPP and CDP than on the level of the other annotations.

Taken together these observations imply that TGF-β stimulation in all cell types yields the response of a transcriptional core program, which besides several metabolic pathways, appears to be related to the *insulin signaling* and *adipocytokine signaling* pathways as well as *immune response*, *apoptosis* and *cell proliferation* (see above).

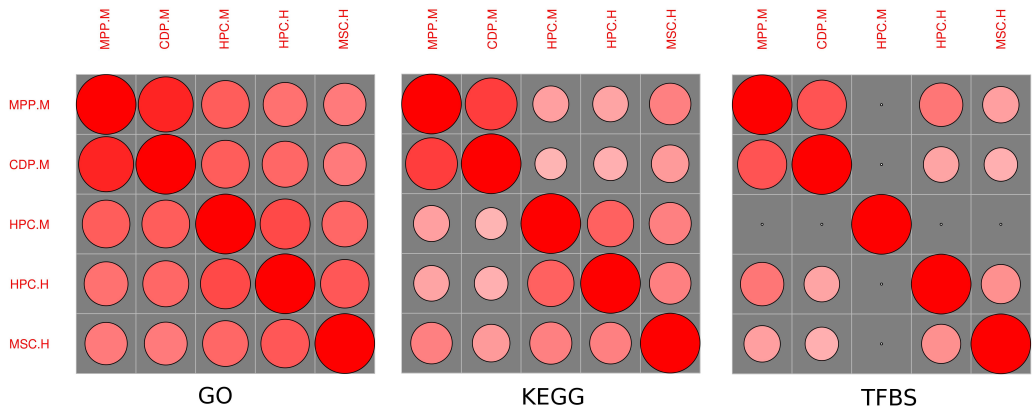

Figure 4.12: Functional similarity maps visualizing the proximity of different cell types based on significant GO® Terms (left), KEGG® Pathways (middle) and predicted Transcription Factor Binding-Sites (TFBS) (right). Circles size indicate the degree of similarity or dissimilarity (larger size = higher) and circles color indicate the direction of similarity or dissimilarity (red color mean similarity of 1 'exactly the same' and zero values are represented with small white circle).

The identified transcription factors seem to be rather species-specific. The exception is the transcription factor KROX, which represents genes Egr1 and Egr2. In human target genes of EGR1 are known to play a major role in cell differentiation and mitogenesis. Moreover, EGR1 is involved in several signaling pathways (*BMP signaling*, *cytokine mediated signaling*, *interleukin-1-mediated signaling*, *type-I interferon mediated signaling* – see GO® annotation).

Network Analysis Suggests Possible Signal Transduction Pathways in Mouse and Human

In order to better understand how TGF-β may influence the commonly identified transcription factor, biological processes and the PPAR-pathway a network analysis is conducted. Using protein-protein interaction information from the BioGRID database (Stark et al., 2006) a mouse and a human specific networks are constructed. These networks depict dys-regulated paths from TGF-β to SKIL, SMAD7, EGR1 as well as genes involved into *glutathione metabolism*, *purine metabolism*, *PPAR signaling*, *oxidation-reduction process*, *innate immune response*, *negative regulation of apoptotic process*, *angiogenesis* and *positive regulation of cell proliferation and positive regulation of cell migration* (Figure 4.13 and Figure 4.14; further details in Material and Methods part).

The network analysis suggests pathways, by which TGF-β stimulation is possibly propagated via protein-protein interactions to the commonly identified biological processes. Due to the organism specificity of interactome information these pathways show certain differences: Far less protein-protein interactions are known in mouse than in human. In human, for example, *negative regulation of apoptosis* might be mediated via SMAD3 and SOX9 (Yanagisawa et al., 1998). In contrast, the GO® and network analysis in mouse suggests a direct role of TGFBR1.

4.3.6 Enrichment of Biological Processes, Pathways and Transcription Factor Binding-Sites (TFBS) is Reproducible on an Independent Dataset

In order to validate the central finding from the data, namely the existence of commonly affected biological processes, pathways and transcription factors in all cell types, comparisons to results from an independent data are made. For this purpose dataset is downloaded. This dataset comprise gene expression data measured at 9 time points (0, 0.5, 1, 2, 4, 8, 16, 24, 72h) after TGF-β stimulation in human A549 lung adenocarcinoma cell-lines (CRL, GSE17708). The dataset was analyzed in the same manner as described for our data before. High fractions of the 11 GO® terms

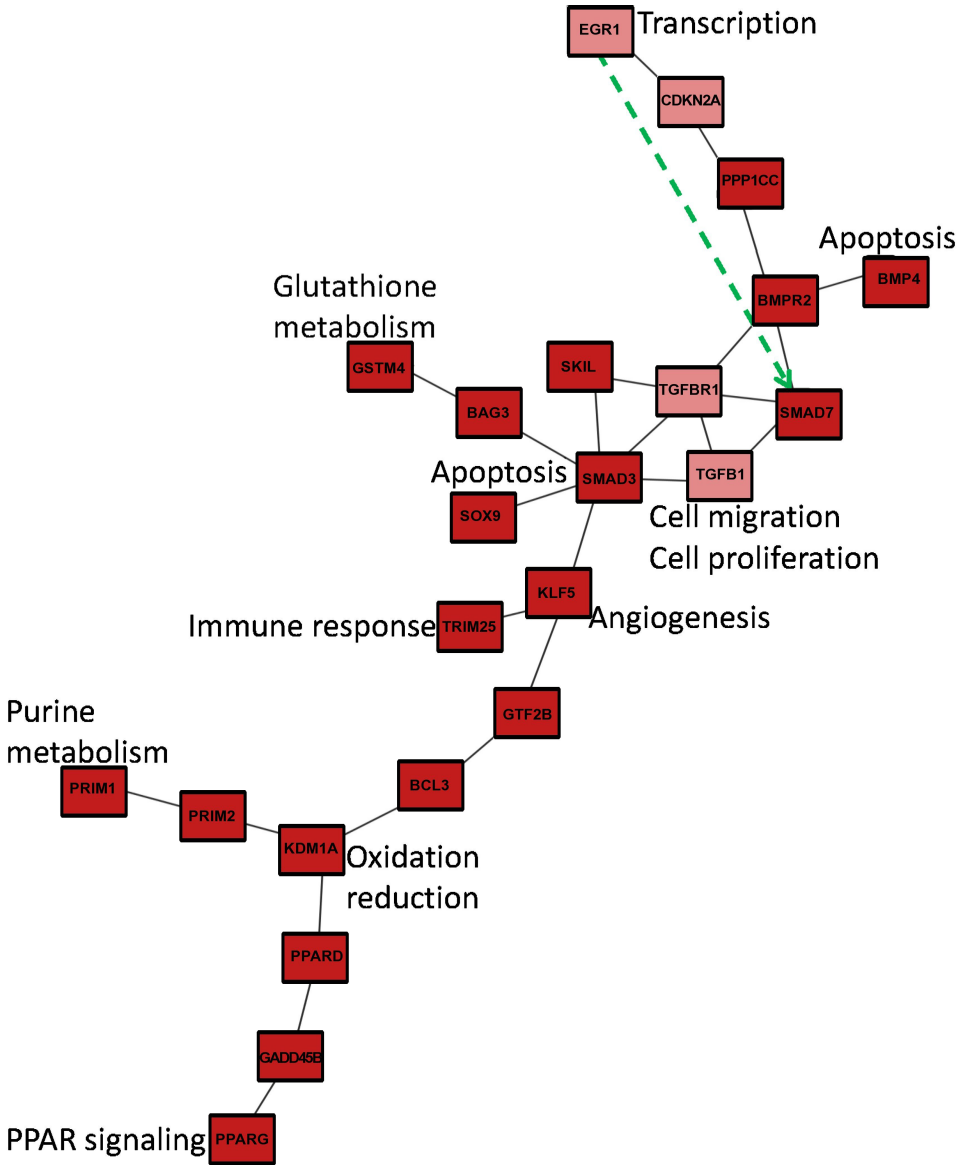

Figure 4.13: Human protein-protein interaction network connecting TGFB1, TGFBR1, SMAD7, SKIL, EGR1, PPARG with genes involved into commonly identified biological processes. The dashed green line indicates the putative transcriptional regulation of SMAD7 by transcription factor EGR1. The darker the red color of a node the higher the average probability for differential time course expression.

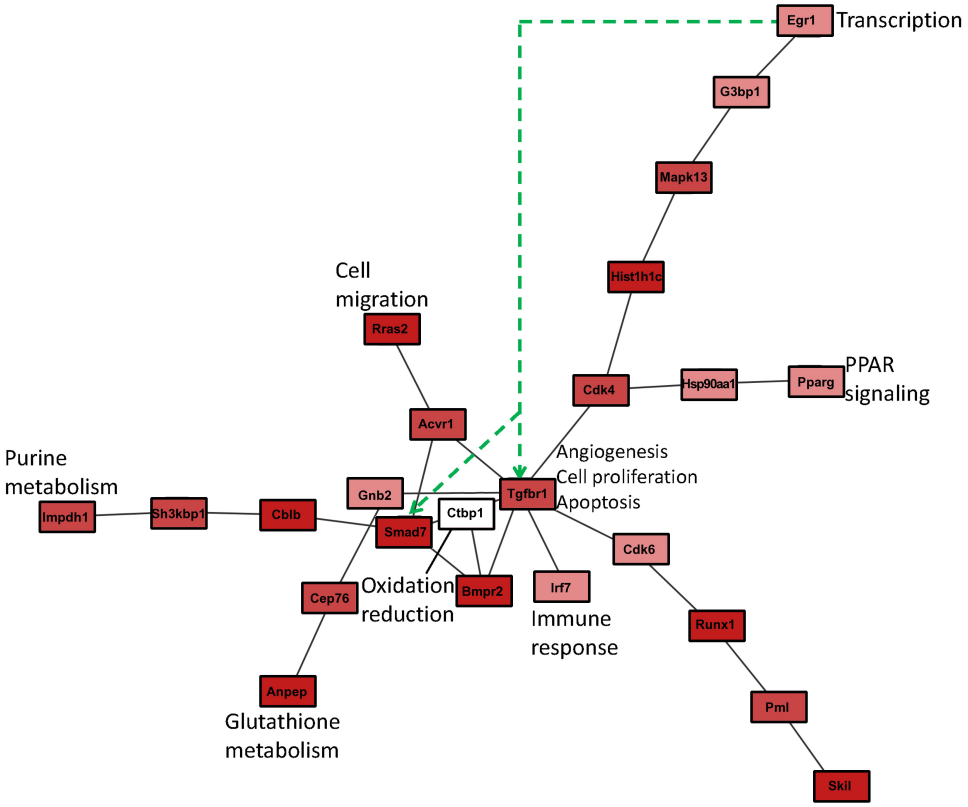

Figure 4.14: Murine protein-protein interaction network connecting Tgbfr1, Smad7, Skil, Egr1, Pparg with genes involved into commonly identified biological processes. The dashed green line indicates the putative transcriptional regulation of Smad7 and Tgfbr1 by transcription factor Egr1. The darker the red color of a node the higher the average probability for differential time course expression.

and 6 KEGG® pathways commonly identified in all of the cell types were also found in GSE17708 (Figure 4.15, details in appendices A Excel file 10).

Out of the KEGG® pathways and GO® terms associated to all of the human cells 70% and 74%, respectively could be reproduced on the independent dataset (Figure 4.15, details in appendices A Excel file 12). Notably, 11 (61%) out of the 18 genes which exhibiting differential time courses in both the human MSC and HPC cells were found also to have differential time-courses in GSE17708 cells, these were ROR1, SMAD7, FSTL3, GADD45B, JUNB, ZFP36, ID1, LMCD1, GXYLT2, SKIL and HES1. This corresponding fraction is significantly larger than expected by chance ($p < 1E\text{-}9$, hypergeometric test).

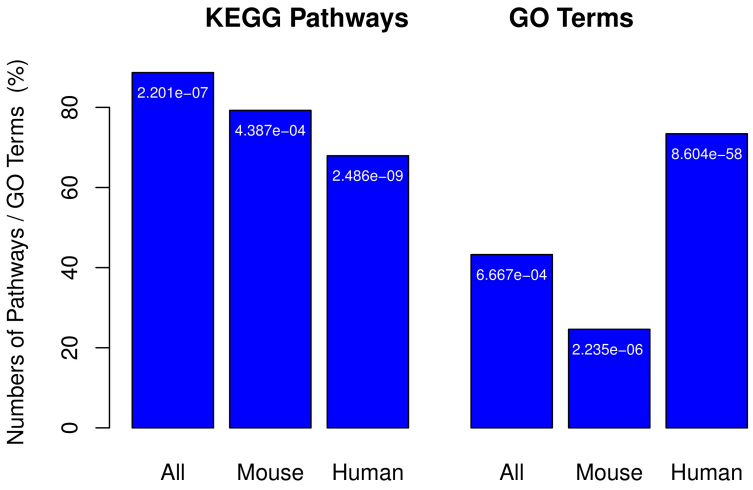

Figure 4.15: Percentages of KEGG® pathways (left) and GO® terms (right) enriched commonly in the cell types that could be reproducibly identified in GSE17708. The numbers in tip of the bars are the P-value for the null-hypothesis to see the corresponding overlap just by chance (hypergeometric test).

The KROX TFBS (corresponding to transcription factors EGR1 and EGR2), which was enriched in all of the cell types, was also found in GSE17708. Moreover, the other two TFBS that identified in the human cells (HFH4, PAX4) were also enriched in the A549 lung cancer cell line (details in appendices A Excel file 17). Taken together this analysis reveals a high reproducibility of the commonly identified biological processes, pathways as well as transcription factors.

4.4 Conclusions

An in-depth comparison of the dynamical TGF-β response profile on gene expression level across several cell types have been conducted in this work. Despite of a generally high degree of cell type specificity, there appears to be a common functional response, which is conserved across cell types and species (i.e. mouse and human). Our analysis suggests a common effect of TGF-β stimulation on apoptosis, cell proliferation, immune response, angiogenesis, cell migration, PPAR signaling, oxidative stress as well as purine and glutathione metabolism. Network analysis gives hints to possible pathways, by which these effects could be mediated.

On the level of individual genes the SKI-like oncogene and Smad7 were differentially expressed in most (Smad7) or all (SKI-like oncogene) cell types and thus appear to play a major role. Smad7 is involved into the canonical TGF-β pathway (Kanehisa and Goto, 2000). It is a general antagonist of the TGF-β family (for review see Yan et al. (2009)). The SKI-like oncogene is a direct target gene of Smad2, which regulates its transcription (Lee et al., 2011). It plays a role in cell growth and differentiation. Notably, a high fraction of the biological processes, pathways and TFBS that have been identified to be enriched in all cell types was found also in an independent dataset from a lung cancer cell line. This strengthens the confidence in our results.

In summary the findings indicate that despite a high variability of transcriptional response across cell types and organisms there appears to be a set of commonly affected processes and pathways. In addition, the TFBS analysis suggested a major role of the transcription factor EGR1 in the TGF-β response in human and mouse. Indeed the induction of EGR1 via TGF-β stimulation has been already reported earlier (Chen et al., 2006) and thus fits to the existing knowledge about TGF-β induced transcriptional response in other cell systems.

Previous studies of TGF-β stimulation were mainly limited to one specific cell type, e.g. fibroblasts (Clark et al., 1997; Petrov et al., 2002). In this work we went beyond this point and conducted perturbation experiments in different cell types under as much as possible comparable conditions. In consequence it was possible to compare transcriptional responses across cell types and organisms, which revealed common patterns. The identification of common and specific signal transduction pathways that are affected by TGF-β in human and mice will allow us to define potential therapeutic targets and will further enable us to characterize gene expression patterns and complex regulatory networks. In addition, future work using our and other transcriptome data can, for example, address the identification of TGF-β dependent mesenchymal or epithelial gene signatures or the definition of cell specific cancer signatures.

CHAPTER 5

INVASIVE AND NONINVASIVE MICRORNA BIOLOGICAL MARKERS IN CHRONIC AND ACUTE EPILEPSY

5.1 Introduction

Creating pathogenic status in living organisms through controlled moods of perturbations allow for standardized experimental designs in the study of disease models. Proper statistical methods for the analysis of high-throughput transcriptome measurements can help to better understand cellular malfunctions, identify biological disease markers and investigate therapeutic approaches.

In-vivo pro epileptic micoRNA markers are investigated in expression data in a number of epilepsy models in this work. Lower data dimensionality, compared to the typical gene expression microarray data, required adapting normalization and differential expression analysis methods. Experimental implications resulting in incomplete and censored high-throughput qPCR (HT-aPCR) data impairs the performance of analysis methods. A designated test procedure, that involve estimation of distribution parameters for censored data, is proposed to resolve this issue. The method showed higher detection power at lower false positive rates based on simulated data where differentially expressed features are known.

Epilepsy is a severe chronic neurological disorder that is usually manifested in repeated transient occurrence of signs and/or symptoms. The disease affects over 50 million people worldwide. The most common type of the disease is the *Temporal Lobe Epilepsy* (TLE) which is characterized by spontaneous recurrent seizures (Weiss et al.,

1986). A seizure (*Status Epileptics (SE)*) is a sudden unprovoked occurring event that usually takes few seconds to few minutes. Epileptic seizures are caused by complex processes in the body. The molecular processes that contribute to epilepsy involve transcription factors (McClelland et al., 2011; Mazzuferi et al., 2013), chromatin methylation processes (Kobow and Blümcke, 2011) and small none-coding RNAs (Jimenez-Mateos et al., 2011, 2012; McKiernan et al., 2012)

The diagnosis of epilepsy based on the symptoms is notoriously difficult. Disease related Biological markers can help improve better disease recognition, drug targets specification and patient treatment management. MiRNAs represent a class of biological markers that have not been thoroughly investigated in the context of epilepsy disease (Mazzuferi et al., 2013; Kretschmann et al., 2015a). The aim of this study is to investigate the role of microRNAs after generalized seizures and assess their suitability and potential as invasive (from brain hippocapmus tissues) and non-invasive (from blood serum samples) biomarkers for epilepsy. For this purpose experiments are done in mouse and rat disease models and high-throughput data are generated from hippcampal tissues and blood serum samples.

MicroRNAs (miRNAs) are short (17-28 nucleotides long) single-stranded and highly conserved none-coding RNA molecules. They are involved in post-transcriptional regulation (repression and silencing) of genes (Aravin and Tuschl, 2005). Many studies showed that one miRNA may regulate hundreds of protein-coding genes thus playing central role in many important biological processes (Hutvágner and Zamore, 2002; Pillai, 2005). They are found to be highly abundant in brain tissues (Kosik, 2006; Im and Kenny, 2012; McNeill and Van Vactor, 2012) and have been detected in organisms body fluids such as urine, saliva and blood (Ross and Davis, 2011; Chen et al., 2012). Circulating (non-invasive) miRNAs are part of the cell-cell communication system and are found to have stable expression in blood (Schöler et al., 2011). They are transported in plasma and delivered to recipient cells by high-density lipoproteins (HDL) (Vickers et al., 2011). Recent studies suggested that miRNAs play essential roles in neurogenesis, in particular epileptogenesis processes and maintenance and progression of epileptic state (Schaefer et al., 2007; Mitchell et al., 2008; Di Stefano et al., 2011; Jimenez-Mateos et al., 2011; Margis et al., 2011; Creemers et al., 2012; Jimenez-Mateos et al., 2012; McKiernan et al., 2012; Pritchard et al., 2012; Volvert et al., 2012; Wang et al., 2015). Therefore, they can be used as sensitive biological markers for epilepsy.

MiRNA profiling studies in epileptic tissues have revealed highly selective and spatiotemporal alteration in expression patterns. However, inconsistencies in expression of deregulated miRNAs have been observed. The reason for this could be the multi-species tissues (rat versus mouse), different epilepsy models (chemical-induced versus electrical-induced SE) used in the studies and the comparison of different time-points

(Kretschmann et al., 2015a). Therefore, more experimentally controlled studies are required in order to unveil miRNAs functions in the epilepsy disease.

In order to limit these problems a number of TLE animal models have been developed that are useful for the standardized experimental designs (Turski et al., 1983; Mazzuferi et al., 2012). Each of these models have distinct characteristics regarding the pathophysiological parameters such as onset of seizures, occurrence of recurrent seizures, seizure severity and hippocampal sclerosis (Jimenez-Mateos and Henshall, 2013). These parameters are expected to influence the pattern of miRNA activities during epileptogenesis. Chronic SE is caused in animals of these models via long-term chemical (by drugs) or electrical stimulation, so that a recurrent epileptic seizure can be triggered when needed. The animals show pathophysiology that is typical for epilepsy such as neuronal inflammation. Further animal models for acute TLE are obtained by triggering a single seizure in naive animals which do not show TLE-related pathophysiology.

In this work, miRNA expressions at different post-seizure time-points are compared to their pre-seizure expressions in different naive chronic and acute epilepsy perturbation models in rat and mouse. Samples are taken from hippocampal tissues and from blood serum. The following mouse models are used in this study: (I) Pilocarpine SE model; where status epilepticus (SE) is provoked chemically by the drug Pilocarpine. (II) Self-Sustained Status Epilepticus (SSSE); where SE is provoked via weak and irregular electrical stimulation "kinlding" of brain in animals through implanted electrodes in the amygdala. It is assumed that temporary neuronal discharges leads to long-termed changes in the nerve cells and create SE. The kindling procedure is increasingly used to study epilepsy. These two models represent chronic types of epilepsy. (III) 6-Hertz model; where a single acute seizure is triggered by monopolar pulses at frequency of 6 hertz and current intensity of 44 mA. This model represents an acute type of epilepsy. (IV) Another SSSE model in rat is also used together with two control groups: a "sham" control group where animals have undergone electrodes implantation surgeries but have not been kindled, and a "naive" group which have not been subjected to surgeries. Profiling was performed using microarray in the mouse models and high-throughput qPCR technologies in the rat models.

The miRNA profiling data are of a lower dimensionality compared to gene expression mircoarray data and HT-qPCR data have many undetermined Ct values. Therefore, appropriate advanced normalization techniques are applied to the data. In addition, a differential expression analysis procedure is proposed which uses distribution parameters estimates for censored data. The procedure showed higher detection power compared to generic methods and enabled detection of novel deregulated miRNAs in the data.

The results of the profiling analyses revealed many deregulated miRNAs in the different animal models and time points. Significantly high proportion of the deregulated miRNAs in the Pilocarpine chronic model were also active in the SSSE chronic model but very few of them were affected in the 6-Hertz model. This delineate the difference between the two chronic models and the 6-Hertz acute model. The activity of many deregulated miRNAs was confirmed via RT-PCR experiments. Pathway and gene ontology analyses revealed significant enrichment of target mRNAs of deregulated miRNAs in biological processes terms and biochemical pathways that are relevant to neurological malfunctions. This might help focus the exploration tactics for novel therapeutic applications in epilepsy.

5.2 Material and Methods

5.2.1 Experimental Design

Two large experiment studies have been performed in different epilepsy models in mouse and rats. miRNAs are profiled via microarray (see section 2.1) and high-throughput qPCR technologies. Significantly deregulated miRNAs were validated via RT-PCR experiments (details in section 2.3).

MiRNA Profiling by Microarray: Experiments are done in mouse. Hippocampal Samples of $n = 8$ of each set were taken at 24 hours and 28 days post seizure in both chronic models. In the acute 6-Hertz model samples were dissected at 0, 3, 6, 24 & 72 hours. Control samples for Pilocarpine model are taken from naive animals at 24 hours and 28 days. A single control set was prepared for the SSSE model at 28 days. Samples at time-point 0 hours were considered the control set in the 6-Hertz model. Expression profiles of 579 mature miRNAs in hippocampal tissues are measured via two-channel cDNA Microarrays in two chronic epilepsy mouse models, namely the Pilocarpine SE model and the SSSE model, and in an acute seizure 6-Hertz model.

MiRNA Profiling by High-Throughput qPCR: Experiments are done in rat. Samples are taken from kindled animals (rats subjected to a surgery to implant electrodes into the amygdala which are used later to stimulate (kindle) status epilepticus SE in animals and to trigger generalized seizure), sham animals (animals that are subjected to surgery but no kindling) and from naive animals. Profiling of 752 mature miRNAs was performed at 4 weeks before triggering generalized seizures in animal models and at 2 minutes, 4 hours, 24 hours, 1 week and 4 weeks after the seizure. Samples are

taken from blood serum and hippocampal tissues. High-throughput qPCR (CYBER-Green® and ABI-7900®) Exiqon® 2-panels chips are used for profiling miRNAs from blood and hippocampus samples. An overview of the experimental design is found in Figure B.1.

5.2.2 Differential Expression Analysis Procedure for Censored Expression Data

The maximum number of amplification reaction cycles in qPCR is typically limited to 40 cycles (details in section 2.3). This intrinsic feature of qPCR technology poses the problem that RNA expression is sometimes not fully quantifiable, i.e. Ct values are not determined for those features whose amplifications have not reached the threshold after 40 cycles. This is particularly the case in our study data of cell-free blood serum with low DNA concentration. Many miRNAs were not expressed across the samples in the data. A common way to treat these undetermined features is by imposing their expression Ct value to 40 or by excluding them from the analysis. However, this renders strongly negative-skewed and often bi-modal distributions of Ct values in the samples which cannot be corrected by post experiment analysis including advanced normalization techniques (Figure B.2). This lead to biased results.

Imposing Ct values of 40 for undetermined features or excluding these from the data is therefore not appropriate, because the true expression values of these features are simply not known and could be higher than 40. This implies that conventional t-test based approaches cannot be applied directly, because the statistical distribution of signals is truncated at the right-hand side.

In order to account for this issue we developed the *double detection* procedure, which comprises a *detection test* as first step and in the second step a subsequent moderated t-test based approach is used either via the "limma" method or via estimating parameters form censored data (Figure 5.1). This procedure can be used in differential expression analysis for any other data where expressions of the samples or the features are partially not determined. It can also be formulated for left-censored or interval-censored data.

First Step: The *detection test* is a one-sided t-test to investigate whether Ct values of a particular feature in a given sample group are statistically significant lower than a defined threshold, meaning there is a significant detected RNA expression. The threshold was set to the average of the maximal observable, normalized Ct values per sample in a given experimental condition. Formally, this means testing for the null hypothesis: $H0 : Arg\_Avr(Ct)_{c,m} \geq Arg\_Avr(max(Ct)_c)$, where m denote

features and the conditions c are factor combinations of treatments and replicates factors. It is worth mentioning that due to normalization the maximum Ct value in a given sample could be higher than 40. The *detection test* aims to filter out features that are undetectable in a given sample group. Thus, it ensures the applicability of t-test based significance tests in the subsequent step.

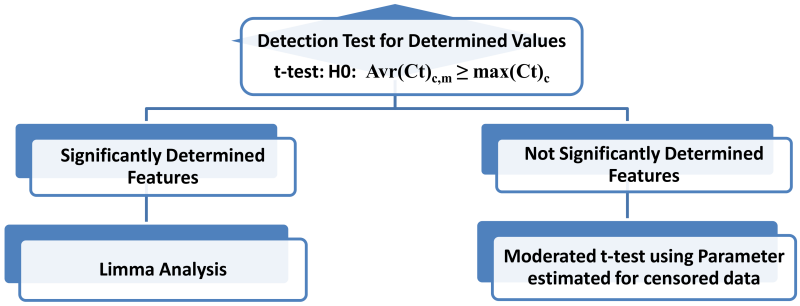

Figure 5.1: Flow chart of processing and testing procedure for HT-qPCR data. The *detection test* decides whether a feature is at all expressed in a given experimental condition. Differential expression analyses using "limma" procedure is performed for features, which passed the *detection test* in both experimental conditions. Variance and mean of Ct values for features that did not pass the *detection test* are estimated via univariate distribution estimation method for censored continuous data. The estimated parameters are used for moderated Student t-test for differential expression

Second Step: The *Moderated test* Features, which passed the detection test are checked for their differential expression using the "limma" method utilizing empirical Bayes (details in subsection 3.2.3). However, typically there will be also many features that do not pass the *detection test*. The data for these features can be conceived as continuous randomly right-censored data (of type I). There are several methods to fit distributions to data with different types of censoring using various strategies (Turnbull, 1975; Greene, 2005; Leha et al., 2011; Busschaert et al., 2010; Commeau et al., 2012). We used the method described in Delignette-muller and Dutang (2015) to estimate univariate distribution parameters (i.e. for individual features) form the censored HT-aPCR data. The normal distribution is chosen as candidate and parameters are estimated from the cumulative distribution function of the parametric distribution using maximum likelihood method (Andersen, 1970; Aldrich, 1997). Variances and means estimated this way for the individual features which did not pass the *detection test* are used instead of the empirical means and variances. This is done by plugging the estimates following the idea of Smyth (2005) in a test statistics of the form:

$$t_{gj} = \frac{\hat{\beta}_{gj}}{\sqrt{s_g^2}\sqrt{c_{gj}}}\,, \tag{5.1}$$

where c_{gj} is the estimated combined variance of the two conditions A and B of feature g in array j of the design matrix, and the coefficient $\hat{\beta}_{gj} = \hat{x}_A - \hat{x}_B$ is the difference between the estimated means. The prior value s_g^2 is estimated from the uncensored part of the data.

Validation via Simulated High-Throughput qPCR Data: In order to examine the performance of the above procedure 10,000 data sets with two conditions, 12 samples per condition and 750 features are simulated. A modified version of R package "madsim[1]" is used for this purpose (Dembélé, 2013). The function uses a beta distribution with additive Gaussian noise to originally simulate microarray data. The parameters of the distributions are modified to generate Ct values bound between 19 and 50. The package allows the use of sample seeds to estimate means and standard deviations required to initiate the simulation, the real HT-qPCR data set is utilized to this end. 2% of the features are considered differential and symmetrically distributed between over and under expressed. A simple cut at Ct value 40 is used to define the undetermined values.

On average 37% of the features have undetermined expression in at least one sample. As in the real HT-qPCR data, this percentage increase for cut values lower than 40. The percentages of undetermined expressions per feature in the simulated data is comparable to those of the real data, in particular to hippocampus data Figure 5.2. In addition, the simulated data show mean-variance distribution patters that are similar to these of the real HT-PCR data (see appendices Figure B.3).

The *double detection* procedure starts by applying the *detection test* on the normalized observations of the individual features in each perturbational condition after filtering out all features which are entirely not determined in all the samples of at least one of the two testing conditions. The *detection test* decides whether a feature is at all expressed in a given experimental condition or not. This results in two sets of features; a set of features which passed the *detection test* in both experimental conditions, and another set of features that did not. Differential expression analyses using "limma" procedure is performed on the features in the first set. Variance and mean for individual features of the second set are estimated using the method explained above. These values are used in a moderated t-test instead of the empirical values Figure 5.1.

[1] `https://cran.r-project.org/web/packages/madsim/index.html`

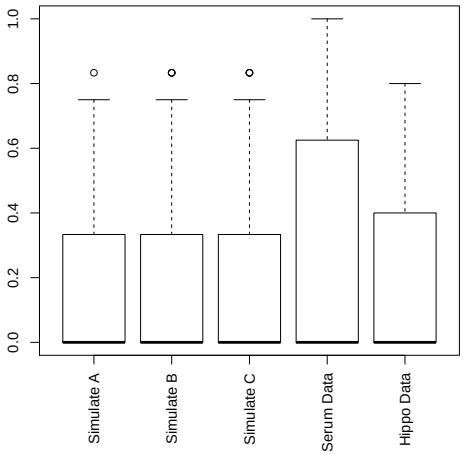

Figure 5.2: Distribution of censored (undetermined) expressions in the simulated and real data. The boxplots shows the percent of undetermined *Ct* values per feature in three exemplary simulated data sets, in serum and in hippocampus data.

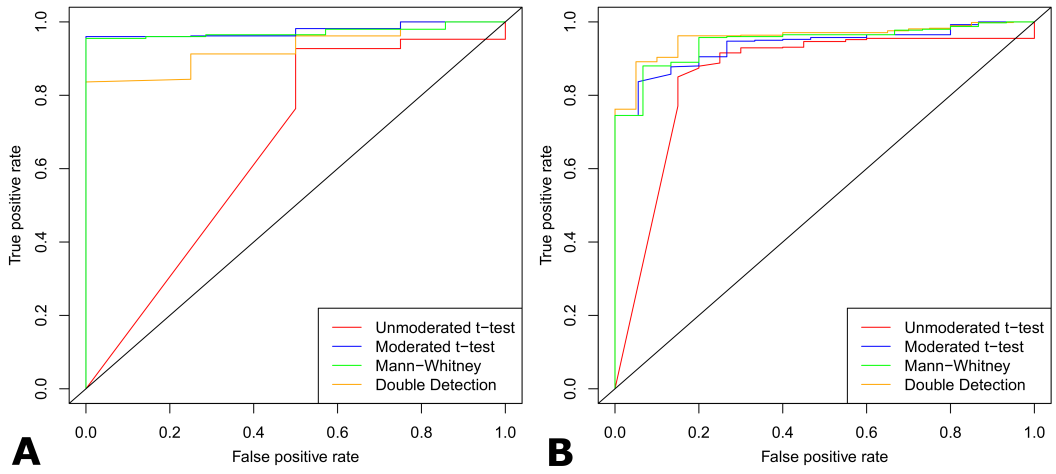

Figure 5.3: Performance of the *double detection* procedure based on ROC curves from simulated data. The method is compared to the nonparametric Mann-Whitney test, the unmoderated and moderated t-tests at **A** 10% censoring and **B** 30% censoring

The *double detection* procedure is compared against standard methods that simply ignore the data censoring i.e. consider $Ct\text{-}value = 40$ for undetermined read-outs. These methods, which are often employed, are the moderated & unmoderated two-sided Student t-test and Mann-Whitney test. The performances of these tests are compared based ROC curves in Figure 5.3 (Akobeng, 2007; Parikh et al., 2008; Naeger et al., 2013). The figure shows that for low percentage of censored observations in the data ($< 15\%$) the detection power of the *double detection* procedure is only better than the unmoderated t-test (Figure 5.3 **(A)**). However the performance of the method seem to improve with increasing level of the censorship. The *double detection* is found superior to the other methods in heavily truncated data with censoring percentage between 30% and 40% (Figure 5.3 **(B)**). The performances of the tests are also given via sensitivity, specificity, Positive Predictive values and Negative Predictive values in appendix Table B.1).

However, attention should be paid to the number of samples per condition and the distribution of the missing values since the maximum likelihood methods requires sufficient sample size in order to produce unbiased estimates.

5.2.3 Normalization and Differential Expression Analyses in Microarray Data

Normalization, Preprocessing & Quality Control: Red and green intensities and their respective background values were extracted from two-channel arrays. In order to avoid negative corrected intensities and to reduce variability of low intensity log-ratios, the normal-exponential convolution 'Normexp' method was used for background correction (see subsection 2.2.1). The LOWESS normalization procedure is used for within-array normalization, and the variance stabilizing normalization (VSN) for between-array normalization (see section 2.1 & Figure B.4). A rigorous quality assessment before and after the normalization confirmed the quality of the chips. The 'GAL' files from Exiqon together with the 20th release of miRBase® (Griffiths-Jones, 2004; Kozomara and Griffiths-Jones, 2011) were used for chip annotation.

Differential Expression Analysis: Differential expression analyses for miRNAs were performed using linear models for microarray data analysis "limma" utilising the empirical Bayes method (see section 3.2) Statistical dependencies of samples between different conditions and replicates were considered via a factorial design matrix in "limma" using a 'condition-replicate' factor. Contrasts were considered for interaction effects. Correlations between the technical quadruplicates in the chips were taken into consideration, and spot quality weights were used (manually flagged spots, empty, poor and negative spots are downweighted with 0.7, 0.4, 0.2 and 0.1 factors,

respectively). Corrections for multiple testing (subsection 3.2.4) were done using Benjamini and Hochberg (1995) method. Significant differentially expressed miRNAs were reported at $FDR_{BH} \leq 0.05$ and visualized via Volcano plots (e.g. Figure B.5)..

Pilocarpine-treated mice samples are compared to their counterpart naive samples at 24 h and 28 days, respectively. SSSE samples at 24 h and 28 days were compared to a single SSSE control group. In 6-Hertz data, samples at the subsequent time points (3, 6, 24 and 72 hours) were compared to the samples at the initial time point (0 hours). For further investigation of the differences between late and early time-points contrast of these time points (i.e. 28 days versus 24 hours) are compared withing each of Pilocarpine and SSSE model.

5.2.4 Normalization and Differential Expression Analyses in High-Throughput qPCR Data

Normalization Preprocessing & Quality Control: Ct values of miRNAs are extracted via Sequence Detection System (SDS 2.4) software. Inter-plate calibration of the two panels was performed by normalizing signals relative to the feature UniSp3 IPC. Singal quality was checked by the defined spike-in Controls in the first panel (UniSp2, UniSp4, uniSp5, UniSp6 & cel-miR-39-3p). The Exiqon 2-panels chip contains several potential reference RNAs (RNU1A1, RNU5G and U6-snRNA). These are usually used as indigenous control to normalize miRNA expressions in cell material. However, these RNAs are degraded in serum samples and therefore could not be used here. Instead, a geometric mean based global normalization (division of raw Ct values by the geometric mean expression of all features) was used to remove non-biological and systematic variations in fractional Ct values between samples (Dvinge and Bertone, 2009; Vandesompele et al., 2002; Yuan et al., 2006). This is equivalent to converting data to linear scale. A rigorous quality assessment confirmed the good performance of the normalization method. The plate layout and annotation files supplied by the manufacturer for the two panels together with the 21st miRBase® release (Griffiths-Jones, 2004, 2010; Kozomara and Griffiths-Jones, 2011) were used for annotation.

Differential Expression Analysis: The procedure devised in subsection 5.2.3 is used for the differential expression analysis of HT-qPCR data. Time dependency was modeled via a grouping factor "time point", and an additional factor "animal model" accounted for the different treatment conditions. Finally, a random effect was added to pool variances for the same animal, because several animals had been measured repeatedly. Benjamini and Hochberg (1995) method was used for multiple testing correction (subsection 3.2.4) and significant differentially expressed miRNAs are

reported at $FDR_{BH} \leq 0.05$. MiRNA(s) which are detected in one testing condition but have undetermined expressions in all the samples of the other condition are ranked based on detla-significance measure (compared to same feature in the control condition, or compared to a reference gene; $\triangle Ct = Ct_{Target} - Ct_{Reference}$. A threshold of $|\triangle Ct| \geq 4$ is considered to delineate miRNA disease relevance (Livak and Schmittgen, 2001; Schmittgen and Livak, 2008).

Samples of sham and kindled types in the different time points (pre-kindling (0 hours), 2 minutes, 4 & 24 hours, 1 & 4 weeks) are compared to their counterpart groups in Naïve (wile type). Samples of kindled type are compared in the same manner to samples of sham type. In addition, different sample sets in subsequent time-points in each individual rat model are compared to their corresponding reference time point set.

5.2.5 Normalization and Differential Expression Analyses in RT-PCR Data

RT-PCR data are produced to validate microarray results of the different time-points in the mouse models. RT-PCR data are normalized using the reference gene *RNAU6*. A two-samples paired *t-student test* is used for differential expression analyses. In addition, fold changes are calculated using the $\triangle\triangle$ Ct method (Livak and Schmittgen, 2001). Fold change is defined as $FC = 2^{-\triangle\triangle Ct}$. PCR technology and normalization methods are briefly explained in section 2.3.

5.2.6 Functional Analysis of miRNA Target Sets

To investigate the associations of deregulated miRNAs ($FDR_{BH} \leq 0.10$ and $\log_2(FC) \leq 0.5$) to biological processes and cell activities a statistical enrichment analyses of Gene Ontology (GO®) terms (Ashburner et al., 2000) and in Kyoto Encyclopedia of Genes and Genomes (KEGG®) pathways (Kanehisa and Goto, 2000)) are performed (details in section 3.5). This was done via gene set enrichment analysis with regard to the annotation of miRNA(s) to their target genes. MicroRNA predicted target genes were retrieved from TargetScan® database release 6.2 (Lewis et al., 2003). These target sets were tested for overrepresentation of biological process ontology terms in GO® and of pathways in KEGG® comparing them to the gene universe of all predicted microRNA targets. A hyper-geometric test was used for this purpose (subsection 3.5.1) and multiple test correction conducted via Benjamini and Yekutieli (2001) method for false discovery rate control under dependency (subsection 3.2.4). Significant KEGG® pathways and GO® terms are reported at $FDR_{BY} \leq 0.10$.

5.3 Results

5.3.1 Global Expression Comparisons Revealed Differences between Chronic Epilepsy and Acute Seizure Models

Principal component analyses (PCA) are performed to analyze the variance of miRNA expression in all three models in more detail. The expression values of all detected miRNAs were included in this analysis.

The PCA plots showed a segregation of the different sample groups between early and late time points and in treated and naive set of the Pilocarpine model (Figure 5.4 A). Within each group the samples were clustering together which indicates that similar miRNA expression patterns are observed. Furthermore, both control groups gathered closely and are clearly separated from the pilocarpine 24 hours and 28 days time points. Interestingly, the 24 hours pilocarpine mice represented the lowest variability within the group and were more separated from both control and the chronic time points (28 days). Furthermore, early as well as late time points are clustered away from their respective controls and from each other in the pilocarpine model.

Similar to the pilocarpine model, the samples of SSSE model within each experimental group are gathered together (Figure 5.4 B). However, the samples in the SSSE model showed a higher relative variability compared to the experimental groups of the pilocarpine model. Overall, as observed also in the pilocarpine model, the early and late time points in the SSSE model can clearly be separated from each other and from the control group. The clustering observed in both chronic models indicates that differences in the miRNA expression pattern are detectable between the analyzed sample groups.

The PCA for the acute seizure model (6-Hertz) is shown in Figure 5.4 C. In contrast to the chronic epilepsy models, the samples groups of the 6-Hertz model showed minor separation. However, partial clustering within the control, 24 hours and 72 hours groups on one side and within the 3 and 6 hours groups on the other side can be observed. This suggests that miRNA expression differs between the early and late time points in the 6-Hertz model.

Overall, a good separation of the control and treatment groups in the chronic models is evidenced, whereas this clustering was less pronounced in the acute seizure model. Furthermore a PCA analysis using all samples of all models and all time-points show clear separation between acute and chronic models. A moderate separation between the two chronic models could also be observed, though this might be caused by a batch effect.

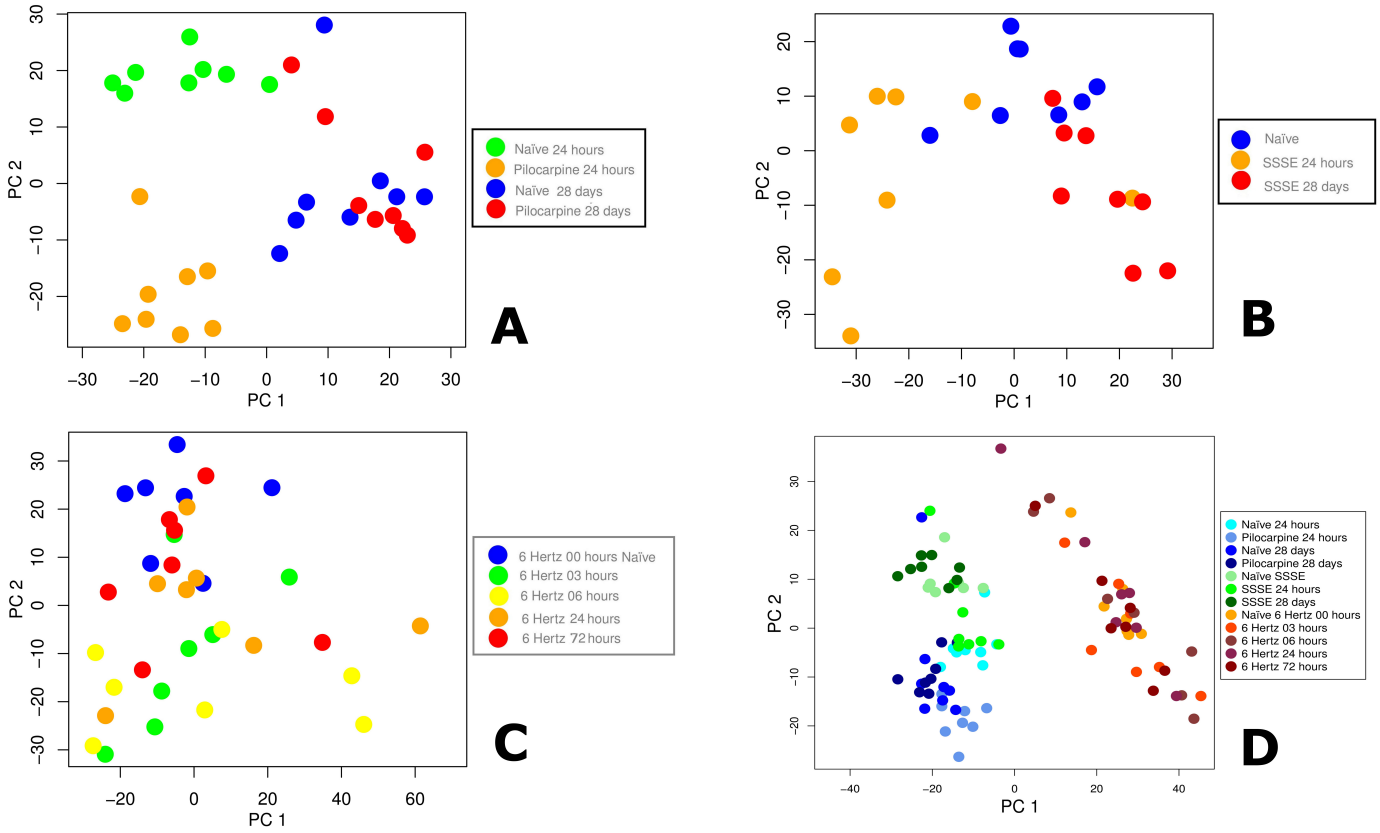

Figure 5.4: Principal component analysis (PCA). Plots of the first and the second principal components based on all detected miRNAs. **A** samples of a animals of the pilocarpine model: 24 h and 28 days naive versus 24 h and 28 days after pilocarpine-induced status epilepticus (SE); **B** animals of the SSSE model: 24 h and 28 days naive versus 24 h and 28 days following electrically induced SE; **C** animals of the 6-Hertz model: 0 (naive), 3, 6, 24 and 72 h following seizure induction; **D** showing sample of all animal models at all time-points.

A hierarchical clustering of all samples confirmed the finding suggesting closer correlations of expression patterns in chronic model samples (Figure 5.5). This highlights the differences in miRNA expression patterns between the chronic models and the acute seizure model. However, expression patterns in the control samples of each one of the chronic models are more similar to the expressions of their counterpart treated groups.

Furthermore, within the chronic models, early time-point (24 hours) is separated from late time-point (with the control group in between) suggesting that distinct miRNA expression patterns are potentially involved in the early development of chronic epilepsy. With progression of the disease (28 days), the two models are more diverging. Interestingly, in the pilocarpine model, early and late time points seem to be less related to each other compared to the same time points in the SSSE model. This suggests that differences in miRNA expression pattern during disease development are more pronounced in the pilocarpine model.

In addition, the dendrogramm show similarity of expression patters of early time-point 0 hours and late time-points 24 hours & 72 hours in 6-Hertz model. Expression patterns in this model at 3 hours are more similar to these at 6 hours. This suggests that particular similar miRNA are potentially involved in the early development of acute epilepsy.

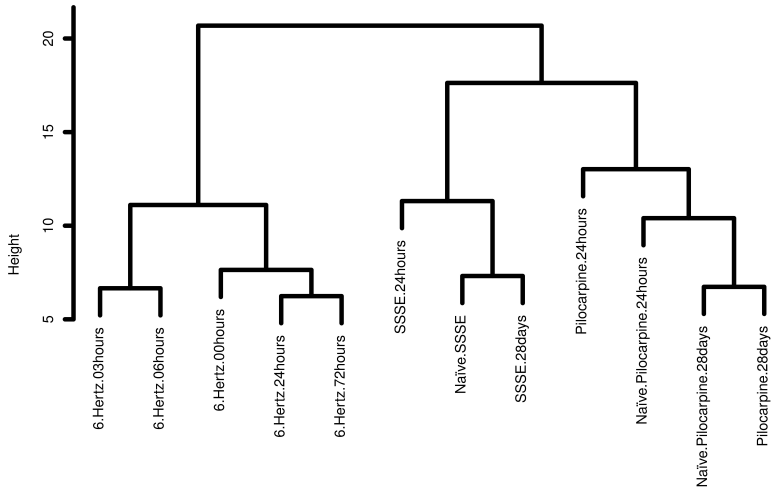

Figure 5.5: Dendrogram of hierarchical clustering. Using Euklidean distance and complete linkage agglomoration of normalized miRNAs expression values for mouse hippocampus samples of the pilocarpine and SSSE models at 24 hours and 28 days following SE, as well as of the 6-Hertz model at 3, 6, 24 and 72 hours following seizure

5.3.2 Double Detection Procedure Identified Differentially Expressed miRNAs in Rat Serum High-Throughput qPCR Data

The expression of 752 miRNAs are profiled in serum samples from naive, sham and kindled groups using high-throughput qPCR technology. Due to low concentrations of biological materials in the samples and qPCR experimental implications RNA expression Ct values in 42% of the features could not be determined in at least one sample. The *double detection procedure* for differential expression enabled the detection of deregulated miRNAs in serum data. For each of the naive, sham and the kindled groups post seizure samples at different time-points are compared to the corresponding post seizure groups (Table 5.1).

Table 5.1: Numbers of differentially expressed miRNAs at the different post seizure time-point compared to preseizure time point in serum samples from rat naive, sham and kindeled groups at $FDR_{BH} \leq 0.01$ and at $FDR_{BH} \leq 0.05$.

Rat Model Type	Naive				
	02 Minutes	04 hours	24 hours	1 week	4 weeks
Significance $FDR_{BH} \leq 0.01$	0	0	0	3	0
Significance $FDR_{BH} \leq 0.05$	4	3	0	21	21
Rat Model Type	Sham				
	02 Minutes	04 hours	24 hours	1 week	4 weeks
Significance $FDR_{BH} \leq 0.01$	1	0	0	2	1
Significance $FDR_{BH} \leq 0.05$	1	10	1	15	14
Rat Model Type	Kindled SE				
	02 Minutes	04 hours	24 hours	1 week	4 weeks
Significance $FDR_{BH} \leq 0.01$	18	7	0	1	1
Significance $FDR_{BH} \leq 0.05$	34	27	3	2	19

Higher numbers of significant differentially expressed miRNAs are detected in the different time-points in kindled rat samples compared to the corresponding time-points in naive and sham groups. This highlight the kindling effect in the former group. In general higher regulation of miRNAs could be observed at early time-points 2 minutes and 4hours compared to late time-point. This indicate higher acute effects of provoked generalized seizure.

5.3.3 MiRNAs Show Different Expression Patterns in Pilocarpine, SSSE and 6-Hertz Mouse Models

The expression of 579 miRNAs are profiled in the different mouse models using the Exiqon miRNA microarray profiling platform. For each model the treatment groups at different time-points are compared to the corresponding control groups. These comparisons brought up high-to-moderate numbers of significant differentially expressed miRNAs (Table 5.2, details in Appendices B Excel files).

In general all mouse models show higher regulation of miRNAs at early time-points compared to late time-point. In Pilocarpine and SSSE models the numbers of deregulated miRNAs at 24 hours (at $FDR_{BH} \leq 0.05$; 99 & 87 respectively) are almost as double as their numbers at 28 days (51, 57 respectively). This gives two indications; firstly: that epilepsy seizure affect large numbers of microRNAs, secondly: The effect is higher immediately after the seizure and fades in the course of time.

Table 5.2: Numbers of differentially expressed miRNAs at the different time-point in hippocapal samples from each of the epilepcy models at $FDR_{BH} \leq 0.01$ and at $FDR_{BH} \leq 0.05$ (appendices CHAPTER 5 APPENDICES Excel files).

Mouse Model Type	PiloCarpine SE		SSSE SE	
	24 hours	28 days	24 hours	28 days
Significance $FDR_{BH} \leq 0.01$	68	28	57	39
Significance $FDR_{BH} \leq 0.05$	99	51	87	57
Mouse Model Type	6-Herz SE			
	03 hours	06 hours	24 hours	72 hours
Significance $FDR_{BH} \leq 0.01$	78	111	4	1
Significance $FDR_{BH} \leq 0.05$	99	146	17	4

The numbers of deregulated miRANs in the 6-Hertz model confirms this finding. These numbers are even higher at early time points 3 & 6 hours (99, 146 respectively) compared to later time-points 24 hours and 3 days (17, 4 respectively). The 6-Hertz acute model, unlike the chronic models (Pilocarpine and SSSE), is characterized by one-time induced seizure. Therefore, the 6-Hertz model is expected to be sensitive at early time-points, however, long-term changes of miRNA expression in it are expected to be less pronounced at later time-points.

Deregulated miRNAs were subjected to a two-way hierarchical clustering showing up-regulation and down-regulation of miRNAs. The created heatmaps, one for each

comparison, showed patterns of altered miRNA expression indicating clearly that up- and down-regulated miRNAs are present in each comparison (Figure 5.6, Figure 5.7, Figure 5.8).

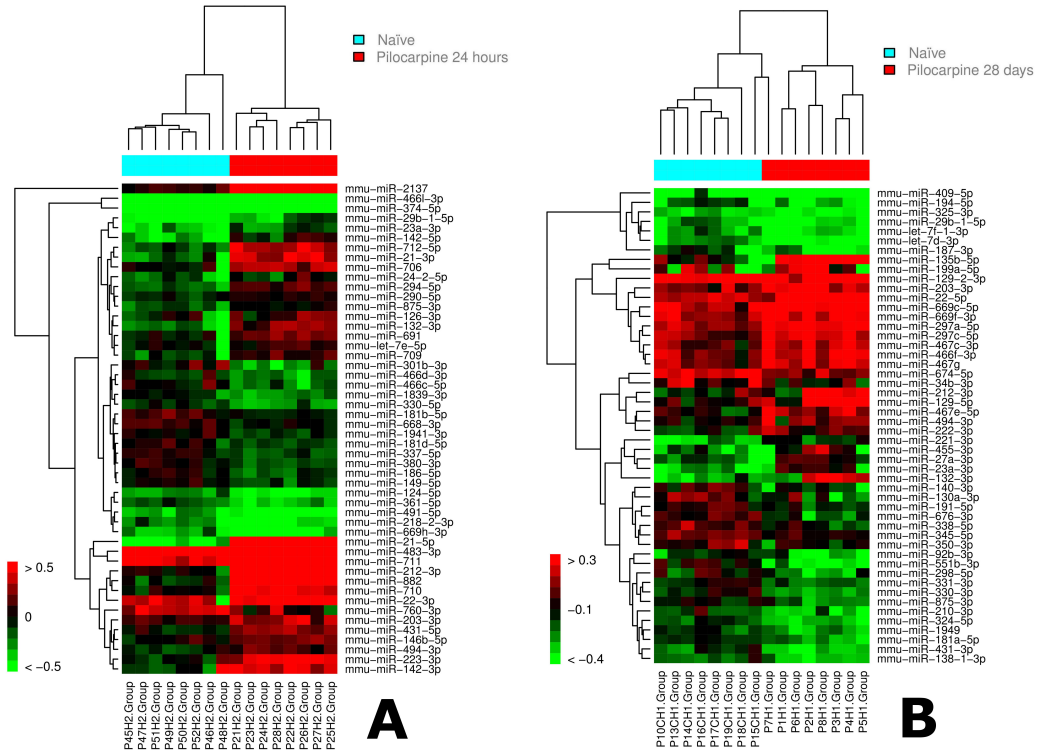

Figure 5.6: Heatmap plots. The result of the two-way hierarchical clustering of miRNAs and samples. Expressions of the top 50 differentially expressed miRNAs ($FDR_{BH} \leq 0.05$) are shown. **A** shows Pilocarpine induced SE mice at 24 hours versus the naive group. **B** shows Pilocarpine induced SE mice at 28 days versus the naive group.

In the pilocarpine model especially the early time-point (24 hours) showed clustering of the naive and treated animals and a good separation between both groups (Figure 5.6 A). At the later time point clustering and group separation are less pronounced (Figure 5.6 B) suggesting that changes in miRNA expression are more distinct at the early time point. The same can be observed in SSSE model where a clear clustering and group separation of the naive and treated animals at early time-point 24 hours is better than in late time point 28 days (Figure 5.7 A & B).

In the 6-Hertz epilepsy model the separation of naive and treatment samples is best at time-point 6 hours (Figure 5.8). However, the 4 identified miRNAs at time-point 72 hours also separate the two groups very well in the hierarchical clustering (Figure 5.8 D). Up-regulated and down-regulated miRNAs are identified in all models and all time-points. The numbers of down-regulated miRANs are generally higher in all models and time points.

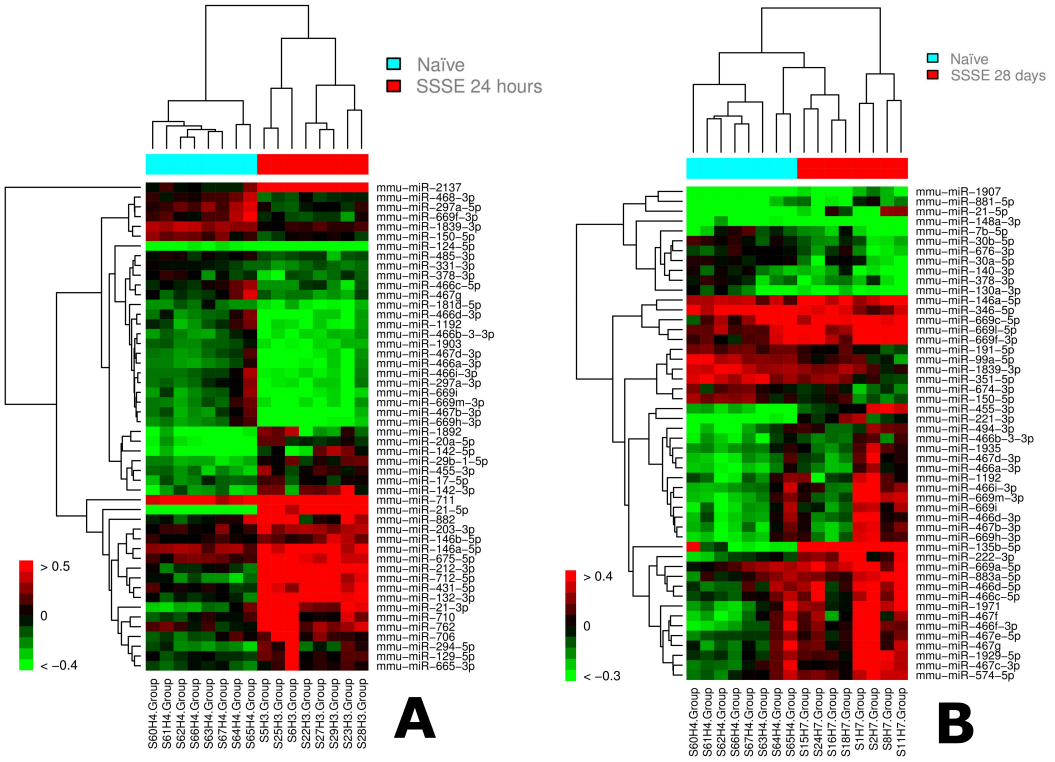

Figure 5.7: Heatmap plots. The result of the two-way hierarchical clustering of miRNAs and samples. Expression of the top 50 differentially expressed miRNAs ($FDR_{BH} \leq 0.05$) are shown. **A** shows electrically induced SE mice (SSSE) at 28 days versus the naive group. **B** shows electrically induced SE mice (SSSE) at 28 days versus the naive group.

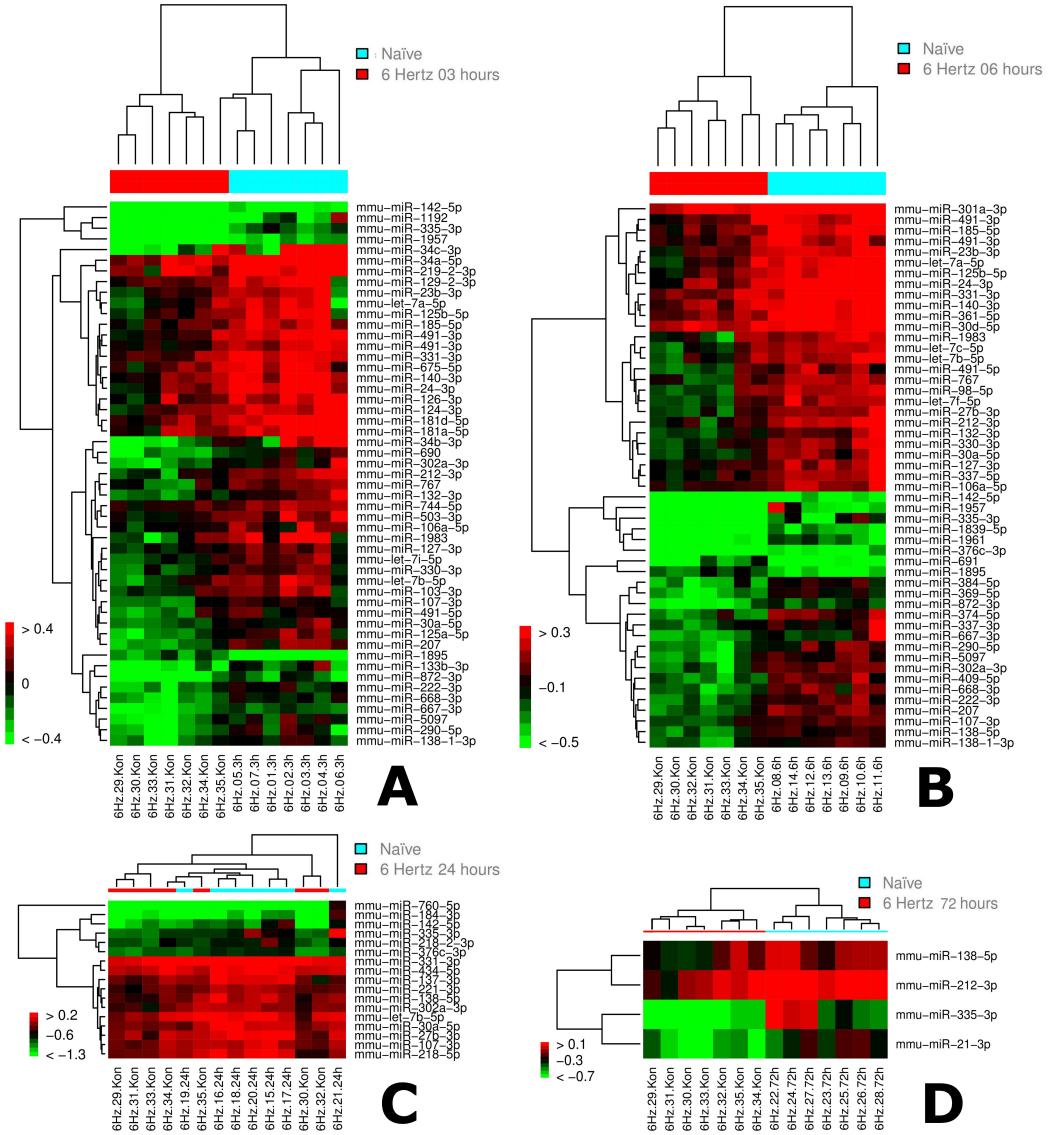

Figure 5.8: Heatmap plots. The result of the two-way hierarchical clustering of miRNAs and samples. Expressions of at most the top 50 differentially expressed miRNAs ($FDR_{BH} \leq 0.05$) are shown. The different heatmaps (**A-B**) show miRNA expressions in sound-induced SE mice (6-Hertz) at different time points versus the control naive group. **A** 03 Hours, **B** 06 Hours, **C** 24 Hours, **D** 72 Hours.

5.3.4 Overlap Analyses of Deregulated miRNAs in Mouse Models

In order to identify miRNAs linked to the disease development across all models the overlap of miRNA between the two chronic epilepsy models (pilocarpine and SSSE) and the acute seizure model (6-Hertz) was performed. Here, the analysis was based on all differentially expressed miRNAs (at $FDR_{BH} \leq 0.05$) for the comparisons of treatment and control groups within each model.

A first comparison was made between the three models at time-point 24 hours upon induction of the status epilepticus (Figure 5.9 A). Only 3 miRNAs (miR-142-5p, mmu-miR-331-3p & mmu-miR-30a-5p) are found differential in common in all models. Further 2 miRNAs (miR-218-2-3p & miR-335-3p) are shared between Pilocarpine and SSSE models. In contrast 36 miRNAs are found differential in Pilocarpine and SSSE models. This is might be due to the fact that fewer miRNAs are differentially expressed in the 6-Hertz model. This indicates that the chronic epilepsy and acute seizure models can be characterized by different miRNA expression patterns at the time-point 24 hours.

To investigate the both chronic models a detailed overlap analysis of their miRNAs sets was performed. At the early time-point (24 hours), the majority of altered miRNAs were exclusively present in the pilocarpine model (55 miRNAs; 55%), while 27 miRNAs (31%) were exclusively present in SSSE model. At the late time point (28 days), 18 miRNAs (35%) were exclusively specifically altered in pilocarpine, while 17 miRNAs (30%) were SSSE model-exclusive (Figure 5.9 B). In contrast, only a single miRNA is in common in all groups and 15 miRNAs are shared between the two chronic models at time-point 28 days. This demonstrates that big proportion of all deregulated miRNAs were model and time point specific.

MiRNAs that are differentially expressed in both chronic models and at both time points might be the most interesting miRNAs as targets for disease intervention. However, only a single miRNA (miR-494-3p) is overlapping. For this reason, all miRNAs that show up in both models at the early or late time-points are considered for further investigation. At 24 hours 36 miRNAs and further at 28 days 14 are identified as commonly deregulated for both models. The annotation of these miRNAs which are potentially therapeutically interesting are shown in appendices B Excel tables.

Analysis of the different time points in the acute seizure model 6-Hertz showed that early time points (3 and 6 hours) displayed higher significant miRNA deregulation, while at later time-points few miRNA were changed (Table 5.2). The majority of the miRNAs (67; 46%) were significantly deregulated uniquely at the 6 hours time point whereas at the 3 hours 26 miRNAs (26%) were exclusively deregulated (Figure 5.9

C). 73 miRNAs are in common in both groups and two (miR-335-3p & miR-138-5p) are found in all time-points. Taken together, our data show that in contrast to the chronic models, the majority of changes in differential miRNA expression were detected during the very early time points in the acute seizure model. Indeed, as the acute model is characterized by the induction of a single seizure long term changes in miRNA expression are expected to be less pronounced in the 6-Hertz model compared to the chronic models.

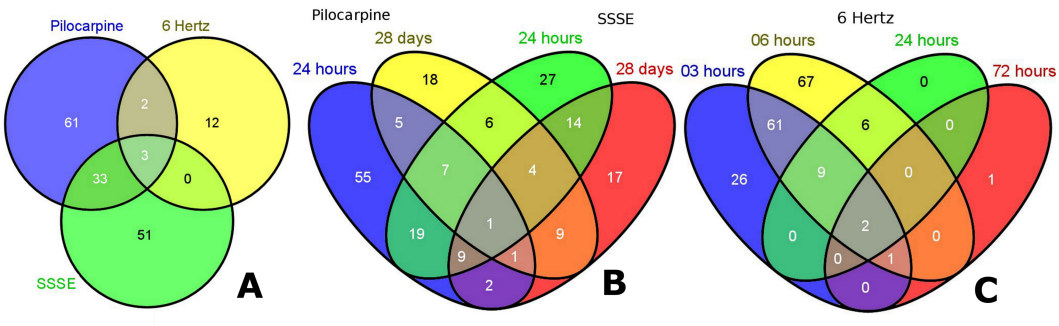

Figure 5.9: Venn Diagrams of all significantly deregulated miRNAs ($FDR_{BH} \leq 0.05$) in the different mouse models. **A** shows the numbers and overlaps of commonly deregulated miRNAs in 6Hertz, pilocarpine and SSSE models at 24 hours following induction of seizure. **B** shows numbers of deregulated miRNAs in the chronic pilocarpine and the SSSE model at 24 hours and 28 days following SE. **C** shows deregulated miRNAs in the acute 6-Hertz mouse model at 3, 6, 24 and 72 hours following seizure

5.3.5 Deregulated miRNAs Successfully Validated via External Experiments

RT-PCR technology is used for validation of the results of the microarray data. Additional independent samples were taken for this purpose from Pilocarpine and SSSE models at each time-point. RT-PCR runs are done in triplicates for miRNA sets in both models at each time points. In particular miRNAs displaying differential expression identified using microarray platform to span a range of high signals ($FDR_{BH} \leq 0.05$ and $\mid \log_2(FC) \mid \geq 1$) and deregulated-overlapping in chronic epileptic models at early and late time point, as well as overlapping in acute epilepsy models at early time points (Table 5.3).

Table 5.3: Validation of deregulated MiRNA's via RT-PCR experiments. Lists of highly up-regulated and down-regulated miRNAs in Pilocarpine and SSSE epilepcy models at time-points 24 hours and 28 days validated through RT-PCR procedure.

PiloCarpine SE		SSSE Model	
24 hours	28 days	24 hours	28 days
mmu-miR-21	mmu-miR-130a	mmu-miR-142-3p	let-7b
mmu-miR-124	mmu-miR-132	mmu-miR212	mmu-miR-221
mmu-miR-124*	mmu-miR-181	mmu-miR-297-3p	mmu-miR-222
mmu-miR-142-3p	mmu-miR-221	mmu-miR-882	mmu-miR-298
mmu-miR-142-5p	mmu-miR-222		mmu-miR-676
mmu-miR-212	mmu-miR-298		
mmu-miR-882	mmu-miR-382		
	mmu-miR-409-5p		

Endogenous control gene *RNU6*, which is also measured in parallel with RT-PCR from the samples, is used as reference gene for normalization. Fold changes are calculated using the $\triangle\triangle\ Ct$ method (Livak and Schmittgen, 2001). Fold change is defined as $FC = 2^{-\triangle\triangle Ct}$. A paired t-student test is used to test for differential expression and p-values are computed. See section 2.3 for details about PCR technology and normalization methods for PRC data.

Tha majority of moderately deregulated miRNAs ($FDR_{BH} \leq 0.5$ and $1 \leq \mid log_2(FC) \mid \leq 0.5$), e.g. miR-331-3p, miR-30a-5p & miR-211-5p in Pilocarpine model at 24 hours, showed no statistical difference between control and treatment groups in RT-PCR data (i.e $\mid FC \mid \leq 0.5$). In contrast, a number of highly deregulated miRNAs in Pilocarpine and SSSE models at 24 hours and 28 days are clearly confirmed in the RT-PCR data in terms of deregulation direction and the rough deregulation magnitude (Figure 5.10). For instance, the Ct values of miRNA-212 and miR-142-3p (which were found highly up-regulated in microarray data at 24 hours) increased significantly in both chronic epileptic models at 24 hours in comparison to the control groups.

While the fold change (FC) of miR-882 in Pilocarpine 24 hours animals is hardly above 0.5, it is over 3.5 in SSSE samples at 24 hours. This actually corresponds to the microarray signals of this miRNA in the models at that time-point. Similar patterns

could be observed in miR-221 at time-point 28 days; its FC is high in Pilocarpine animals and low in SSSE. Overall, the interesting sets of miRNAs listed in Table 5.3 have at least fold changes of 0.5 (Figure 5.10).

In general, the qRT-PCR results widely confirmed the validity of findings of the microarray experiments. These validated miRNAs can therefore be considered potential biological markers for epilepsy and their putative target mRNA might be investigated for further validation.

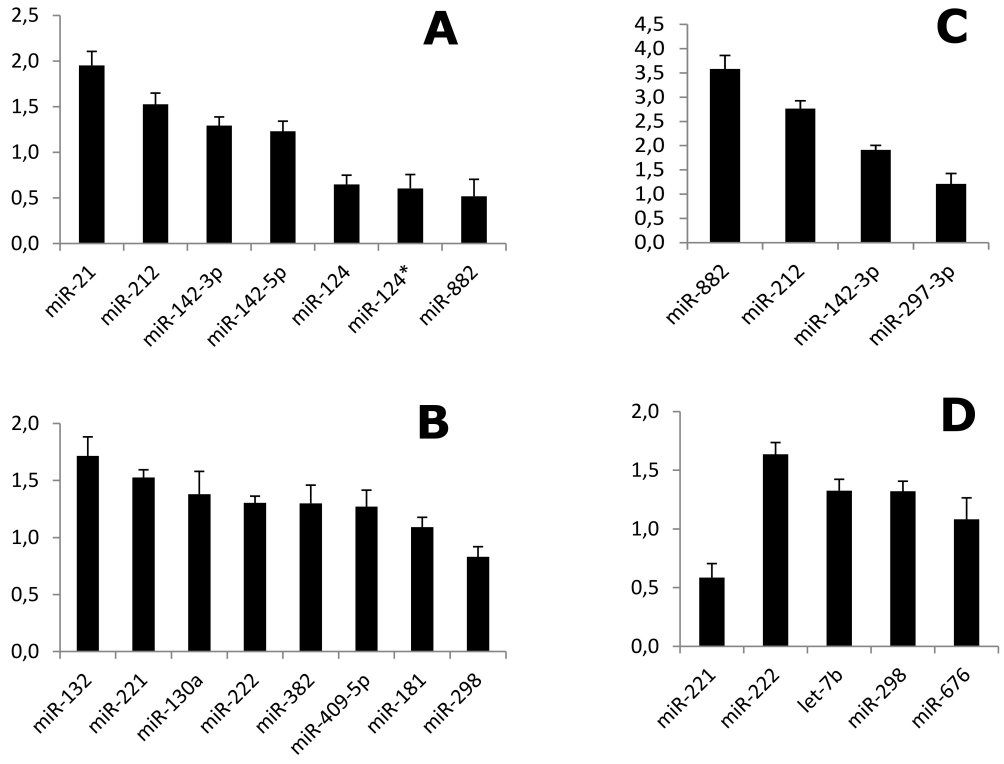

Figure 5.10: RT-PCR expressions of selected validated miRNAs which have been identified as differentially regulated in samples of the different mouse models at different time-points. Bar-plots show the absolute fold change values ($\mid FC \mid$). Fold changes defined as $FC = 2^{-\triangle\triangle Ct}$, where the Ct values are normalized to the reference gene *U6-RNA* ($\triangle Ct$) and compared to control Ct values ($\triangle\triangle Ct$). Sample size in the SE samples and control sets $N = 8$, RT-PCR reaction are done in triplicates. **A** & **B** show results in Pilocarpine mouse model at 24 hours and 28 days respectively. **C** & **D** show results in SSSE mouse model at 24 hours and 28 days respectively.

5.3.6 Targets of Deregulated miRNAs are Enriched in Meaningful Biological processes and Biochemical Pathways

Mapping of mRNAs targeted by miRNAs using 'targetScan' (Lewis et al., 2003) identified hundreds of genes ($\sim$ 279 - 500, depending on comparison) including many mRNAs known to be involved in epilepsy. The target mRNAs are used to perform over-representation enrichment analysis of gene ontology and KEGG pathways.

In the obtained results for KEGG and GO analyses only few terms and pathways show striking significance ($FDR_{BY} \leq 0.10$) in the different models and time points. Nevertheless, the presence of many pathways related to etiology of epilepsy (such as endocytosis, actin cyroskeleton remodeling, mTOR etc) support validity of used approach.

Our finding put in focus *ErbB signaling pathway*, that was most enriched at both 24 hours and 28 days as well as between those two time-points in Pilocarpine animals (Table 5.4). Interestingly, a recent report by Li et al. (2011) linked ErbB family (in particular ErbB4 protein) to etiology of epilepsy. Down-regulation of neuregulin 1 (NRG1)-ErbB4 pathway decreased the excitability of fast spiking paravalbumin (FS-PV) interneurons via regulation of voltage-gated potassium channel (Li et al., 2011). Furthermore, the authors showed that mice deficient for ErbB4 receptor in paravalbumin interneurons (Pvalb-cre;Erbb4-/- mice) were more susceptible to seizures after treatment with colvusant agent pentylenetetrazole (PTZ) and pilocarpine. Interestingly, the gene ErbB4 is down-regulated in temporal lobe epilepsy patients which makes ErbB pathway interesting pathway to target for anticonvulsant therapy.

Detailed analysis of *ErbB signaling pathway* and miRNA targeting it raised ten miRNA candidates that of particular relevance to this pathway. These are; *miR-223, miR-184, miR-431, miR-186, miR-142-3p, miR-873, miR-126-3p, miR-149, miR-324-5p* and *miR-31*. The first 7 of these miRNAs are detected significantly deregulated for early onset of epilepsy.

Another pathway, the *actin cytoskeleton regulation* pathway was found enriched at early pilocarpine stage. This finding agrees with previously reported critical role of actin cytoskeleton dynamics and glutamatergic synaptic function in dendritic spine structural integrity and stabilization which in turn leads to the development of recurrent seizures in pilocarpine-treated animals (Ferhat, 2012). Also in kainite model depolymerization of actin and a corresponding changes in dendritic morphology is found to promote seizure (Zeng et al., 2007; Guo et al., 2012). This implicates importance of actin cytoskeletal maintainance as a crucial factor for prevention of seizures.

mTOR and *insulin* pathway were found enriched in both early and late stage in pilocarpine model. Both pathways are implicated in the development of epileptic seizures (Huang et al., 2010; Cho, 2011). *MAPK signaling pathway* was significantly altered in Pilocarpine. Although the downstream targets of MAPK in epilepsy are still unknown, recent studies demonstrated the MAPK activation in animal models of epilepsy. Further *in-vitro* and *in-vivo* rat studies suggested that up-regulating p38 and MAPK could help treat epilepsy (Jung et al., 2010).

Similar pathways were affected at the early stage of disease development in SSSE model. For instance, *ErbB pathway* was found enriched in SSSE model and thus again pointing at the importance of ErbB for the development of epilepsy. In contrast to Pilocarpine model, at late stage of SSSE there were no changes in either of the previously discussed pathways (Table 5.4).

KEGG pathway enrichment analyses for early phases (3 & 6 hours) in 6-Hertz model pointed towards *Natural killer cell mediated cytotoxicity*. Bauer et al. (2008) pointed out that these natural killers (NK) are increased following epileptic seizures in the peripheral blood of TLE patients with hippocampal sclerosis levels (Bauer et al., 2008). *ErbB* and *mTOR* pathway were also enriched at the 6 hours stage in 6 Hertz model. Functional changes disappear with time in 6-Hertz mouse model, i.e. there are no significantly enriched pathways or GO terms at 24 and 72 hours. This corresponds to the low numbers of deregulated miRNA detected by the differential expression analyses of these time-points comparing them to the reference time-point 0 hours.

Further nonspecific pathways such as *pathways in cancer* were also detected as enriched. Only few and general gene ontology terms were weakly significant enriched in Pilocarpine late time-point (Table 5.5). However, relevant GO terms were found enriched in late time-point, these include; *negative regulation of nucleobase-containing compound metabolic process*, *nuerogenesis*, *negative regulation of transcription from RNA polymerase II promoter* and *negative regulation of biosynthetic process*. The findings of the enrichment analyses might help focus the exploration tactics for novel therapeutic applications.

Table 5.4: Pathways enrichemtn analyses results. Significantly enriched KEGG pathways by the miRNA targets in the different comparisons in each epilepcy model and time point at $FDR_{BY} \leq 0.10$.

ID	Term	Size	count	FDR_{BY}
	Pilocarpine-24H vs. Pilocarpine-Naive-24H			
4012	ErbB signaling pathway	61	32	0.013895
4360	Axon guidance	119	35	0,020150
4660	T cell receptor signaling pathway	77	24	0,028920
5215	Prostate cancer	68	22	0,033800
4010	MAPK signaling pathway	174	87	0.041871
4810	Regulation of actin cytoskeleton	143	59	0.043464
4150	mTOR signaling pathway	37	14	0.072392
4144	Endocytosis	140	29	0,076122
4910	Insulin signaling pathway	100	20	0,086693
	Pilocarpine-28D vs. Pilocarpine-Naive-28D			
4350	TGF-beta signaling pathway	60	9	0,040943
5200	Pathways in cancer	220	109	0.048769
4310	Wnt signaling pathway	123	14	0,053270
4012	ErbB signaling pathway	61	22	0.058952
4062	Chemokine signaling pathway	107	32	0,061088
230	Purine metabolism	70	19	0,061382
4120	Ubiquitin mediated proteolysis	83	47	0.061871
5210	Colorectal cancer	46	29	0.070986
4144	Endocytosis	140	31	0,072039
4916	Melanogenesis	67	30	0.076450
4910	Insulin signaling pathway	100	16	0,089920
4150	mTOR signaling pathway	37	12	0.092738

Continued on next page

Table 5.4 – *Continued from previous page*

ID	Term	Size	count	FDR_{BY}
	Pilocarpine-28D vs. Pilocarpine-Naive-24H			
4120	Ubiquitin mediated proteolysis	83	47	0.046015
4010	MAPK signaling pathway	174	66	0.046910
5200	Pathways in cancer	220	109	0.048769
5210	Colorectal cancer	46	29	0.061871
4916	Melanogenesis	67	38	0.063293
4012	ErbB signaling pathway	65	9	0,070366
	SSSE-24H vs. SSSE-Naive			
4810	Regulation of actin cytoskeleton	143	19	0.076620
	SSSE-28D vs. SSSE-24H			
4012	ErbB signaling pathway	61	24	0.044570
5200	Pathways in cancer	220	63	0.057812
	6-Hertz.03H vs. 6-Hertz.00H			
4360	Axon guidance	110	24	0.034053
4650	Natural killer cell mediated cytotoxicity	51	16	0,040189
	6-Hertz.06H vs. 6-Hertz.00H			
4120	Ubiquitin mediated proteolysis	83	39	0.006018
4150	mTOR signaling pathway	37	20	0.019201
4012	ErbB signaling pathway	61	29	0.037442
4650	Natural killer cell mediated cytotoxicity	51	13	0,057200

Table 5.5: Gene Ontology enrichemtn analyses results. Significantly enriched Gene Ontology (GO) terms by the miRNA targets in the different comparisons in each epilepcy model and time point at $FDR_{BY} \leq 0.10$.

ID	Term	Size	count	FDR_{BY}
	Pilocarpine-24H vs. Pilocarpine-Naive-24H			
GO:0080090	regulation of primary metabolic process	1974	386	0.08296
	Pilocarpine-28D vs. Pilocarpine-Naive-28D			
GO:0001702	gastrulation with mouth forming second	14	9	0.092096
GO:0045934	negative regulation of nucleobase-containing compound metabolic process	487	106	0.092096
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	299	70	0.092096
GO:0022008	neurogenesis	315	72	0.092096
GO:0051094	positive regulation of developmental process	245	59	0.092096
GO:0033077	T cell differentiation in thymus	16	9	0.092096
GO:0002053	positive regulation of mesenchymal cell proliferation	26	12	0.092096
GO:0009792	embryo development ending in birth or egg hatching	336	76	0.092096
GO:0009890	negative regulation of biosynthetic process	529	111	0.092096
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	414	90	0.092864
	SSSE-24H vs. SSSE-Naive			
GO:0050794	regulation of cellular process	512	41	0.079812

5.4 Conclusions

The advantage of our study is in overstepping single epileptic model currently used for miRNA screening. This facilitated an approach for systematic comparisons of three different epilepsy models at different time points mimicking distinctive forms and phases of disease development. Adapted normalization methods and a devised procedure for differentially expression analysis for censored data enabled the identification of novel miRNAs and relevant target genes as key players in epilepsy.

Alterations in miRNA patterns were found more pronounced in the animals of chronic epilepsy models in comparison to the 6-Hertz animals. Significantly altered miRNA expression patterns were detected in the early as well as in the late time-points following SE in chronic models. The number of miRNAs co-regulated in both models is much higher in early than in the later stage of disease progression. This might be due to the underlying biological processes taking place in the respective disease progression stage.

In the acute seizure model, deregulated miRNAs represent stress-induced biomarkers changed after a single seizure. Peak of biological response in this model is taking place immediately following seizure and returns back to basal levels comparable to control as time progresses. Deregulated miRNAs sets in the acute model have small to no overlaps to the sets of the chronic models. The low correlations between expression signatures and the small overlapping in the deregulated miRNAs between acute and chronic models clearly delineate chronic models as appropriate epilepsy models. The microarray expression patterns of a number of the overlapping deregulated miRNAs in the chronic epilepsy animals were successfully validated by independent RT-PCR experiments. Indeed, some of the detected miRNA have been already implicated in animal models of epilepsy. However, many of the miRNAs common for both chronic models have not been tested *in-vivo* before.

Enrichment analyses revealed biological processes and pathways that are of great relevance to the disease. These were more pronounced at early/acute phase of disease due to the severity of neuronal changes taking place during the onset. At the later/chronic time point lower number of pathways comes to focus. This can be due to the fact that processes underlying pathophysiology of epilepsy present in later/chronic stages of disease have cumulative effects. This could mean that not only alteration in one biochemical process is responsible for spontaneous recurrent seizures, but rather additive effects of more than one biochemical network contribute to epileptic episodes. Alterations in miRNA signature at late/chronic stages could therefore cause subtle effects that per se together lead to onset of symptoms.

CHAPTER 6

BIOLOGICAL EFFECT SIMILARITIES OF COMPOUND TREATMENTS BASED ON INTEGRATED INFORMATION FROM MULTIPLE SOURCES

6.1 Introduction

Disease and drug perturbations affect biological systems by interacting with different types of molecules leading to regulation of expressions on cellular level (Hopkins and Groom, 2002; Iskar et al., 2010; Schneider and Klabunde, 2013). Perturbations involve modifying drugs in *lead optimization* for better biological activity or changing proteins (e.g. receptor proteins) when adding chemical ligands in order to determine location of binding, affinity of the ligand or the structure of the resulting complex (Brader et al., 1997; Williamson, 2013).

In this chapter a new method for drug-drug similarity assessment based on drug-proteins interaction network and drug pharmacological effects on disease related targets is introduced and applied to detect useful patterns. The integration of interaction information and perturbation data can potentially help reveal how certain compounds work and which proteins are affected (Chu and Chen, 2008). In return, this might aid the difficult and expensive process of drug discovery through drug repurposing or the direct development of novel drugs (Bakheet and Doig, 2009; Iorio et al., 2010; Wang et al., 2013a).

Given the large number of targets and the multitude of drug-like chemical (synthetic small molecules), enzymes, antibodies, organic- & inorganic-derived and oligopeptide etc, the drug-target space is immense. It is estimated that around 3M human proteins are potential drug targets (Hopkins and Groom, 2002; Russ and Lampel, 2005; Overington et al., 2006; Li and Lai, 2007; Landry and Gies, 2008). ChEMBL alone has bioactivity evidences for over 1.4 M distinct compounds with around 13 M bioactivity evidences (Gaulton et al., 2012). Older computational methods aiming to predict potential drug-target interactions can roughly be classified into ligand-based approaches and protein-based approaches. The former is based on *similar properties* principle, i.e. compounds of similar structure usually share similar physicochemical properties and biological activities (Cramer et al., 1978; M. Johnson, 1990). The latter, the *targets first* principle approach, focuses on protein-ligand docking.

Structure-based analysis of proteins exploit structural and physicochemical protein properties (Hopkins and Groom, 2002; Orth et al., 2004; Russ and Lampel, 2005; Keller et al., 2006; Bakheet and Doig, 2009) in order to predict protein druggability (Nayal and Honig, 2006; Cheng et al., 2007; Mehio et al., 2010; David Andersson et al., 2011). However, this approach requires protein structure, which is not available for many proteins (e.g. membrane proteins). Moreover, druggable proteins are not necessarily drug targets. Ligand-based methods like QSAR (Quantitative Structure-Activity Relationships) are based on *similar properties* principle and aim to predict binding and ADME-Tox qualities of compounds from its chemical structure (Cramer et al., 1978). These methods compare candidate compound to known drugs of certain target(s) e.g. using machine learning method (Butina et al., 2002; Byvatov et al., 2003) or partial least squares (PLS) (Cramer et al., 1988, 1989). However, compounds with different structures might nevertheless have similar effects and vice versa. Moreover, most of the physico-chemical properties and quantum-chemical descriptors do not fully describe the structural characteristics of compounds (Bajorath, 2014). Furthermore, the performance of these methods decrease rapidly with decreasing number of known drugs for a target protein (Wang et al., 2013a).

Further statistical methods have been proposed that use the drug-target space and pharmaceutical information e.g. in a graph-model settings (Cheng et al., 2012), using supervised and semi-supervised learning methods (Yamanishi et al., 2008; Bleakley and Yamanishi, 2009; Xia et al., 2010), utilizing chemical structure topology (Keiser et al., 2007) or utilizing genomic sequence and pharmacological effect information (Yamanishi et al., 2010). However, these methods either ignore protein interaction information and/or require compound & protein structure. Moreover, missing drug-target relationships are treated as negative samples. Gottlieb et al. (2011) suggested a large-scale prediction method for compound indications using multiple drug-drug and disease-disease similarity measures.

Methods that rely on chemical structure similarity are thought to produce less sensitive and accurate search results as these similarities are overly conservative and usually put much weight on irrelevant details. It is also important to understand that a chemical ligand's effect in the living system is not limited to direct target proteins. Much more chemical activation or inhibition of a particular protein target can alter downstream signaling cascades and transcriptional responses (Schneider and Klabunde, 2013).

Hence, a new line of research focuses on the use of treatment effects information. For instance, side effects information of marketed drugs gained from text mining data have been used in many methods to predict cooperative effects (Campillos et al., 2008; Kuhn et al., 2010; Lee et al., 2011; Brouwers et al., 2011; Yang and Agarwal, 2011; Hurle et al., 2013; Wang et al., 2013a,b; Ye et al., 2014; Iwata et al., 2015). Other methods used transcriptional responses following treatment information from array data (e.g. from the Japanese Toxicogenomics Project (Uehara et al., 2010) or the *Connectivity Map* (Lamb et al., 2006)) to predict novel drug indications (Hu and Agarwal, 2009; Iorio et al., 2010; Iskar et al., 2010; Sirota et al., 2011; Qabaja et al., 2014; Zhao et al., 2014). However, side effect information is limited to approved drugs where side effects have been sufficiently assessed, and treatment response data are noisy, inconsistent and limited to few hundreds of compounds and a couple of specialized cell lines.

In this work a novel similarity measure for chemical compounds, the Biological Effects Similarity (BES), is presented. This measure captures treatment induced biological effects via a large number of target protein features. A consensus clustering (Monti et al., 2003) is performed based on the BES measure to identify statistically stable compound sub-groups. This large-scale computational framework can be used to predict approved and novel molecules and does not explicitly require knowledge about protein or ligand structures.

An additional feature of the BES is the direct integration of compound-target affinities. Hence, it is in particular applicable for unsupervised, exploratory analysis of compound collections with heterogeneous binding affinities, focusing e.g. on detection of compound sub-groups with similar induced biological effects. This situation is exemplified here by consensus clustering of almost 4,700 compounds, which have been screened against protein targets that have been associated to HIV and cancer. Both diseases comprise a large number of associated protein targets and accordingly tested compounds. Consensus clustering with our BES measure was able to identify well separated and statistically stable compound clusters with several so far untested compound-target combinations. BES clustering thus suggests to experimentally explore novel compound-target combinations.

6.2 Material and Methods

6.2.1 Dataset

This study focuses on compounds targeting proteins, which have been associated to cancer or HIV. These are two relevant diseases with sufficiently high number of associated targets and compounds being developed. Both diseases are seemingly quite different, and communalities between compounds with respect to induced biological effects are thus not a priori expected, but cannot be excluded either. DisGenNet (Bauer-Mehren et al., 2011) is used to retrieve disease related human proteins for both, HIV and cancer (accessed 12/2013). These were then linked to pharmacological data from ChEMBL® (Gaulton et al., 2012) and DrugBank® (Law et al., 2014), yielding 60 human target proteins putatively related to HIV and 405 proteins (with available Entrez gene IDs) putatively related to cancer.

Altogether there were 74,606 compounds with 115,007 compound-target indication in HIV and 431,566 molecules with 1,199,265 indications in cancer. Only biological activities assessed by the *inhibition constant* (Ki) and measured in saturation or competitive binding assays are considered. Methods for assessing compound's biological activity towards target protein are explained in section 2.4. Compounds with no or very weak activity ($Ki > 10^{-}6M$) are removed. Moreover, compounds for which no MACCS fingerprints could be computed are omitted (see subsection 6.2.6). Target specific biological activities (Ki values) of compounds were standardized to affinity values (subsection 6.2.3).

Thus, 511 compounds for HIV and 4,299 for cancer are considered at the end. Altogether the compounds targeted 22 proteins in HIV and 198 proteins in cancer. They showed an overlap of 154 compounds and 7 targets. The complete dataset contains 4656 unique compounds (containing 122 FDA approved drugs according to DrugBank®) comprising 72,925 interactions with 213 target proteins. The whole dataset can be visualized as a large scale compound-target network (Figure 6.1).

6.2.2 Biological Effects Similarity (BES)

General Idea Our novel BES measure has two major aspects: First, protein-protein similarities are calculated based on several information sources and then combined into one consensus (Fig. 6.2). The second aspect of the approach is the integration of compound-target affinities, which can be retrieved from public databases such as ChEMBL (Gaulton et al., 2012). In this context, compound-target relationships and relationships among proteins can be conceived as a network comprising two

interconnected graphs. One graph is a directed, bipartite graph indicating connections between compounds and targets. This graph is typically sparsely connected as potency information is not available for every compound-target combination (Fig. 6.2). The other one is an undirected complete graph that describes similarities between each pair of proteins.

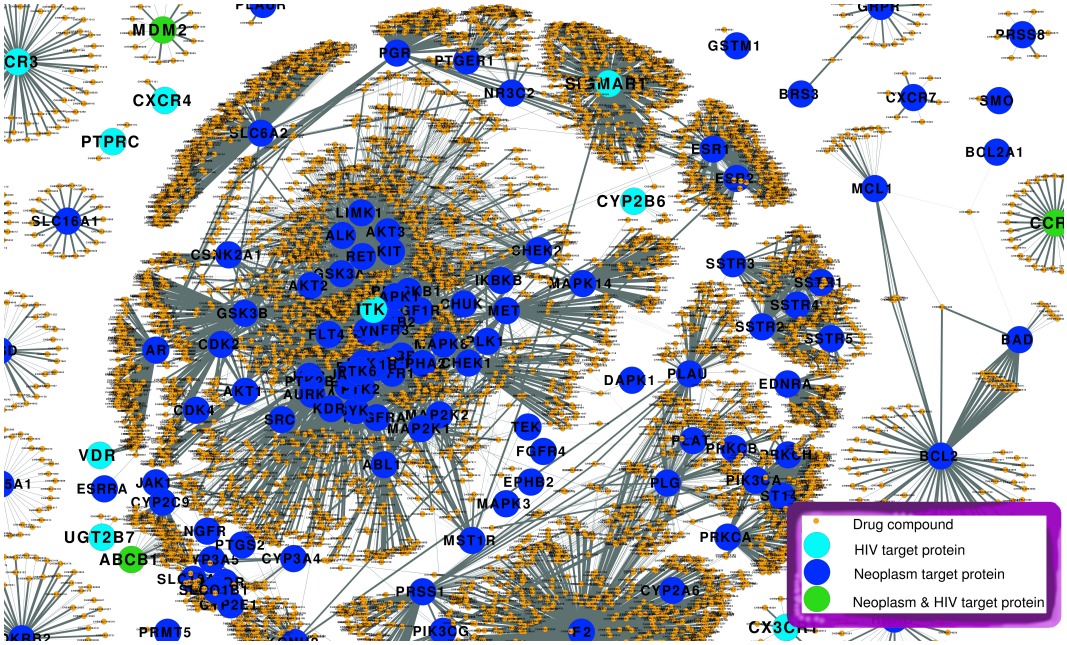

Figure 6.1: Compound-target network. Partial view of compound-target network used in this study: Bioactivities are visualized by edge thicknesses.

The BES algorithm begins by assigning weights to edges in the combined graph. An edge between a compound and a target protein in the first graph corresponds to an appropriately scaled compound potency value (see subsection 6.2.3). An edge between any pair of proteins in the second graph is weighted by the consensus similarity from several information sources (Praveen and Fröhlich, 2013). The intuition behind the approach is to describe biological effects similarity of a compound pair by the biological similarities of their targets, which is weighted by the affinities of each compound to its targets. While doing so it has to be taken into account that one and the same protein can be targeted by both compared compounds at the same time. The method is described in detail in the following passage.

Step 1: Similarity of Protein Targets Let T_i be the set of proteins targeted by compound c_i and T_j be the set of proteins targeted by compound c_j. For each protein pair $(p, q) \in T_i \times T_j$ the similarity with respect to several annotations is calculated, namely:

- Similarity of Gene Ontology terms (biological processes)
- Similarity of protein domains
- Path distance in protein-protein interaction network and canonical pathways
- Potentially interacting protein domains

Details about the way in which this is done are described later. The result is a set of $\mathscr{L}$ similarity matrices of size $\mid T_i \mid \times \mid T_j \mid$. These matrices are combined into one target similarity matrix S using the Borda counts approach described in Marbach et al. (2012). Values in S are always between 0 and 1. Notably, the intersection between T_i and T_j can be non-empty and thus a value of 1 can appear at any position in S.

When integrating data from different information sources a major difficulty is the missing annotation information in one or more of the sources. How this problem is addressed is explained in section subsection 6.2.4.

Step 2: Weighting with Compound Affinities Compounds c_i and c_j may have rather different affinities to their targets. It is assume that these affinities have been sufficiently re-scaled to the interval $[0, 1]$ in a pre-processing step (subsection 6.2.3). Let us denote the vector of affinities to proteins T_i and T_j by , $\overrightarrow{a}_i \in \mathbb{R}^n$ and $\overrightarrow{a}_j \in \mathbb{R}^m$, respectively. The affinity weighted $n \times m$ similarity matrix between protein targets can then be computed as

$$BES^0 = (\overrightarrow{a}_i \otimes \overrightarrow{a}_j) * S, \tag{6.1}$$

where $\otimes$ and $*$ denote the outer and element-wise products, respectively.

Step 3: Integration into BES Matrix BES^0 consists of protein target similarities, which are weighted by products of compound affinities. These values need now to be combine into the desired Biological Effects Similarity (BES) measure. The intuition behind the approach is that; any given $p \in T_i$ is matched with $p \in T_j$ to which it is most biologically similar and most similar with respect to compound affinities.

Thus, highest possible similarities (i.e. best possible "matches") of affinity weighted target pairs $(p, q) \in T_i \times T_j$ are considered. In other words, the row and column maxima in BES^0 are separately computed. The arithmetic mean of all these $m + n$ similarity values is then the *raw* BES value (Figure 6.2).

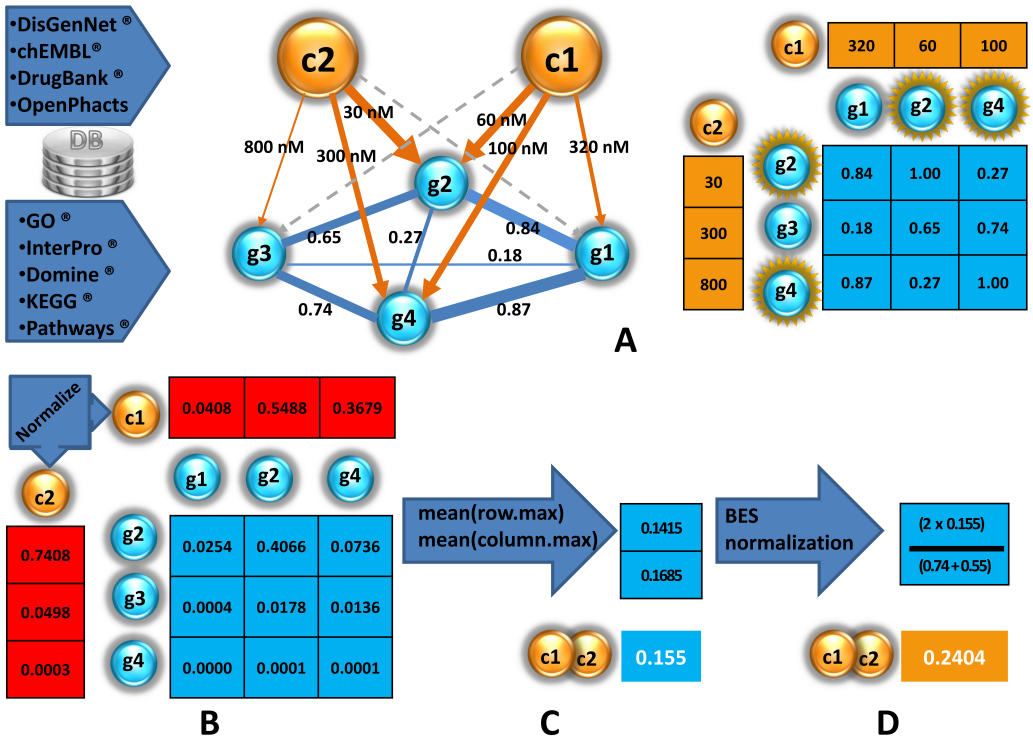

Figure 6.2: Example calculation of Biological Effects Similarity (BES) for two compounds c1, c2: First step: Similarities of all target proteins are integrated in a probabilistic manner on the basis of different biological information sources (A). Second step: Target similarities are weighted by compound affinities (bioactivities standardized to $[0, 1]$) (B). Third step: The BES is computed based on best matches of affinity weighted protein target pairs (C). Fourth step: The BES is normalized to $[0, 1]$ (D).

Step 4: Normalization of BES The raw BES is not guaranteed to lie in the interval $[0, 1]$, and self-similarities can be different from one. Therefore, a suitable normalization of the raw BES values is needed. In this context the normalization used is defined as:

$$BES.norm(c_i, c_j) = \frac{(2 \times BES(c_i, c_j))}{(BES(c_i, c_i) + BES(c_j, c_j))}. \tag{6.2}$$

6.2.3 Drug-Target Affinities

The efficiency of the method depends not only on the protein confidence relationships and the compound-target bioactivity values, but also it depends on the chosen normalization method for the bioactivity values. Ki values are used here to describe compound-target affinities. Raw affinity values are first scaled to the same concentration unit (e.g. nM). The Ki values are normalized to the range $[0, 1]$ via:

$$norm(Ki_i) = \frac{e^{-Ki}}{e^{-min_i(Ki_i)}} \tag{6.3}$$

where min_i runs over the set of all Ki values in the dataset. Higher normalized affinity values thus indicate higher potency of the compound for the respective target. Because the chosen concentration unit can numerically affect the normalized affinity, a number of concentration units for Ki bioactivity values are investigated (Figure 6.3 right).

The scaling adopted here seems to give most favorable characteristics: Scaling bioactivity values to $10^{-3}nM$ yields an affinity curve that appropriately reaches a plateau of zero affinity approximately in the middle of Ki value range that defines weakly potent compounds (Figure 6.3).

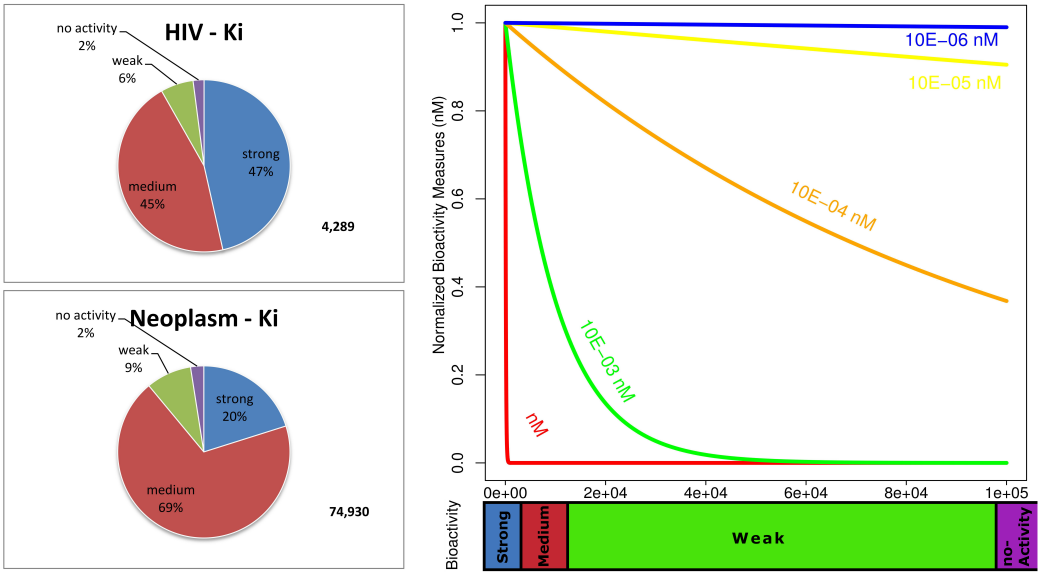

Figure 6.3: (Left plots) Overview of HIV and cancer compounds and their bioactivity levels. Compounds are stratified by disease and activity classes, which are commonly used in pharmaceutical research (Roider et al., 2014): $Ki \leq 10^{-8}M$ **Strong**, $10^{-8}M < Ki \leq 10^{-6}M$ **Medium**, $10^{-6}M < Ki \leq 10^{-5}M$ **Weak** and $Ki > 10^{-5}M$ **No-activity**. (Right plot) Standardization of compound affinities with different concentration units. Compounds Ki concentrations in nM are represented in X-axis, Y-axis represents the normalized bioactivity (affinity) $[0, 1]$. The scaling represented by the green curve to $10^{-3}nM$ is used, as it appropriately reaches a plateau of zero affinity in mid of the class "Weak" (according to classification of compounds based on bioactivity strength)

6.2.4 Biological Similarity of Compound Targets

The question, how to quantify the similarity between two compound targets, can be answered from different points of view. Accordingly different features can be taken into consideration. Here, annotations with respect to pathways (KEGG, PathwayCommons - (Kanehisa and Goto, 2000; Cerami et al., 2011)), biological processes (Gene Ontology - (Harris et al., 2004)), protein domain annotation (InterPro - (Mulder et al., 2002, 2007; Hunter et al., 2012)) and protein domain interactions (DOMINE - (Yellaboina et al., 2011)) are employed.

For GO annotation (biological processes) the default similarity measure for gene products implemented in the R-package GOSim was used (Fröhlich et al., 2007), based on the information theoretic GO term proximity measure proposed by Lin

(1998). Furthermore, protein domain annotations obtained from DOMINE were compared on the basis of a binary vector representation via the cosine similarity. The relative frequency of interacting protein domain pairs was considered as another scoring measure for compound target similarity. Finally, network information from pathway databases was integrated by computing shortest path distances between pairs of proteins. For that purpose KEGG pathways were converted to a graph structure using KEGGGraph (Zhang and Wiemann, 2009) and then joined into one directed network. Similarly a large scale protein-protein interaction graph was constructed form PathwayCommons. The idea behind the use of shortest path distances and protein domain interactions was that targeting closely interacting proteins should increase the chance to induce similar biological downstream effects.

Effectively, all considered features (pathways, GO, protein domains and protein domain interactions) yielded a separate similarity matrix. To integrate these different similarity matrix into a consensus, the rank based approach proposed in the work of Marbach et al. (2012) is applied. Importantly, the method in principle allows for handling missing values in one of the used similarity matrices, which can happen due to incomplete annotation: For each pair of protein targets only completely available features were considered. However, the set of available features could in principle differ between protein pairs. For example, KEGG pathway annotation may be available for protein pair (p1, p2), but not for both (p1, p3). Hence, for pair (p1, p3) KEGG annotation has no numerical influence.

6.2.5 The Consensus Clustering

The consensus clustering approach proposed by Monti et al. (2003) is used (see methods in subsection 3.4.2). The number of iterations was set to 200, and in each of these iterations a complete linkage clustering was conducted on sub-samples of size 80% of the data.

The area under the CDF (AUC) is then reported as a measure of clustering stability. Monti et al. (2003) propose the AUC as a method to select a suitable number of clusters. However, in our implementation it was not always easy to determine the maximum relative increase in AUC. Empirically the AUC was not always a monotonic increasing function in the application example. We therefore determined the number of clusters by the following criterion; higher AUC values correspond to higher number of clusters until the AUC reaches a first maximum and is not increasing in 3 subsequent steps (see subsection 3.4.2 for more details).

After the number of clusters have been determined the quality of the produced clustering is investigated independently by silhouette plots (Rousseeuw, 1987). The

resulting cluster groups are investigated through enrichment analyses of individual clusters. More specifically, this is done by looking for statistical overrepresented GO terms, KEGG pathways, protein domains and sequence motifs (subsection 6.2.8).

Apart from consensus hierarchical clustering, the *affinity propagation* based clustering is used (Frey and Dueck, 2007). This is a method which based on the concept of "message passing" between data points and tries to determine a good cluster number in an automated fashion.

6.2.6 Chemical Structure Similarity

Similarity searching techniques to aid the ligand-based virtual screening of large compound collections are usually derived solely from compounds chemical structure. Descriptors such as fingerprints, which can reflect spatial relationships between features of a molecule, together with different similarity measures can be used to asses similarities between compounds. If chemical structure are available for the set of compounds related to HIV and Cancer then compounds chemical structure coded in Simplified Molecular-Input Line-Entry (SMILE) specification system can be queried via the pubchem http protocol. 166 bit MACCS finderptings of compounds are then computed based on their SMILES after parsing. Given dichotomous MACCS fingerprints, pair-wise compound similarities can be computed using distance metrics defined for binary strings. The frequently employed Tanimoto-Jaccard index T_{AB} $[0, 1]$ was chosen to assess molecular fingerprints similarities between compounds. For dichotomous fingerprints it is defined as:

$$T_{AB} = \frac{c}{a + b - c} \tag{6.4}$$

where a and b denote number of bits in each of the two fingerprints, and c is the number of common bits.

6.2.7 Maximum Common Substructure Analysis

Maximum common subgraph isomorphism (MCS) is a graph-based similarity concept used to identify the largest substructure (subgraph) shared among two or more compounds. More formally, given two graphs $G1$ and $G2$ representing the structures of two compounds MCS tries to find largest usually connected subgraph of $G1$ isomorphic to subgraph of $G2$. This problem is known to be NP-hard, for at least k isomorphic vertices it is NP-complete (Karp, 2010). The problem is usually solved

by finding cliques in the product graph. The maximum found clique corresponds to largest induced subgraph of graphs $G1$ and $G2$.

In this study MCS was used to investigate common structural properties of compounds being grouped together in one cluster. The FMCS® algorithm (`https://bitbucket.org/dalke/fmcs`) in the paython® tool which successively searches MCS of each compound pair in a cluster. This algorithm employs a back-tracking strategy to enumerate pairwise MCSs. A branch & bound procedure limits the search space. Default settings and a maximum structure match threshold of 0.7 were used. ChemMine® (`http://chemmine.ucr.edu` - (Backman et al., 2011)) was taken to visualize the results.

6.2.8 Enrichment Analyses

Enrichment Analyses of Gene Ontology Terms and Biochemical Pathways

The R-package GOstats (Falcon and Gentleman, 2007) was employed to look for overrepresented biological processes and KEGG pathways of targets in individual clusters and in the whole dataset. GOstats uses a hypergeometric test, and additionally incorporates parent-child relationships between GO terms. To control the false discovery rate in multiple testing, the p-value was adjusted using Benjamini and Yekutieli (2001) method. Gene set enrichment analysis methods are explained in section 3.5.

The enrichment was investigated with respect to all human genes. In order to assess the true significance of individual terms for a defined cluster, the same analysis for the whole dataset is also conducted. Terms being significant in the whole dataset were not further considered within a defined cluster (details in appendices C; Figure C.1, Figure C.2, Figure C.3, Table C.2 and Excel tables).

Protein Domain Enrichment Analyses

Protein domains are retrieved from the *InterPro* database (Mulder et al., 2002). The InterPro database integrates predictions from multiple diverse source repositories and provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites. The full human genome wide proteome was retrieved using the org.Hs.eg.db annotation package, which contains 43827 unique Entrez Gene identifiers. The InterPro domains annotated to each of these genes were retrieved using the biomaRt interface (Durinck et al., 2009). Accordingly, 7131 unique InterPro domains were found in the full human genome. A

hypergeometric test was performed to statistically evaluate the significance of the enrichment (see details in subsection 3.5.1). To control the false discovery rate in multiple testing, the p-value was adjusted using the Benjamini and Yekutieli (2001) method.

Once again enrichment analysis was first done with respect to the whole human genome, and then corrected with respect to the enrichments founds in the whole dataset (appendices C; Figure C.4, Table C.2, and Excel tables).

Motif Enrichment Analyses

A wrapper is programmed that uses the MEME Suite (Bailey et al., 2009) which is a tool for discovering motifs in a group of related protein sequences. Assuming no constrains or pre-knowledge about the motifs looked for, this method applies a de novo motif discovery, in which short subsequences that occur more frequently than would be expected by chance are identified. That information is then used to build a Position Weight Matrix (PWM) describing the motif. Note that the de novo discovery phase is not performing an exhaustive search. Much more it uses heuristics to make some good initial guesses about which subsequences are likely to be instances of a motif, it then adjusts and optimizes the subsequences considered and the components of the PWM. The next step uses the PWM to search for hits in a list of given sequences, identifying statistically significant matches to the given motif.

The wrapper also included the MEME Suite tool called FIMO (Grant et al., 2011), that searches a sequence database for occurrences of user provided motifs, treating each motif independently. The program uses a dynamic programming algorithm to convert log-odds scores into p-values, assuming a zero-order background model. The p-values for each motif are then converted to q-values following the method of Benjamini and Hochberg (1995). Using both tools, it is possible to conduct an enrichment analysis using a hypergeometric test (subsection 3.5.1). It has to be stressed here that motifs have no function of their own; function arises exclusively out of context in the mammalian system. Therefore, interpretation of these entities should be seen on the view of disease peculiarity and the suggested drugs.

As before, enrichment analysis was done with respect to the whole human genome and then corrected with respect to the enrichments founds in the whole dataset (appendices C; Figure C.5, Table C.2, and subsection C.2.1).

6.3 Results

6.3.1 Characterization of Target Proteins

As an initial step protein targets in the employed dataset are characterized with respect to statistically enriched KEGG pathways, GO terms, protein domains and sequence motifs (at cutoff $FDR_{BY} \leq 0.05$). Target proteins were, for example, enriched for *protein kinases* (including tyrosine kinases), *chemokine receptors*, *GPCRs*, *nuclear hormone receptors*, *HDACs* and *immunoglobulins*. They were mainly involved in *signaling*, *metabolic processes*, *immune response*, *regulation of cell growth* and *angiogenesis*. While *dys-regulation of cell growth* and *angiogenesis* are known be important in cancer development (Birbrair et al., 2013), HIV is causing an immune deficiency. Hence, these findings are in agreement with common knowledge. KEGG pathway analysis yielded statistically significant associations to several cancer types as well as *apoptosis*, *chemokine signaling*, *cytokine-cytokine receptor interaction* and *epithelial cell signaling in helicobacter pylori infection* (appendices C Excel files). *Chemokine signaling in conjunction* with other processes, such as *cytokine-cytokine receptor interactions*, *promotes cellular morphology changes through transmembrane receptors* trigger prevention of HIV-1 infection (Berger et al., 1999; Lachgar et al., 1998; Mellado et al., 2001; Wu, 2010). *Helicobacter pylori* inhibits HIV in CD4 T cells by producing VacA toxins (Cover and Blanke, 2005).

A consensus clustering of the joint set of all protein targets based on their biological similarities was also performed based on the approach explained in subsection 6.2.2. This resulted in 37 clusters with an AUC of ~ 0.95 and silhouette widths that indicate stable and fairly well separated clusters (Figure 6.4 C1 & C2). Notably, consensus clustering, which is based on a re-sampling strategy, focuses on clustering structure that is statistically stable detectable.

Interestingly, the cluster analysis did not separate the two diseases: Clusters 1 – 11 contained targets from both diseases, the others comprised only cancer proteins (Appendices C Excel files). The functional analysis of these sets of proteins confirmed their relatedness. For example, Cluster c01, which is apparently related to cancer, contains 18 cancer targets (e.g. SRD5A1, SRD5A2, SLC5A1, SLC16A1, SLCO1B1) and a one HIV protein (SIGMAR1). Cluster c02 contained 10 proteins from both diseases (HIV: CCR3, CX3CR1, CXCR1; cancer: CCR1, CXCR3, LPAR1, GRPR, LHCGR, BDKRB2, SSTR2). These are exclusively all receptor proteins of which 5 are different types of chemokine receptors. All 8 proteins in cluster c03 are subfamilies of the cytochrome polypeptide (cancer: CYP2A6, CYP2C9, CYP2C19, CYP2E1, CYP2J2, CYP3A4, CYP3A5; HIV: CYP2B6). Cluster c09 contains 19 proteins of which 7 are MAP kinases (MAPK, MAP2K) and targeted by cancer compounds.

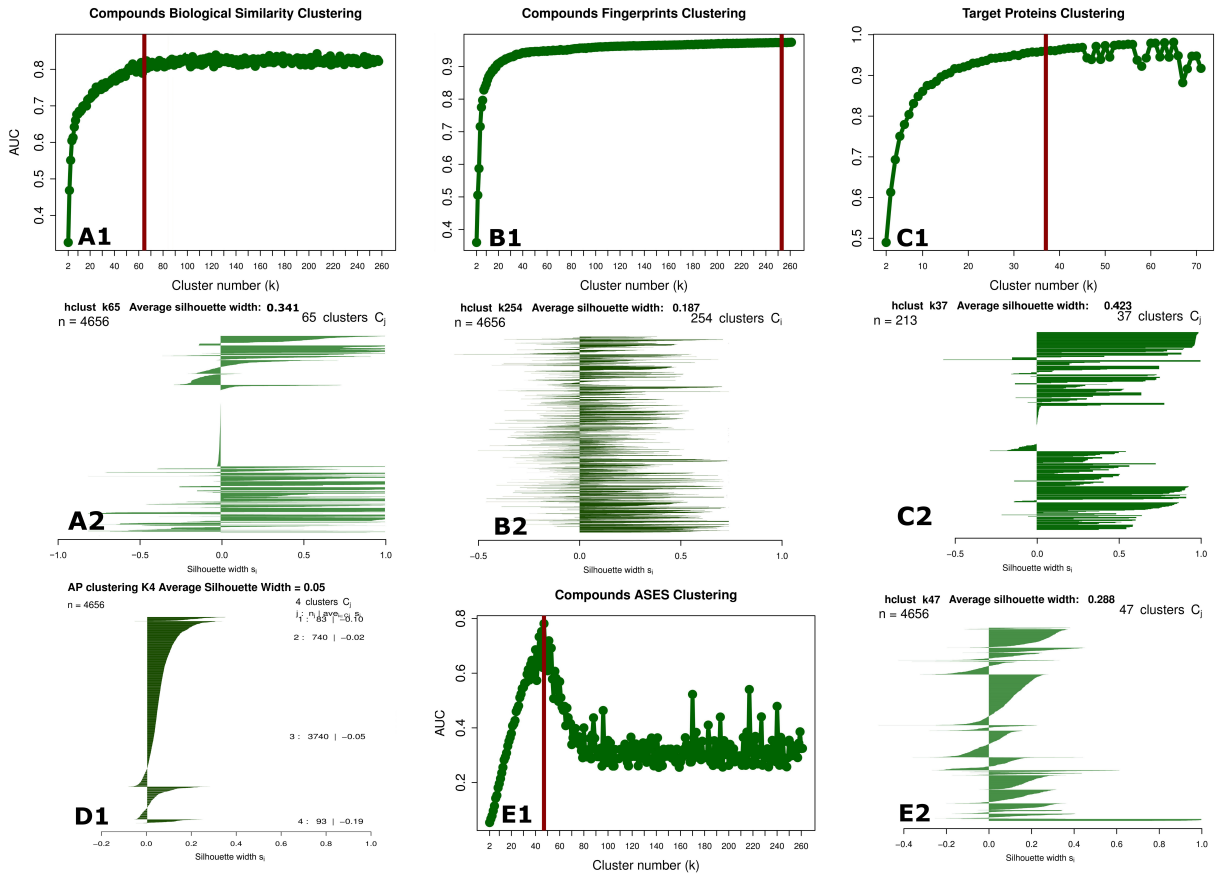

Figure 6.4: Area Under the Curve (AUC) & silhouette plots for complete-linkage consensus clustering of compounds based on BES (A1, A2) and based on chemical fingerprints similarities (B1, B2). Plots C1 & C2 show AUC & silhouette widths for complete-linkage consensus clustering of target proteins themselves based on protein similarities. Silhouette plot for affinity propagation clustering of compounds based on BES is shown in D1. Plots E1 & E2 show AUC & silhouette widths of clustering of compounds (subset) using complete-linkage consensus clustering based on Adverse Side Effect Similarity (ASES).

6.3.2 Validation of BES with Gene Expression Data and Comparison to Existing Similarity Measure

In order to validate the developed BES as a measure for biological effect similarities of compounds, the agreement of the results with gene expression data is checked. For that purpose data from the *Connectivity Map* are downloaded (Lamb et al., 2006). There were 215 compounds found in our dataset as well as in the *Connectivity Map*. In the *Connectivity Map* gene expression for each of these compounds has been profiled in a concentration dependent manner in different cell lines. Using the "limma" (linear models for microarray analysis) method (Smyth, 2004a) gene expression data are modeled using a two-way ANOVA model comprising a cell line and a concentration effect. Based on this model the compound specific treatment effect was estimated using the rest of the data (not treated with the respective compound under consideration) as controls. The result was a t-statistic for each gene and compound. The absolute value of the t-statistics of the top 250 genes is used as a feature vector for each compound. These feature vectors were compared via the cosine similarity. Consequently, a gene expression profile similarity for compound pairs is achieved, which is compared against the proposed BES measure. This resulted in a highly significant ($P\text{-}value < 1E-16$) Pearson correlation of ~ 0.51, indicating a general good agreement of both measures.

For comparison the agreement of the network based adverse side effect similarity (ASES; (Brouwers et al., 2011)) to gene expression data is assessed. The authors of the ASES method used a confidence weighted functional protein-protein interaction network compiled from STRING (Jensen et al., 2009) to predict adverse side effects of compounds via the closeness of targeted proteins in the network. In contrast to our BES the ASES measure does not include bioactivity information. The ASES could be computed for 215 compounds tested in the *Connectivity Map* database, yielding a lower correlation of ~ 0.28 to gene expression data. Hence, the BES agrees better with similarities of transcriptional downstream effects than ASES. The direct correlation between ASES and the proposed BES measure was ~ 0.2.

6.3.3 Influence of Different Features on BES

The BES is influenced by the way in which biological similarities of compound targets are assessed. Hence, the different influence factors (KEGG and PathwayCommons pathways, biological processes, protein domains) are investigated in more detail. This was done by correlating the entries of the BES similarity matrix considering only one of these factors with the combined BES similarity matrix (Figure 6.5). This demonstrated a high influence of GO and pathway information from KEGG

and PathwayCommons (Pearson correlation ~ 0.75). Information from protein domain annotation (InterPro) and protein domain interactions (DOMINE) have comparably lower positive correlations with the combined BES matrix (~ 0.5 and ~ 0.6, respectively).

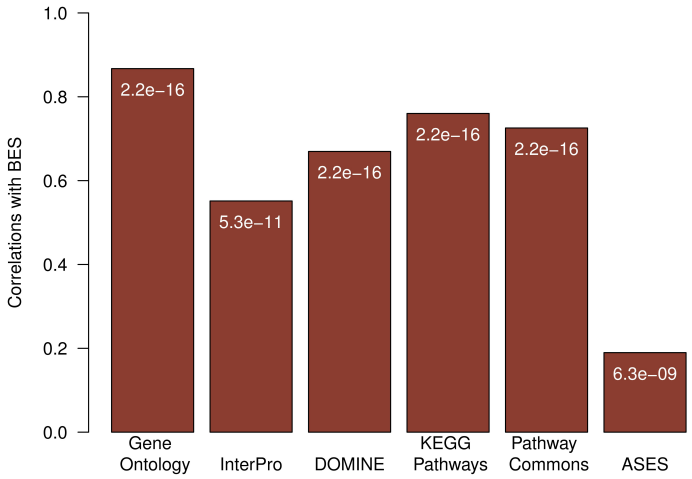

Figure 6.5: Pearson correlation coefficient of BES with BES calculated for one particular information source and ASES, respectively. The significance (p-value) of each correlation is shown on top of each bar.

6.3.4 Application of BES for Compound Consensus Clustering

After the BES measure have been evaluated, it is applied within the earlier described consensus clustering algorithm to explore the dataset with respect to potentially existing compound sub-groups. This resulted in 65 clusters, which are called the *BES clusters* in the following. The BES clustering was highly stable with an AUC of ~ 0.82 and resulted in relatively well separated clusters with an average silhouette width of ~ 0.34. Around 50% of the clusters showed a silhouette width larger than 0.5, many even close to 1 (Figure 6.4, A1 & A2 and appendix Table C.1). Four of the 65 clusters were singleton clusters. Apart from one large cluster (1,944 compounds) clusters contained between 2 and 324 compounds. The large cluster contained compounds, which in most cases were found to have zero or very small BES similarities, indicating biologically divergent effects.

The four singleton compounds have been screened against far fewer target proteins than the rest of the compounds in the dataset, and targets are limited to certain

protein classes (membrane receptors and enzymes). Moreover, bioactivities of singleton compounds are often rather weak. Compound CHEMBL1977148 has been screened only against enzymes and mostly with no or weak bioactivity. Compound CHEMBL1081312 has been measured for a limited number of enzymes (69) with few medium bioactivities. Compound CHEMBL435523 has been only screened against 3 transcription factors showing strong bioactivities. The activity of compound CHEMBL434159 has been tested for 5 proteins (Somatostatin receptor 1-5), where the bioactivity was weak.

In order to assess the influence of the selected clustering algorithm on the results, the cluster analysis is repeated using affinity propagation as described in Frey and Dueck (2007) and using the ASES dissimilarity measure as described in Brouwers et al. (2011). Applying consensus clustering on the ASES dissimilarity matrix produced a lower number of clusters (47 clusters) than our BES clustering. The AUC of 0.8 demonstrated roughly comparable stability (Figure 6.4 E1), however, the average silhouette of 0.29 is significantly lower (Figure 6.4 E2). The overlap of the ASES clustering with the BES clustering is small (only 6 clusters had a relative overlap $\leq 25\%$). Thus, it is concluded that the ASES similarity can also produce statistically stable clusters although it does not involve bioactivity information. However, ASES clusters are of low quality and quite different than the BES cluster. The affinity propagation method resulted in significantly worse silhouette plot (Figure 6.4, D1). Therefore, the focus is put on consensus clustering approach in the rest of this work.

6.3.5 Protein Targets of BES Clusters show Enrichment of Biological Pathways, Processes, Protein Domains and Sequence Motifs

To analyze individual BES clusters further, statistical enrichment analysis were conducted of their target proteins with respect to KEGG pathways, GO terms, protein domains and sequence motifs using a hyper-geometric test. It is worth mentioning that neither sequence motifs nor GO terms from the molecular function category are integrated in the BES measure.

Most compound clusters revealed a statistically significant enrichment of specific biological pathways, processes, protein domains and sequence motifs according to this analysis ($FDR < 5\%$; details in appendices C; Figure C.1, Figure C.2, Figure C.3, Figure C.4, Figure C.5 and Excel files. Table C.2 shows the top 3 vocabularies form each of the categories in the *repurposing clusters*). Notably, significant categories, which also showed up significant in the overall dataset, are excluded.

In most cases sequence motifs, GO terms and protein domains were found enriched, in fewer cases also KEGG pathways. For example, targets of compounds in cluster 1 are strongly enriched in the *renin-angiotensin system*. Clusters 57 and 61 are related to *Wnt signaling*. However, cluster 57 shows an enrichment of RAS, C2, PIK and PKC binding domains, whereas cluster 61 targets are overrepresenting only a subset of these domains, namely C2 and PKC.

Mining the literature for the individual target proteins revealed useful information (details in appendices C Excel tables). Cluster 7, for instance contains many tyrosine kinases that act as cell-surface receptors for fibroblast growth factors and play an essential role in the regulation of embryonic development, cell proliferation, differentiation and migration (Turner and Grose, 2010). Cluster 15 contained a number of nuclear hormone receptors which are ligand-activated transcription factors that regulate eukaryotic gene expression and affect cellular proliferation and differentiation in target tissues (Szanto et al., 2004).

6.3.6 Cluster Analysis Suggests Novel Compound-Target Pairs

There were several BES clusters, in which protein targets were jointly associated to HIV and cancer. Fifteen of them (clusters 2 - 7, 11, 13, 15 - 21, appendix Table C.1) have medium or high silhouette width. These are potentially interesting clusters, specifically, if certain compound-target pairs have not been tested so far. These clusters will be called *repurposing clusters* hereafter. The compounds in these clusters are further investigated by organizing them into three groups:

- those which have been tested for proteins associated to both diseases with strong or medium bioactivities (Table 6.1),
- those which show a strong or medium bioactivity for targets associated to one disease and weak or no activity for targets associated to the other disease (Table 6.2), and
- those compounds, which have only been tested for targets associated to one of the diseases so far (Table 6.3).

Specifically the third group of compounds represents compounds, which could contain interesting candidates for follow-up screening experiments. The latest version ChEMBL (accessed 02/2015) is used to validate some of these candidates. Different assay data (IC50, relative inhibition, AC50) are used. This is different than the one used for our clustering study, hence these data can be seen as independent. Our analysis allowed us to confirm several of our initial findings (for complete lists see; Table 6.1, Table 6.2 & Table 6.3). For example, compound CHEMBL232656

(BAX-471/BAX-741) in cluster 3 has originally only been tested against several cancer associated targets. However, according to the BES clustering it falls together with several other compounds, which have been tested against chemokine receptors, which themselves have been related to HIV infection (Berger et al., 1999; Lachgar et al., 1998; Mellado et al., 2001; Wu, 2010).

Hence, our clustering suggests testing CHEMBL232656 for chemokine receptors. Indeed a strong bioactivity for different types of chemokine receptor (IC50 = $1nM$) has been reported for this compound recently (Pease and Horuk, 2014). Similarly, CHEMBL9298 (FADROZOLE), CHEMBL1444 (LETROZOLE) and CHEMBL1399 (ANASTROLE) in cluster 11 have so far only been tested against cancer associated targets, but the clustering suggests screening them against cytochrome P-450 and estrogen synthase, which are HIV related. The strong bioactivity of these compounds for cytochrome P-450 and estrogen synthase is indeed confirmed by the recent literature (Mayhoub et al., 2012).

The anatomical therapeutic chemical classification (ATC)[1] system is used to further check compounds agreements withing the clusters. Although most of the compounds in the data are not classified in ATC a number of the clusters contained compounds that have similar ATC classes indicating therapeutic homogeneity. For instance, cluster 2 contains 7 compounds the belong to the ATC class *S01A: Orthalmologicals Antiinfectives for Sensory Organs*, and cluster 11 contains at least 4 compounds which belong *L02B: Antineoplastic and Immunomodulating Agenets.*

[1] `http://www.whocc.no/atc_ddd_index/`

Table 6.1: Compounds in repurposing clusters with strong bioactivity for both, HIV and cancer related targets.

Clust	ChEMBL ID
C02	CHEMBL140484, CHEMBL335365, CHEMBL349008, CHEMBL335794, CHEMBL350563
C03	CHEMBL140827, CHEMBL337776, CHEMBL162880, CHEMBL1178786, CHEMBL2178569, CHEMBL115487, CHEMBL109051, CHEMBL323172, CHEMBL138285, CHEMBL139559, CHEMBL105821, CHEMBL432170, CHEMBL164484, CHEMBL138640, CHEMBL140418
C04	CHEMBL472832
C06	CHEMBL2326002
C07	CHEMBL2005886, CHEMBL1982465, CHEMBL2001485, CHEMBL223367, CHEMBL373598
C11	CHEMBL83, CHEMBL578028, CHEMBL575538
C15	CHEMBL475670
C16	CHEMBL2158601, CHEMBL2158599, CHEMBL1774154, CHEMBL481213, CHEMBL2158600, CHEMBL481422, CHEMBL1774162, CHEMBL1774161, CHEMBL1774157, CHEMBL1774031, CHEMBL1241426, CHEMBL1774158, CHEMBL1774163, CHEMBL2158053, CHEMBL1774160, CHEMBL1774164, CHEMBL1774156, CHEMBL456400, CHEMBL516172, CHEMBL2158602, CHEMBL2158606, CHEMBL459541
C17	CHEMBL1980297

Table 6.2: Compounds in repurposing clusters with strong bioactivity for HIV, but unknown bioactivity for cancer related targets. "Cancer targets" here lists proteins that are targets of some (but not all) compounds in a given cluster and are putatively related to cancer. **Bold** compounds have originally not been tested against the listed targets in our data, but are confirmed to have strong bioactivity according to independent data.

Clust	ChEMBL ID	Cancer targets
C02	CHEMBL2158790, CHEMBL2207654, CHEMBL2158784, CHEMBL2158783, CHEMBL2158791, CHEMBL2171047	CCR5, CCR1, EDNRA, CXCR3
C03	CHEMBL2207286, CHEMBL2207660, CHEMBL2158787, CHEMBL2207657, CHEMBL2207283, **CHEMBL2158785**, CHEMBL2207664, CHEMBL2158792, CHEMBL1171008	CCR5, CCR1
C13	CHEMBL1171594	HDAC6, CYP19A1, ACE, GHSR, CXCR3, UTS2R, GRPR, HDAC1, HDAC2, APP, KLK3, KCNH2, MEN1, MME, MTAP, P2RY1, HDAC7, PLAT, PPARA, PPARD, PPARG, AKR1B10, RARB, SLC16A1, SRD5A1, SSTR3, SSTR5, ST14, BRS3
C15	CHEMBL1230584	BIRC2, BIRC3, CHRNA7, ESR1, ESR2, AR, PPARA, PPARG

Table 6.3: Compounds in repurposing clusters with strong bioactivity for cancer, but unknown bioactivity for HIV related targets. "HIV targets" here lists proteins that are targets of some (but not all) compounds in a given cluster and are putatively related to HIV. **Bold** compounds have originally not been tested against the listed targets, but are confirmed to have strong bioactivity according to independent data.

Clust	ChEMBL ID	HIV targets
C02	CHEMBL1513, CHEMBL29346, CHEMBL282628, CHEMBL29972, CHEMBL281977, CHEMBL277447, CHEMBL440780, CHEMBL274489, CHEMBL112624, CHEMBL281659, CHEMBL29793, CHEMBL285832, CHEMBL27855, CHEMBL9194, CHEMBL128818, CHEMBL28863, CHEMBL2113316, CHEMBL282359, CHEMBL30405, CHEMBL281549, CHEMBL28963, CHEMBL8923, CHEMBL29223, CHEMBL303631, CHEMBL8978, CHEMBL109648, CHEMBL437472, CHEMBL431296, CHEMBL29464, CHEMBL8981, CHEMBL1204799, CHEMBL284656, CHEMBL282303, CHEMBL8823, CHEMBL29422, CHEMBL30092, CHEMBL30009, CHEMBL283610, CHEMBL302564, CHEMBL29775, CHEMBL48196, CHEMBL326059, CHEMBL282724, CHEMBL1921858	CCR3, CCR5, CCR2
C03	**CHEMBL232656**	CCR3, CCR5, CCR2
C05	**CHEMBL186101, CHEMBL1086088**	CCR3, BIRC3, CXCR1, CCR2
C06	**CHEMBL213207, CHEMBL1983315, CHEMBL242865, CHEMBL1086736**, CHEMBL2159206, **CHEMBL1977134, CHEMBL570366**	CCR3, CX3CR1, CCR2
C07	**CHEMBL1999153**, CHEMBL67237, CHEMBL158405, CHEMBL158939, CHEMBL367019, CHEMBL106666, CHEMBL289318, CHEMBL277695, CHEMBL705	CCR3, CX3CR1, BIRC3, HSP90AA1, CXCR1, ITK, MDM2, CCR2
C11	**CHEMBL99, CHEMBL96051**, CHEMBL55380, **CHEMBL483254, CHEMBL356066, CHEMBL1213490, CHEMBL430060, CHEMBL1089503**, CHEMBL208212, **CHEMBL1091204, CHEMBL1089630, CHEMBL1444, CHEMBL73279, CHEMBL9298, CHEMBL1957214, CHEMBL1399, CHEMBL73367**, CHEMBL219531, **CHEMBL74339**, CHEMBL198598, **CHEMBL306022, CHEMBL70959, CHEMBL1957217, CHEMBL101540, CHEMBL281433, CHEMBL98998, CHEMBL98537**	CYP2B6, MDM2, ABCB1

Continued on next page

Table 6.3 – *Continued from previous page*

Clust	ChEMBL ID	HIV targets
C13	**CHEMBL2153740**, CHEMBL381642, CHEMBL93124, CHEMBL92615, CHEMBL275311, CHEMBL205807, CHEMBL1080585	HCRTR2
C15	**CHEMBL2151438, CHEMBL2151570, CHEMBL2151569, CHEMBL2179874, CHEMBL2179875, CHEMBL2179877, CHEMBL597241, CHEMBL510275, CHEMBL195190, CHEMBL2179873, CHEMBL2179878, CHEMBL2177538, CHEMBL2177537, CHEMBL364069**, CHEMBL46937, CHEMBL12987, CHEMBL243571, CHEMBL243789, CHEMBL395291, CHEMBL244001, CHEMBL56198, CHEMBL58688, CHEMBL389907, CHEMBL380565, CHEMBL243791, CHEMBL56390, CHEMBL244207, CHEMBL1358, CHEMBL390448, CHEMBL2036560, CHEMBL128654, CHEMBL26865, CHEMBL197188, CHEMBL50241, CHEMBL467790, CHEMBL257379, CHEMBL224204, CHEMBL398226, CHEMBL390849, CHEMBL253535, CHEMBL402835, CHEMBL184133, CHEMBL107367	BIRC2, BIRC3, HSP90AA1
C18	**CHEMBL1994669, CHEMBL298445, CHEMBL187081**	ITK
C19	**CHEMBL1984548, CHEMBL1991734**	ITK
C20	**CHEMBL494089, CHEMBL359794, CHEMBL184510, CHEMBL1083151, CHEMBL1083152, CHEMBL241024, CHEMBL537964**	ITK
C21	CHEMBL292910, CHEMBL55401, CHEMBL2052008, CHEMBL99309, CHEMBL291026, CHEMBL418050, CHEMBL350849, CHEMBL270437, CHEMBL41152, CHEMBL710, CHEMBL435631, CHEMBL1237140	MDM2

6.3.7 BES Clustering Groups Structurally Diverse Compounds

Chemical structural properties of compounds falling into the same BES cluster are investigated. In particular, a Maximum Common Substructure (MCS) search for none-singleton clusters is conducted (Figure 6.6). Only few clusters (clusters 35, 37, 44, 47, 60 & 63) revealed larger MCS scaffolds. Since the relative size of the MCS is an evidence for the structurally relatedness of compounds in a cluster this indicates that structurally diverse compounds were grouped together.

In order to investigate this point further, additional consensus clustering based on MACCS fingerprints and the Tanimoto-Jaccard coefficient is conducted (see subsection 6.2.6). This is called chemical clustering in the following. 254 clusters were produced. The observed AUC was close to one, indicating a highly stable clustering result (Figure 6.4 B1 & B2). However, most clusters showed a much lower silhouette width than in the BES clustering. Also the average silhouette width was significantly lower (~ 0.19; Figure 6.4).

The relative overlap size between BES groups and chemical clusters based on the Tanimoto-Jaccard index is investigated. Only 27 chemical clusters overlapped with at least one BES cluster, the rest of the clusters did not overlap at all or showed a Tanimoto-Jaccard index value below 0.05. Only 3 BES clusters had an overlap of more than 25% (Figure 6.7). This demonstrates again that BES and chemical clustering capture different patterns. Chemical clustering focuses on structural similarity, whereas BES clustering focuses on similar biological effects. Compounds with similar biological effects are in turn not necessarily structurally similar.

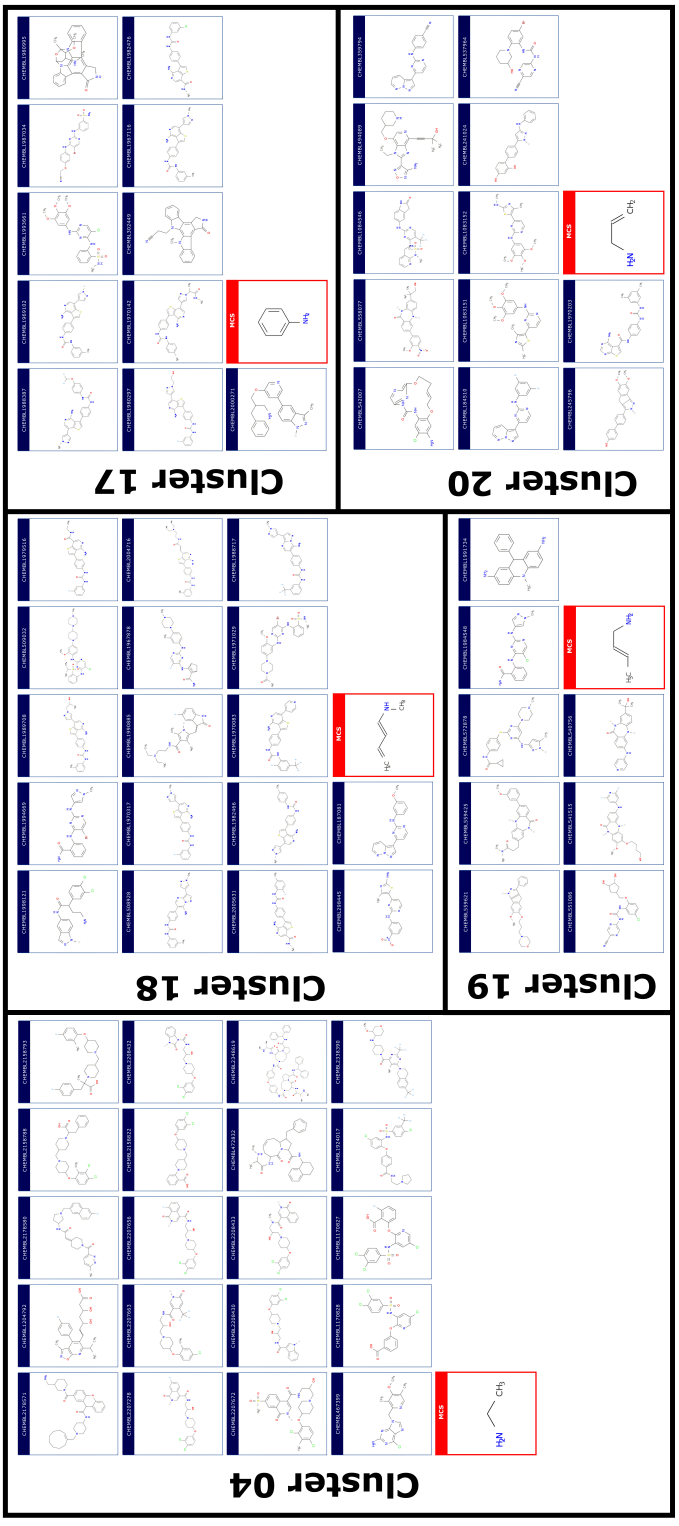

Figure 6.6: Maximum common substructure analysis. Five selected repurposing clusters: Each panel shows ligand structures in blue frames and the maximum common substructure (MCS) in a red frame.

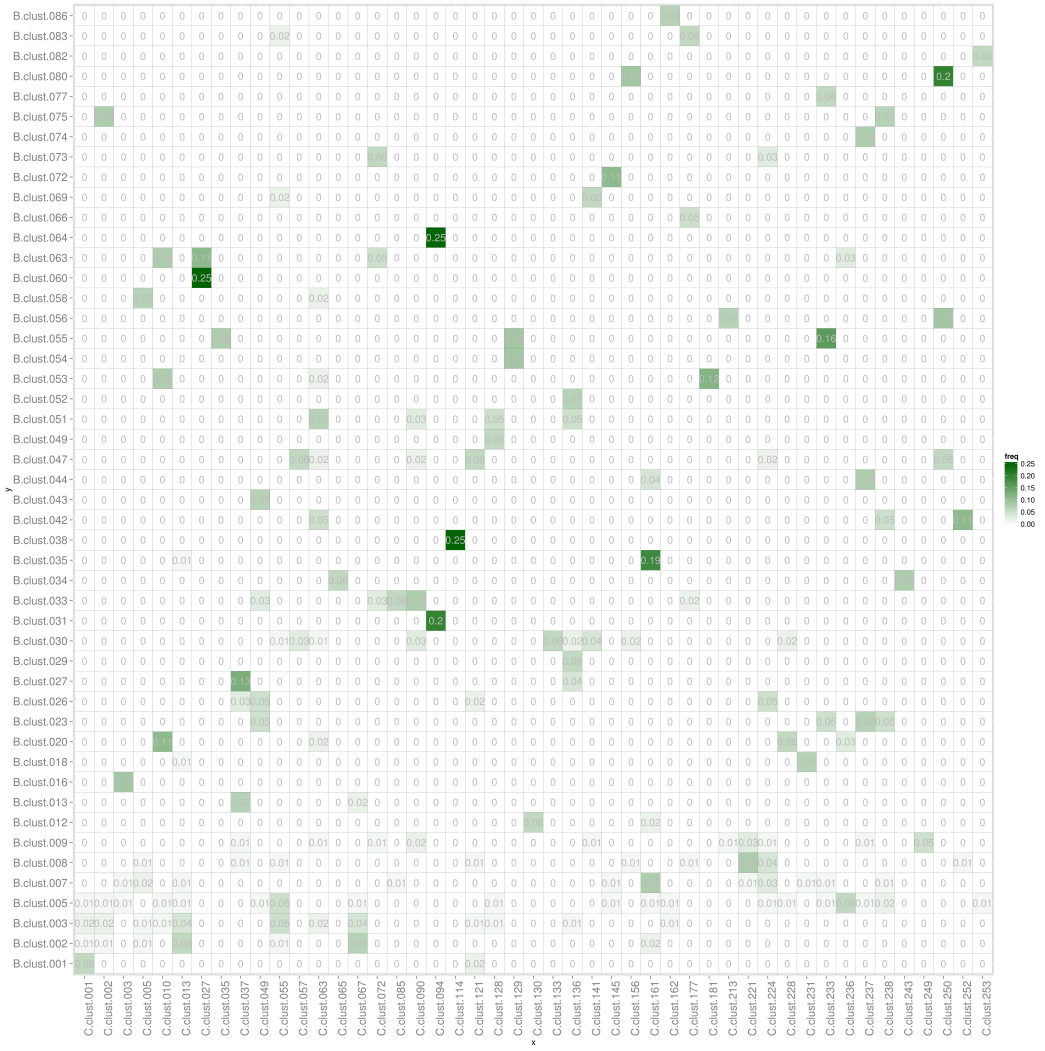

Figure 6.7: Comparison of BES based (Y-axis) and fingerprint (X-axis) based clustering. The color code indicates the fraction of the overlap between cluster pairs.

6.4 Conclusions

The BES is developed to quantify the similarity of biological effects induced by targeted perturbations with chemical compounds. In contrast to existing methods the BES directly integrates compound-target affinities with different features of target proteins. The BES measure is validated by comparing it against gene expression data, showing a highly significant correlation.

The BES measure is applied within a similarity based consensus clustering algorithm. This approach enabled us to identify well separated and statistically stable groups of compounds inducing similar biological effects. The identified clusters could not be found in a traditional ligand based clustering using MACCS fingerprints and the Tanimoto-Jaccard similarity. Our method specifically found compound clusters with targets associated to different diseases, hence indicating similar biological effects despite of different tested medical application areas. It is demonstrated that the cluster analysis may help to identify interesting novel compound-target combinations. Altogether, a two-fold potential of the method can be specified: First, BES based clustering could aid to explore existing compound libraries and identify promising compound-target combinations. Second, the BES may help to optimize compounds towards a desired effect profile.

The present analysis has been restricted to cancer and HIV targets here in order to demonstrate the principal usefulness of our proposed BES measure. Future work should extend the application to a broader dataset covering other disease indications. In that context it is worth mentioning that the investigated dataset with almost 4,700 compounds containing $\sim 73,000$ interactions with more than 200 target proteins is already quite large, specifically if re-sampling based methods like consensus clustering are applied in order to detect statistically stable groups.

A limitation of the present work is that the unavoidable uncertainty about the disease association of a particular protein is currently neglected. Future work should thus extend the present approach in this direction. Despite of this limitation the results demonstrate that useful predictions for compounds can already be retrieved by comparison of biological effects similarities.

CHAPTER 7

CONCLUSIONS

Perturbation experiments and high-throughput technology hold the promise to further advance the research in biology and medicine. Targeted perturbation experiments allow inducing controlled alternations in biological systems, and high-throughput techniques enable measuring cellular and biochemical activity of large numbers biomolecular entities under different perturbational variations. Interventional data gained from multiple such experiments can guide the identification of the functional consequences of induced or natural variations. Statistical methods that exploit variations in replicated measurements are used for inference and prediction of perturbation responses by detecting relevant patterns in the data.

However, although advances in biotechnology permit sophisticated perturbation experiments, particular characteristics of the experimental techniques (such as mircoarray, HT-qPCR and NGS) involved in measuring variety of molecules lead to generating diverse and complicated data. The sheer volume and high dimensionality of high-throughput data pose major challenges for its analysis and computational integration. Further experimental implications add to this severity. Missing or censored observations and different sources of variations in the data, for instance, can impair the performance of generic analysis methods, and very diverse types of data can make the computational integration very difficult. Therefore, understanding the characteristics of experimental methods behind the acquired data is very important in order to be able to the use appropriate analytical methods.

The main objectives of this thesis were to explore, devise and develop proper methodologies for efficient extraction of biologically relevant patterns from diverse perturbation data. Advanced normalization techniques and statistical analysis methods for

gene expression data have been discussed and illustrated on perturbation data from extensive experimental studies. The large-scale characterization of biological systems is aided by computational integration approaches for information from literature and via text mining.

This work is composed of three parts. Each of these is a different perturbation experiment setting with its own distinguishable experimental implications and data analysis challenges.

The first part investigated the dynamical transcriptional responses to TGF-β stimulation in different human and mouse cell types base on time-series microarray data from extensive experiments. A panel of statistical and bioinformatics methods are used to gain biologically relevant insights which could help understand the complex mechanisms by which TGF-β yields phenotypic effects.

A Bayesian approach to determine differentially expressed time-course is used. The method detected transcript response profiles between perturbed and naive conditions that could not be identified by classical time-point specific differential expression analysis. Most of these transcripts, in particular genes that are known to belong to the TGF-β pathway, reacted time-dependent. Although the identified transcriptional responses were highly tissue specific, several commonly affected processes and signaling pathways across cell types and species are discovered. This is done using a logistic regression-based method for gene set association analysis. This method provides comparable results between the different cell types because it does not depend on which or how many genes have been measured in the chips or ranked in the differential expression analysis. Further functional analyses suggested an important role of few transcription factors, which appear to have a conserved influence across cell types and species. A devised visualization tool manifested effective integration of comprehensive results of various functional analyses and helped assessing proximity between tissues and organisms.

A clustering method that considers the dynamics of expression changes successfully grouped similar expression profiles in the different cell types. Transcription factor enrichment analyses within clusters allowed for partial reconstruction of gene regulatory modules. Validation via an independent dataset confirms the findings and network analyses suggest explanations, how TGF-β perturbation could lead to the observed effects.

The analysis of such complex experimental data could be improved by integrating modeling approaches, e.g. hypotheses on the dependencies between gene regulatory modules may be derived via advanced predictive methods such as dynamic Bayesian networks.

The second part investigated miRNA and relevant target genes as potential biomarkers in epilepsy. Longitudinal microarray and high-throughput qPCR data are generated from blood and hippocapus samples of rat and mouse perturbation models after initiating generalized epileptic seizure in them.

I suggested a specific work-flow (the double-detection method) for differential expression analysis of HTqPCR data where part of the data is right-censored for observations only known to be above the detection limit. The method involves modeling the censored data based upon maximum likelihood. The devised procedure showed higher detection power based on simulated data in comparison to methods that exclude the censored observations from the analysis or replace them by a fixed value.

The double detection method enabled the identification of novel pathogenic relevant miRNAs in hippocampal tissues and blood samples. The identified biomarkers are successfully validated via independent RT-PCR experiments. Indirect characterization of epilepsy types is achieved through comparisons of miRNA activities in the animal models at different time points and relevant mRNAs could be identified through annotation of sequence predicted putative targets. Using functional analysis methods a number of disease relevant biological processes and pathways were significantly associated to the identified gene target sets of deregulated miRNAs.

The third part of this work studied drug induced downstream effects on system level based on compound perturbation data. A new method for compound similarity assessment based on drug pharmacological effects and drug-target interaction network is proposed. The *biological effects similarity (BES)* measure of chemical compounds integrates compound-targets affinities and captures compound perturbation induced effects by assessing target-target proximity in a probabilistic manner using target information from multiple sources. Unlike existing similarity measures the BES requires neither chemical structure information nor treatment response data.

The BES showed high correlation to gene expression profile similarities computed from treatment response data. This indicates its validity. The consensus clustering utilizing BES-based distance measure is used for exploratory analysis of large dataset of chemical compounds related to HIV and cancer. This procedure produced separable and statistically stable 65 clusters, of which targets could be related to specific pathways, biological processes, protein domains and sequence motifs. The identified clusters could not be found in a traditional chemical structure based clustering using fingerprints and the Tanimoto-Jaccard similarity. The clustering did not separate the two disease, so that numerous clusters contained compounds of both diseases, hence indicating biologically close effects despite of a very different medical application area.

Targets of compounds falling into one cluster suggested several new compound-target combinations, which could in several cases be confirmed by independent data.

Taken together this work suggests the usefulness of a joint view on the compound-target space and the proposed BES measure in particular. Our results on cancer and HIV demonstrate that the latter may help to uncover drugs with biologically similar effects and in consequence also repurpose existing drugs for novel application areas. However, BES measure need to be scrutinized; whether it puts much weight on the induced effects of targets. As biological systems are robust, drug perturbation effects are restored rather passed to nearby proteins. Investigating the method in datasets of multiple diseases is also recommended.

In conclusion, although a number of approaches and statistical methods have been proposed for pattern detection from high-throughput perturbation data, due to the complexity of such data, choosing the appropriate method and correctly implementing it is still a challenge. One reason is that choosing an appropriate method typically requires sufficient understanding of the data generation process.

In this work I have explored, devised and developed methodologies for the efficient extraction of biologically relevant patterns from diverse perturbation data. These methods which incorporated several analysis stages have been illustrated in a number of compelling studies. Further modeling techniques, including network reverse engineering and mathematical modeling of biological systems, could be used as a follow-up step to the approaches presented here.

Appendices

APPENDIX A

CHAPTER 4 APPENDICES

A.1 Cell and Tissue Types

A.1.1 Mouse Hematopoietic Progenitor Cells (MPP & CDP)

Multipotent Progenitors (MPP) and Common Dendritic Progenitors (CDP) are Hematopoietic stem cells, i.e. they have a property to develop into many types of cells. The latter are derived from the former types. In the process of specialization the cells undergo several intermediary states and differentiate into multiple of cell types. Instead of all getting specialized, part of the MPP proliferated population remains as before. This is indicated by the loop in Figure A.1. CDP cells represents such an intermediary state and is the one that gives rise to Conventional Dendritic Cells (cDC) and Plasmacytoid Dendritic Cells (pDC) as seen in Figure A.2 (Zenke and Hieronymus, 2006). These cells and their specialized different types partially form the immune system. Further details can be found in (Felker et al., 2010) and Abnaof et al. (2014).

Cell culture MPP and CDP were obtained from mouse bone marrow, using *in-vitro* culture with a specific cytokine cocktail and FACS sorting (Felker et al., 2010; Seré et al., 2012).

TGF-β1 stimulation After sorting, MPP and CDP were treated with 10 ng/mL recombinant human TGF-β1 (R & D Systems, Minneapolis, USA) for 2, 4, 8, 12 and

24 hours as described in (Felker et al., 2010) or left untreated. Cells were lysed in 350 μl TRI-Reagent and stored at -80 °C.

RNA Isolation RNA was isolated using the MagMAX-96 for Microarrays kit (Life Technologies, Darmstadt, Germany) according to manufacturer's protocol.

GeneChip Hybridization Replicated time-course microarrays are produced by creating three technical replicates for each one of the cell types at 6 successive time points; 0, 2, 4, 8, 12, 24 hours (detailes in Table A.1). Assays are produced using Affymetrix® GeneChip type "Mouse Gene 1.0 ST Array" with 32,321 probe-sets. Hybridization, wash and staining were done according to manufacturer's recommended standard techniques.

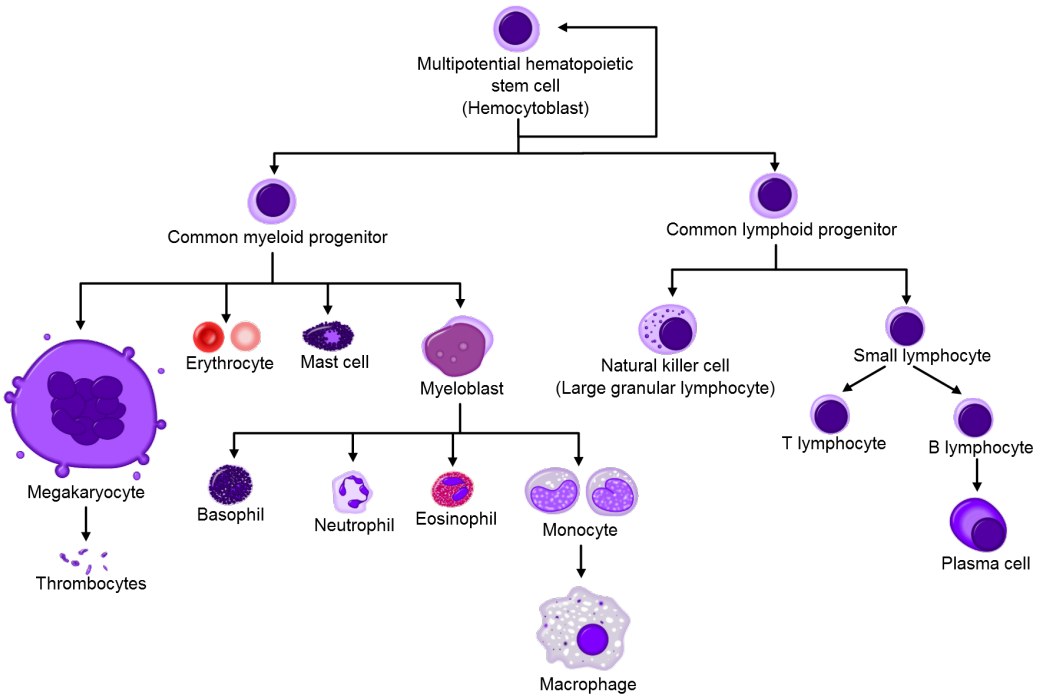

Figure A.1: Development of different blood cells from Haematopoietic stem cells to mature cells[1].

[1] used with permission, see original figure in source (accessed: 24th July 2015): `https://en.wikipedia.org/wiki/Haematopoiesis`

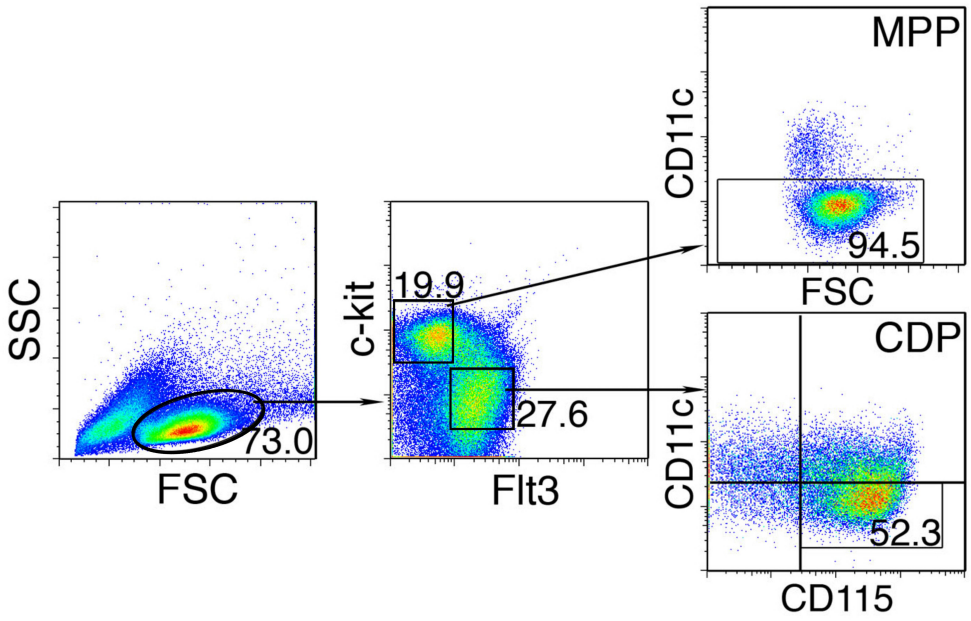

Figure A.2: Multiparameter (10-colors) analysis of murine Dendritic cells using focusing cytometer. Lymphocytes are gated via FSC/SSC parameters (left). CD4+ T cells are separated into two populations of MPP cells & CDP cells based on the expression of co-regulators c-kit and Flc3. MPP cells can be precised via FSC and the expression of CD11c. CDP cells can further be divided into CD11c and CD115 dendritic cell subtypes.

A.1.2 Primary Mouse Hepatocytes and Human HepG2 Cells (HPC)

Hepatocytes (HPC) represent the most prominent cell population in the liver. Primary HPC are sensitive to TGF-β1, and express the corresponding type I (ALK5), type II (TβRII), and type III (betaglycan) receptors. TGF-β1 promotes cell cycle arrest and apoptosis of primary HPC. In addition, *in-vitro* TGF-β1 provokes epithelial-to-mesenchymal transition (EMT)-like processes in this hepatic cell subpopulation, which most likely do not occur *in-vivo* (Chu et al., 2011). HepG2 cells originate from a 15 year old child with primary hepatoblastoma (Aden et al., 1979). They do secrete the major plasma proteins but do not express the hepatitis B virus surface antigen (HBsAg) (Knowles et al., 1980). Figure A.3 shows the application process for HPC cells from mouse (primary) and human cell-lines. Abnaof et al. (2014) gives further details.

Cell culture Primary murine HPC were isolated from male C57BL/6 mice according to the collagenase method of Seglen (Seglen, 1976). Cells were plated in collagen coated 6-well dishes at a density of 1.2 x 106 cells using HepatoZYME-SFM (Gibco, Life Technology, Darmstadt, Germany). Four hours after seeding the medium was renewed and cells were grown for a further 24 hrs culture period. HepG2 (DSMZ: DSM ACC180) were cultured in RPMI (PAA, Pasching, Austria) containing 10% fetal calf serum (PAA), 1 x Penicillin/Streptomycin (Lonza, Cologne, Germany). Medium was renewed every second day. For the experiment, cells were passaged and plated in 6-well dishes using accutase (PAA) at a density of 4 x 105 cells. One day before the experiment, cells were washed with PBS (1x), medium changed to HepatoZYME-SFM (Gibco) and cultured for further 24 hrs.

TGF-β1 stimulation One hour before the experiment, the medium was exchanged and cells stimulated with 1 ng/mL recombinant human TGF-β1 (R&D Systems, Minneapolis, USA) for indicated time intervals; HepG2: 0, 20 minutes, 1, 2, 4, 24 hours; primary murine HPC: 0 min, 1, 2, 4 hours. The cells were harvested using Qiazol for cell lysis (Qiagen, Hilden, Germany), directly frozen and stored at -80 °C.

RNA Isolation RNA was isolated using the RNeasy Kit system (Qiagen), performing a DNAse digestion according to the manufacturer's protocol.

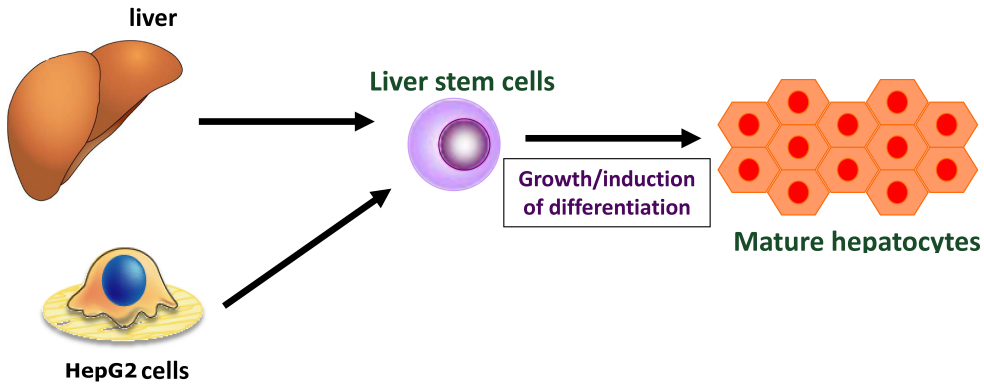

Figure A.3: Primary Mouse Hepatocytes and Human HepG2 Cells (HPC). HPC are cultivated from mouse liver and from human HepG2 cell-line.

GeneChip Hybridization Replicated time-course microarrays are produced by creating three technical replicates for human HPC cell types at 6 successive time points;

0, 20 minutes, 1, 2, 4, 24 hours and in 4 time points; 0 min, 1, 2, 4 hours (detailes in Table A.1). Human samples were assayed using Affymetrix® GeneChip type "Human Gene 1.0 ST Array" with 34,760 probe-sets and mouse samples were assayed in Affymetrix® GeneChip type "Mouse Gene 1.0 ST Array" with 32,321 probe-sets. Hybridization, wash and staining were done according to manufacturer's recommended standard techniques.

A.1.3 Human Mesenchymal Stromal Cells (MSC)

Mesenchymal stromal cells (MSC) are multipotent cells which are characterized by their plastic adherence. They have the potential to renew itself and to differentiate *i*n-vitro and *i*n-vivo into diverse vital stem cell types (Figure A.4). The differentiation potential of MSC include Adipocytes, Osteoblasts, Chondrocytes, Myocytes, Pancreatic islet cells, brain cells and neurons, Blood Cells and Hepatocytes (Sharma et al., 2012). MSC are found in all supportive tissue as in fat tissue, bone marrow and cord. All MSC express the surface markers CD29, CD73, CD90 and CD105 and they lack the expression of CD14, CD31, CD34 and CD45 (Dominici et al., 2006; Horwitz et al., 2005)).

The cells were isolated, incubated and sorted into younger (early passages) and older (late passages) cells. Further details can be found in Abnaof et al. (2014) and in Walenda & Abnaof et al. Gudrun et al. (2013).

Isolation and Expansion MSCs were isolated from mononuclear cells (MNCs) by plastic adherence as described before (Koch et al., 2012; Lohmann et al., 2012; Walenda et al., 2012). In brief, bone fragments from the caput femoris of patients undergoing femoral head prosthesis were flushed with phosphate- buffered saline (PBS) and washed twice with PBS. MNC were then resuspended in culture medium and seeded into tissue culture flasks.

The cells were cultured at 37 °C in a humidified atmosphere with 5% CO2. The first medium exchange was performed after 48 h to remove nonadherent cells. Thereafter, media changes were performed twice per week and MSCs were passaged when reaching 80-90% of confluence.

TGF-β1 stimulation MSC from three different donors were used in an early passage (p3-5) for stimulation with TGF-b1. 1x106 MSC were seeded into 6-well culture plates. When the cells were attached after 24 h 1ng/mL recombinant TGF-β1 (R&D Systems) was added to the culture media at different time points. The cells were

harvested at the same time point with Qiazol (Qiagen) and directly frozen and stored at -80 °C.

RNA Isolation RNA was isolated via phenol/chloroform extraction using the miRNeasy Kit (Qiagen), performing a DNAse digestion.

GeneChip Hybridization Replicated time-course microarrays are produced by creating three technical replicates for each one of the cell types at 4 successive time points; 0, 1, 4, 12 hours (detailes in Table A.1). Experiments are carried under equal conditions. Assays are produced using Affymetrix® GeneChip type "Human Gene 1.0 ST Array" with 34,760 probe-sets. Hybridization, wash and staining were done according to manufacturer's recommended standard techniques.

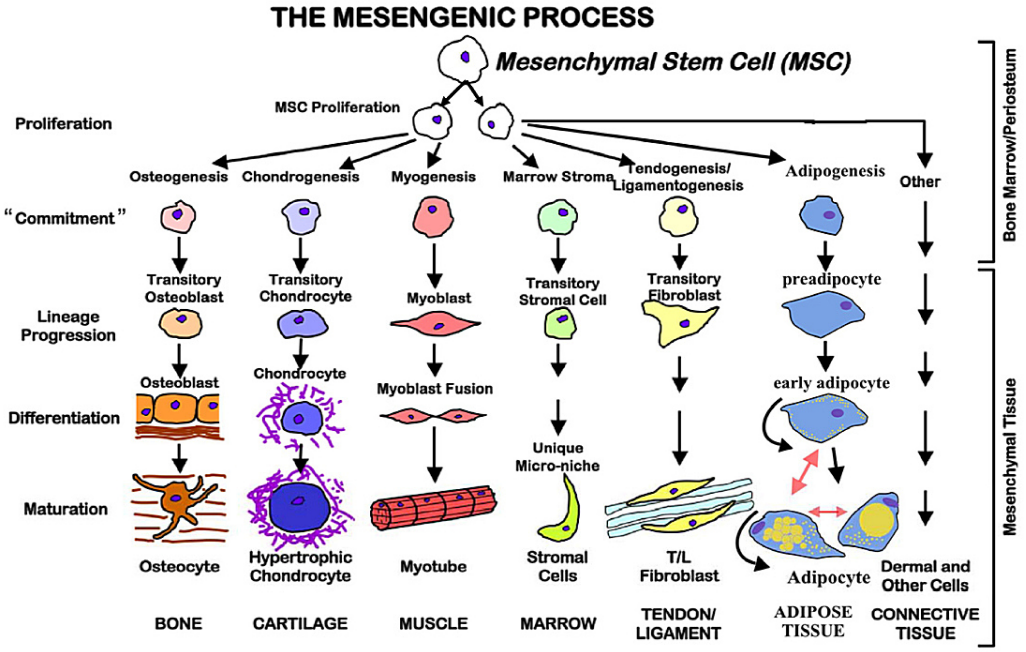

Figure A.4: Mesenchymal Stromal Cells (MSC) Differentiation. MSC are proved to be cultured *in-vitro* and differentiate into variety of cell types. The differentiation and the resulting cell types depends on the culture conditions[2].

[2] used with permission, original figure in source (accessed: 24th July 2015): http://www.discoverymedicine.com/Tracey-L-Bonfield/2010/04/15/adult-mesenchymal-stem-cells-an-innovative-therapeutic-for-lung-diseases

A.1.4 Data Availability

The microarray datasets for the different cell types are deposited in the Gene Expression Omnibus (GeneExpression, 2015) under the corresponding accession numbers (Abnaof et al., 2014):

- Murine Primary Hepatocytes (HPC); GSE45942[3] (Meurer SK, Walenda G, Abnaof K, Joussen S, Lin Q, Zenke M, Hoffmann K, Wagner W, Fröhlich H, 2013b).
- Human HepG2 Cell-line (HPC); GSE45945[4] (Meurer SK, Walenda G, Abnaof K, Joussen S, Lin Q, Zenke M, Hoffmann K, Wagner W, Fröhlich H, 2013a).
- Human Mesenchymal Stromal Cells MSC; GSE46019[5] (Meurer SK, Walenda G, Abnaof K, Joussen S, Lin Q, Zenke M, Hoffmann K, Wagner W, Fröhlich H, 2013c).
- Murine Hematopoietic Progenitor Cell; Multipotent Progenitors & Common Dendritic Progenitors (MPP & CDP); GSE46109[6] (Meurer SK, Walenda G, Abnaof K, Joussen S, Lin Q, Zenke M, Hoffmann K, Wagner W, Fröhlich H, 2013d).

[3] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45942
[4] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45945
[5] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46019
[6] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46109

A.2 Supplementary Tables

Table A.1: Overview of the different experiments. Chips for mouse hepatocytes at time point 1h were removed due to quality issues.

Organism	Mouse			Human	
Cell Type	Multipotent progenitor (MPP)	Comm-Dendritic Progenitor (CDP)	Hepatocytes (HPC)	Hepatocytes (HPC)	Mesenchymal (MSC)
Time Points	6	6	5	6	4
Replicates	3	3	3	3	3
00 Hours	✓	✓	✓	✓	✓
20 Minutes	✗	✗	✓	✓	✗
01 Hours	✗	✗	✓	✓	✓
02 Hours	✓	✓	✓	✓	✗
04 Hours	✓	✓	✓	✓	✓
08 Hours	✓	✓	✗	✗	✗
12 Hours	✓	✓	✗	✗	✓
24 Hours	✓	✓	✗	✓	✗

Table A.2: TFBS analyses, time-course. Significantly predicted transcription factor's binding-sites (TFBS) in each cell type and condition (best match according to STAMP and $E - value \leq 1e - 3$) according to time-course differential expression analysis, (details in Supplementary Excel Files Excel file 17).

Mouse						Human			
MPP		CDP		HPC		HPC		MSC	
TFBS	E-value	TFBS	E-value	TFBS	E-value	TFBS	E-value	TFBS	E-value
SP1SP3_Q4	1.11E-16	KROX_Q6	2.32E-13			KROX_Q6	3.66E-15	KROX_Q6	0.00E-00
FOXP1_01	8.93E-14	PITX2_Q2	1.28E-12			PITX2_Q2	2.61E-12	SF1_Q6	3.24E-11
SP1SP3_Q4	3.19E-13	KROX_Q6	3.73E-12			TBX5_Q5	1.90E-11	KROX_Q6	3.92E-11
KROX_Q6	4.48E-11	FOXP1_01	4.61E-12			POU1F1_Q6	4.56E-11	PU1_Q6	5.49E-09
FOXP1_01	2.61E-09	FOXP1_01	8.81E-12			HFH4_01	3.49E-09	PIT1_Q6	1.75E-08
TEF_Q6	1.81E-08	E2A_Q2	2.36E-08			SP1SP3_Q4	1.74E-08	POU3F2_01	3.90E-08
POU6F1_01	1.54E-07	NFE2_01	1.91E-07			MAZ_Q6	9.64E-08	PAX4_04	1.63E-07
FOX_Q2	4.60E-07	POU3F2_01	1.97E-07			ZNF219_01	1.21E-07	MYC_Q2	6.71E-07
MAZ_Q6	1.89E-06	FOX_Q2	2.07E-07			RBPJK_Q4	1.21E-07	IRF_Q6	1.16E-06
HFH4_01	1.06E-05	POU6F1_01	6.61E-07			HNF1_Q6_01	1.39E-07	PAX4_04	1.89E-05
PITX2_Q2	3.10E-05	TEF_Q6	1.19E-06			AP2_Q6_01	1.76E-07	HFH4_01	3.72E-05
		FOXP1_01	2.72E-06			TFIII_Q6	8.09E-07	SF1_Q6	4.08E-05
		VDR_Q3	7.85E-06			LFA1_Q6	5.72E-06	FOXM1_01	9.98E-04
		YY1_Q6	1.01E-05			GFI1B_01	1.29E-05		
		CACBINDPROTEIN_Q6	1.66E-05			IRF1_01	1.33E-05		
		E2A_Q6	6.02E-05			PAX4_04	2.32E-05		
		E2A_Q2	1.25E-04			CEBPGAMMA_Q6	2.41E-05		
		E2A_Q2	1.25E-04			RREB1_01	5.67E-05		
		CACBINDPROTEIN_Q6	3.39E-04			AHR_01	6.62E-05		
						E2A_Q2	1.69E-04		
						LFA1_Q6	1.81E-04		

Table A.3: TFBS analyses at 4-hours. Significant TFBS in each cell type (best match according to STAMP and $E - Value \leq 1e - 03$) for the significant genes at time point 4 hours (Supplementary Excel Files Excel file 18)

Mouse						Human			
MPP		CDP		HPC		HPC		MSC	
TFBS	E-value	TFBS	E-value	TFBS	E-value	TFBS	E-value	TFBS	E-value
FOXP1_01	1.11E-16	FOXP1_01	8.91E-10	ZF5_01	5.91E-08	KROX_Q6	0.00E+00	CKROX_Q2	1.23E-10
SP1SP3_Q4	3.00E-15	PIT1_Q6	4.08E-08	HFH4_01	1.00E-05	SP1SP3_Q4	1.60E-10	PIT1_Q6	2.94E-10
ZF5_01	8.69E-11	HEN1_01	1.34E-07			AP2_Q6	2.26E-07	HNF1_Q6	6.46E-10
SP1_Q6_01	6.17E-10	DMRT5_01	1.68E-07			POU1F1_Q6	8.11E-06	KROX_Q6	4.57E-09
TEF_Q6	9.35E-07	TEF_Q6	1.73E-06			HFH4_01	2.36E-05	KROX_Q6	1.18E-07
AP2_Q6	3.25E-06	AP2_Q6_01	3.07E-06			SF1_Q6	1.03E-04	PAX4_04	1.79E-07
PPARG_02	2.50E-04	E2F_Q2	7.81E-05			CDC5_01	2.00E-04	ZNF219_01	7.99E-07
								TFE_Q6	2.09E-05

A.2.1 Supplementary Excel Files

Additional tables in excel files containing multiple sheets. The excel file can be found in the attached compact disc.

Excel file 8 Time-course differential expression analyses results of all cell types.

Excel file 9 Time-point differential expression analyses at 04 hours results of all cell types.

Excel file 10 Time-course KEGG pathways analyses results for all cell types.

Excel file 11 Time-course Gene ontology analyses results for all cell types.

Excel file 12 KEGG pathways analyses at 4 hours for all cell types.

Excel file 13 Gene ontology analyses at 4 hours for all cell types.

Excel file 14 Cluster Analyses gene assignment lists.

Excel file 15 Gene-set enrichment analyses of KEGG pathways in cluster groups.

Excel file 16 Gene-set enrichment analyses of gene ontology in cluster groups.

Excel file 17 Transcription factors binding sites (TFBS) analyses of differential time-course genes.

Excel file 18 Transcription factors binding site (TFBS) analyses of differential genes at 4 hours.

APPENDIX B

CHAPTER 5 APPENDICES

B.1 Cell and Tissue Types

B.1.1 Mouse Types

All procedures were carried out according to the Helsinki declaration and conducted according to the guidelines of the European Community Council directive 86/609/EEC. A local Ethics Committee approved all performed experiments.

5-6 weeks old male mice, purchased from Charles River, were subjected to two different models of chronic epilepsy causing spontaneous recurrent seizures (SRSs) (pilocarpine and SSSE models). Single injection of 300 mg/kg pilocarpine (muscarinic agonist), was used to trigger generalized spontaneous seizures in NMRI mice, weighing 28-32 g. The animals were intraperitoneally injected with 1 mg/kg of N-Methylscopolamine bromide 30 minutes prior to pilocarpine treatment. Within 10 to 45 minutes after treatment animals displayed generalized clonic-tonic seizures that progressed to continuous convulsive activity, i.e. status epilepticus (SE). To limit extensive brain damage, SE was interrupted 1 to 2 h after induction by intraperitoneal (i.p.) injection of Diazepam (10 mg/kg). The mice surviving SE typically show SRSs within the few days and continue to display them for several weeks. As controls, age–matched, completely naive mice were used.

Self-Sustained Status Epilepticus (SSSE) is a chronic epileptic model induced by electrical stimulation. As previously described (Niespodziany et al, 2010) C57Bl/6J male mice were surgically implanted with EEG electrodes: depth electrode (bipolar): AP= -1.40 mm; L = -2.65 mm; D = -5.00 mm, cortical electrode (monopolar): AP

= - 4.00 mm, L = + 3 mm and reference electrode in the prefrontal bone. After recovery from surgery, mice underwent electrical stimulation through the amygdala-implanted electrode (90 minutes duration, 100 ms trains of 1 ms alternating current pulses (50 Hz), 2 trains per 1 s, 250 μA peak current intensity). Upon cessation of electrical stimulation, the animals developed SSSE represented as continuous convulsive activity that was stopped after 150 minutes by i.p. injection of Diazepam (10 mg/kg). As controls age-matched, non-stimulated animals were used. For both, the pilocarpine and the SSSE model, 8 animals were used per experimental group.

The third group of mice underwent electrical induction of an acute seizure (6-Hertz model) that caused one psychomotor seizure as a model of partial epilepsy. In the 6-Hertz model, mice were electrically stimulated to evolve acute, focal seizures. As previously described (Kaminski et al, 2004) the induction was triggered by a stimulator (ECT Unit 57800, Ugo Basile, Comerio, Italy) using a current intensity of 44 mA, 0.2 ms monopolar pulses at 6-Hertz frequency for a duration of 3 s through corneal electrodes. Prior to stimulation, a drop of saline with 0.1% Unicaïne was placed on the eyes to ensure good conductivity and mild anesthesia. After stimulation, each mouse was observed for convulsive behaviour (i.e. stereotypy, immobility, and mild myoclonus), typically lasting more than 7 seconds. As controls age-matched non-stimulated animals were used. In this model 7 animals per experimental group were used.

B.1.2 Tissue collection and RNA isolation

Mice were sacrificed 24 h and 28 days (28 d) after the induction of SE in both the pilocarpine and the SSSE model. As control groups, 24 h and 28 d time points were used for the pilocarpine model and a 28d time point was used for the SSSE model for technical reasons. Mice were sacrificed at 3 h, 6 h, 24 h and 72 h following acute seizure in the 6-Hertz model. Mouse hippocampi were extracted from fully anesthetized animals (Nembutal, Ceva Santé Animale, France) and rapidly frozen. Total RNA was isolated from the sonicated tissue using miRVANA miRNA isolation kit from Ambion, Texas, USA. (Ambion Inc., Austin, Texas, USA) and RNA quality was checked on the Agilent Bioanalyzer Lab-on-a-Chip System observing intact 5S, 5.8S and 18S ribosomal RNA. All samples analyzed had a RIN number >7.

B.1.3 MicroRNA Microarray Profiling

Profiling was performed by Exiqon A/S using the miRCURYTM LNA Array microRNA Profiling Service on dual-channel 5th generation miRNA arrays with complete coverage of miRbase v14 according to Exiqon standard operating procedures. For

the pilocarpine and SSSE models miRNA microarray profiling was performed at 24 h and 28 d following SE. For the 6-Hertz model miRNA microarray profiling was done at 3 h, 6 h, 24 h, and 72 h after acute seizure. Based on comparison with the current miRBase® (Griffiths-Jones, 2004; Kozomara & Griffiths-Jones, 2011) release 19 all sequences that are not representing miRNAs anymore were deleted from the analysis.

B.1.4 Real-time RT-PCR

Universal cDNA synthesis kits (No. 203300), SYBR Green master mix (No. 203450) and mercury LNA Universal RT microRNA PCR primer sets purchased from Exiqon A/S were used for quantification of miRNAs (mmu-miR-124*, mmu-miR-132, mmu-miR-142-3p, mmu-miR-142-5p, mmu-miR-21, mmu-miR-212, mmu-miR-221, mmu-miR-222, mmu-miR-298, mmu-miR-882, mmu-miR-2137) Synthesis of cDNA was done according to Exiqon standard protocols using 20 ng of total RNA. Concentration of RNA and cDNA was measured by Nanodrop 2000c Spectrophotometer (Peqlab Biotechnology). The real-time PCR reactions were carried out in 384 well plates with 5 μl SYBR Green master mix, 1 μl of primer mix for each miRNA and 4 μl of 1:80 diluted cDNA per well. Each sample was run in triplicates for the miRNA of interest as well as for endogenous controls (SNORD68 or RNU6). The real-time PCR assays were performed on a 7900HT system (Life Technologies). The real-time PCR settings were 50°C for 2 minutes, 95°C for 10 minutes, 40 cycles of 95°C for 10 seconds followed by 60°C for 1 minute, and 25°C for 1 minute. Data were calculated using the $\triangle\triangle$Ct method of Schmittgen and Livak.

B.2 Data Availability

The microarray datasets for the different mouse models and time-points are deposited in the Gene Expression Omnibus (GeneExpression, 2015) under the corresponding accession numbers (Kretschmann et al., 2015a):

- 6-Hertz Mouse Model Data Set; GSE51840[1] (Kretschmann et al., 2015b).
- Pilocarpine Mouse Model Data Set; GSE51841[2] (Kretschmann et al., 2015c).
- SSSE Mouse Model Data Set; GSE51842[3] (Kretschmann et al., 2015d).

[1] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51840
[2] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51841
[3] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51842

B.3 Supplementary Figures

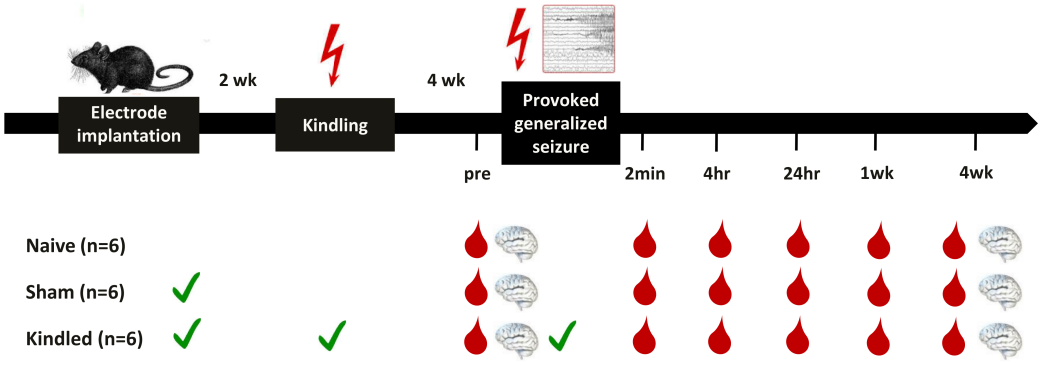

Figure B.1: Experimental design in rats. Three rat groups are considered; naive, sham and kindling groups. The sham and kindling groups are subjected to surgeries and electrodes are implanted in the amygdala of each animal in these two groups. Only the kindling animals undergone kindling. Hippocampal tissues and blood serum samples at taken form all the groups at 4 weeks pre-seizure and at 2 minutes, 4 hours, 1 day, 1 week and 4 weeks after generalized seizure. High-throughput qPCR (CYBER Green® and ABI-7900) Exiqon 2-panels chips are used for profiling miRNAs from blood and hippocampus samples.

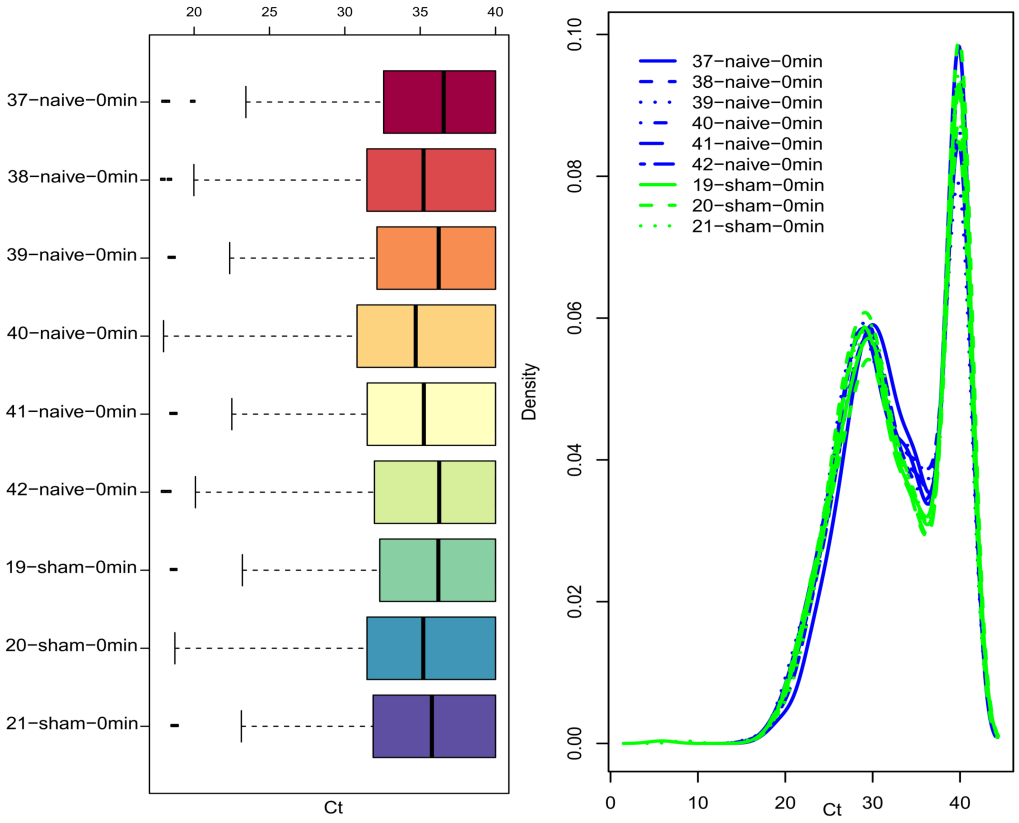

Figure B.2: Density and box plots of high-throughput qPCR miRNAs profiling for the first 9 chips. Imposing Ct value 40 to undetermined features renders strongly negative-skewed and often bi-modal density distributions of the data

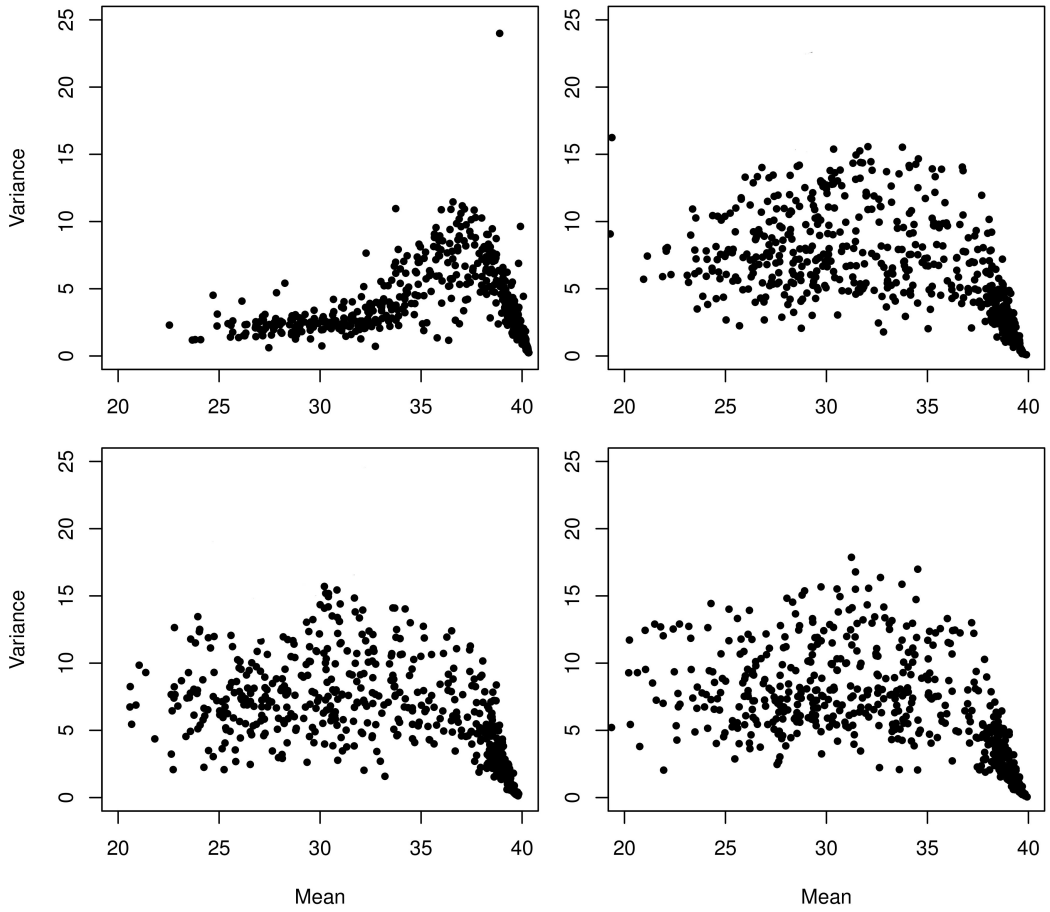

Figure B.3: Mean-variance plots of real and simulated high-throughput qPCR. The top left plot is for the real qPCR data of the kindled animals at 4 hours, the other three plots are 3 exemplary simulated qPCR data sets (out of 100,000 data sets)

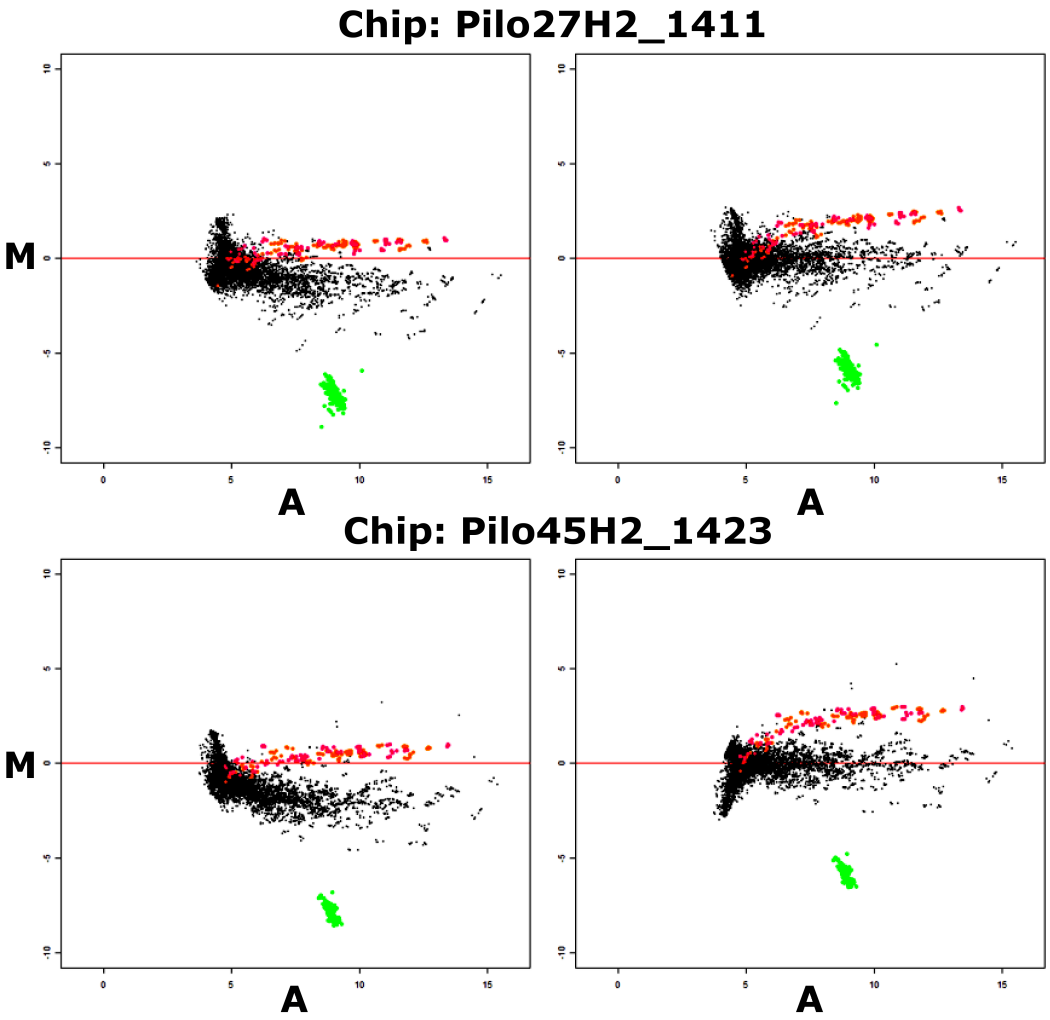

Figure B.4: MA plots for 2 exemplary chips from the mouse two-channels array data. The normalization is performed to minimize differences between the colors in an intensity-dependent manner. The plots show M versus A plots 'MA plot'; $M = \log_2(\frac{Hy5}{Hy3})$ (log-ratio; difference between the chennels) and $A = \log_2(\frac{Hy5 \times Hy3}{2})$ (combined signals from both channels). The green spots are Hy3 controls spotted directly on the array surface the orange spots are the 52 different spike-in controls. The plots show 2 exemplary chips from Pilocarpine animals before (left) and after normalization (right).

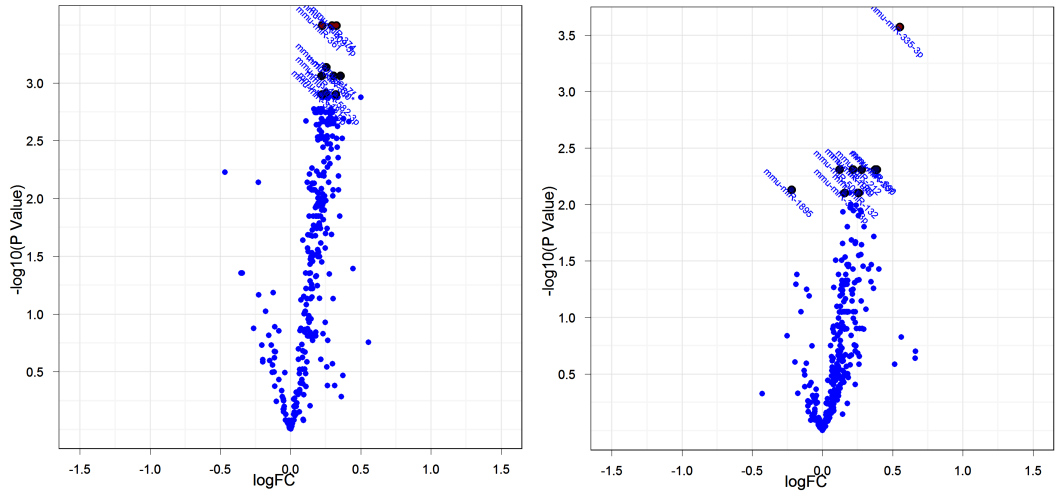

Figure B.5: Volcano plots show the relation between the logarithm (base 10) of fold change between treatment and control groups on the x-axis and the negative logarithm of the p-values on the y-axis. The top selected microRNAs are marked with annotation on the plot. The plots show the comparisons 6-Hertz 3 hours versus 0 hours (left) and 6-Hertz 6 hours versus 0 hours (right).

B.4 Supplementary Tables & Excel Files

Table B.1: Performance of the *double detection procedure*, without (A) and with (B) the use of empirical Bayes method, based on simulated data. The method is compared to the unmoderated and the moderated t-tests and the nonparametric Mann-Whitney test. The table gives the sensitivity, specificity, Positive Predictive values (PPv) and Negative Predictive values (NPv). Standard deviations of these measures are given beside each value.

Test Type	Sensitivity	specificity	PPv	NPv
Unmoder. T.test	0.650 ± 0.021	0.978 ± 0.001	0.500 ± 0.000	0.928 ± 0.002
Moderated T.test	0.833 ± 0.005	0.971 ± 0.006	0.523 ± 0.020	0.949 ± 0.007
Mann-Whitney	0.800 ± 0.001	0.973 ± 0.000	0.500 ± 0.000	0.953 ± 0.005
Double Detection A	0.834 ± 0.011	0.988 ± 0.005	0.520 ± 0.000	0.983 ± 0.008
Double Detection B	0.847 ± 0.033	0.985 ± 0.004	0.554 ± 0.045	0.988 ± 0.002

Additional tables in excel files contain multiple sheets. The excel files can be found in the attached compact disc.

Excel file 1 Most differentially regulated miRNAs in the Pilocarpine model at 24 h and 28 d following SE

Excel file 2 Most differentially regulated miRNAs in the SSSE model at 24 h and 28 d following SE

Excel file 3 Most differentially regulated miRNAs in the 6 Hz model at 3 h and 6 h following single seizure

Excel file 4 Overlapping miRNAs between pilocarpine and SSSE model at 24 h and 28 d

APPENDIX C

CHAPTER 6 APPENDICES

C.1 Supplementary Figures

Figures C.1, C.2, C.3, C.4 & C.5 show further plots, which are referenced in the main text.

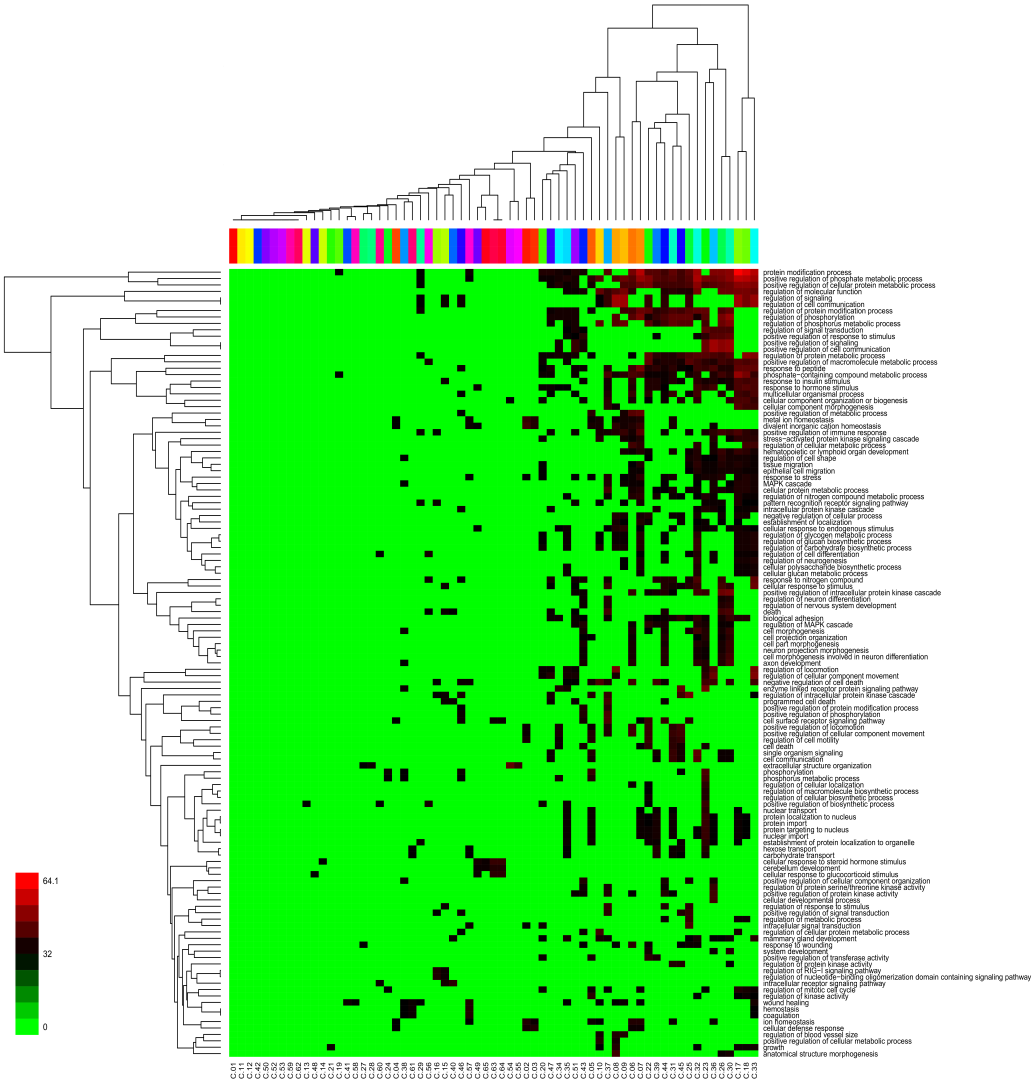

Figure C.1: Gene Ontology (biological processes) terms enriched in target proteins. The heatmap depicts negative log(FDR) values at a 1% cutoff. Individual GO terms can be found in Supplementary Excel Files

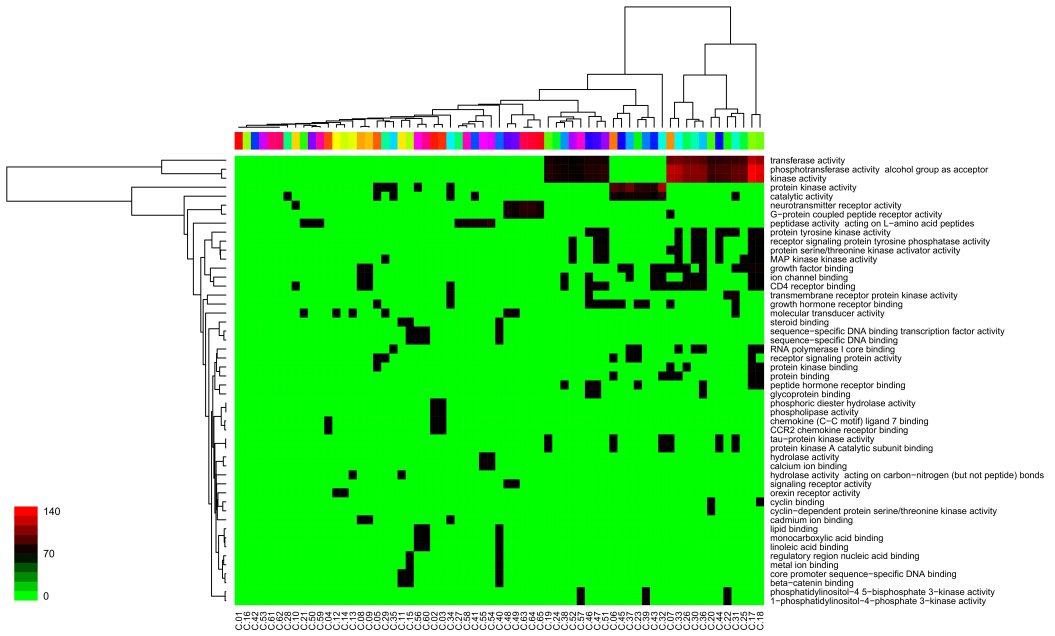

Figure C.2: Gene Ontology (molecular functions) terms enriched in target proteins. The heatmap depicts negative log(FDR) values at a 1% cutoff.

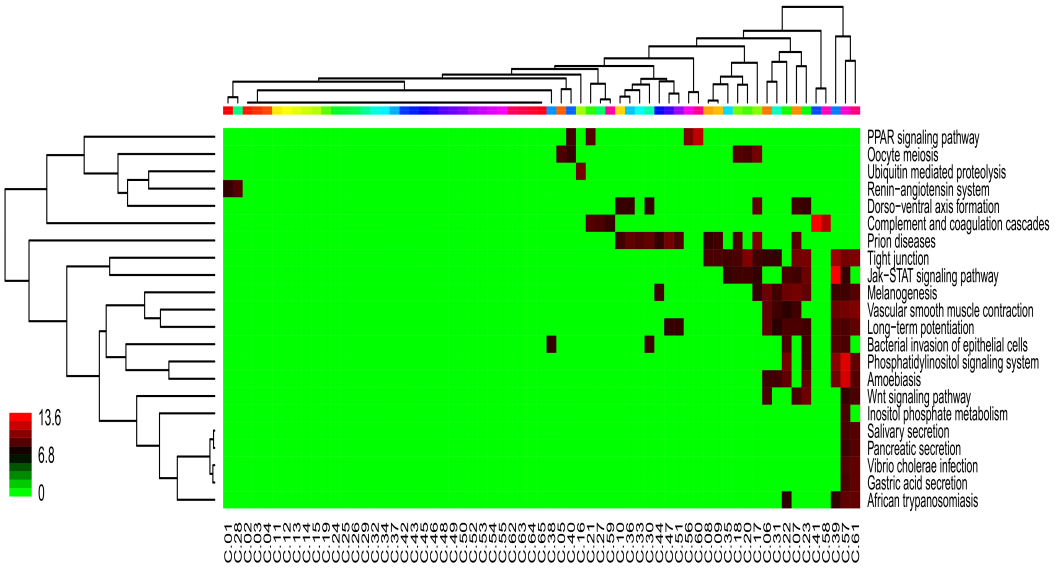

Figure C.3: KEGG pathways enriched in target proteins. The heatmap depicts negative log(FDR) values at a 1% cutoff.

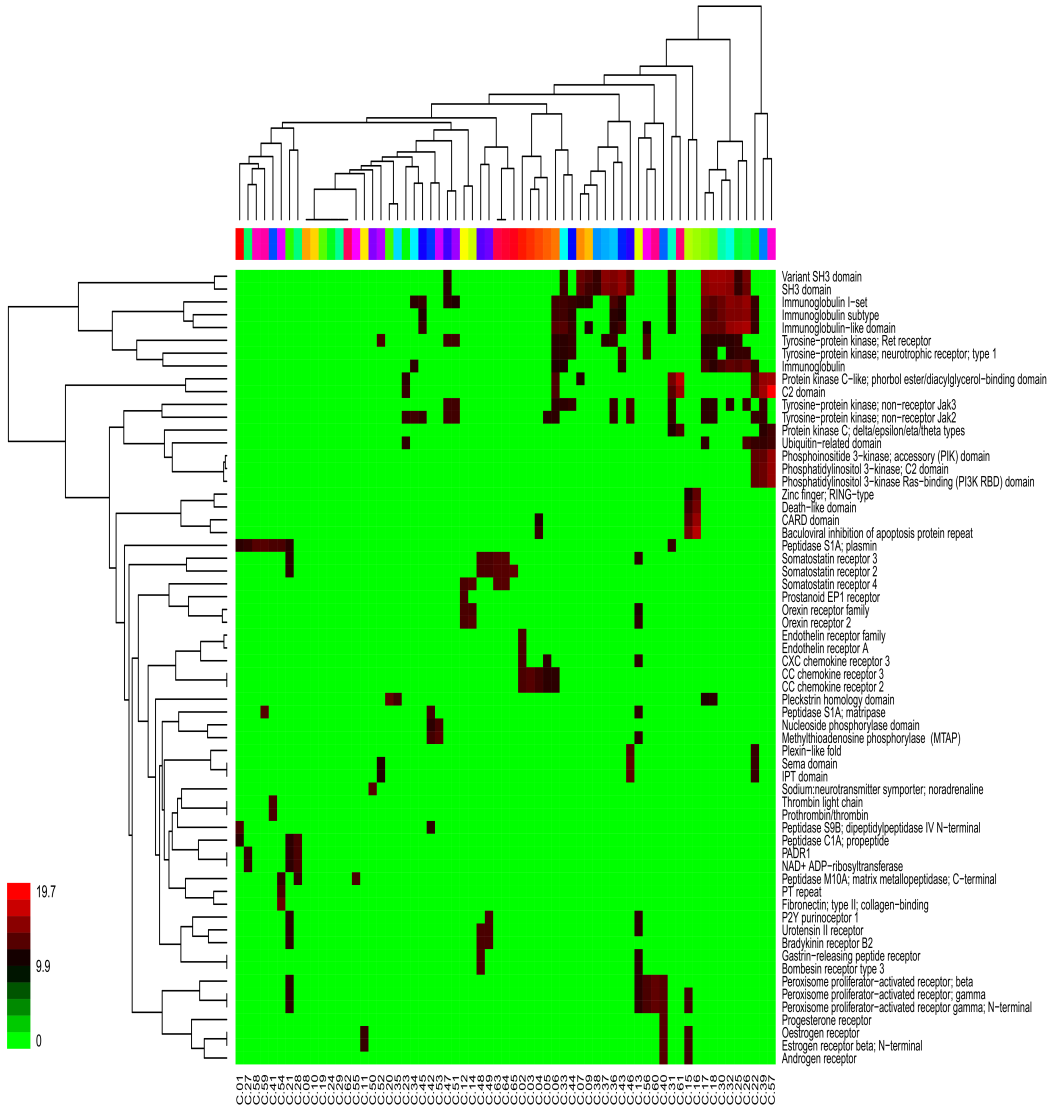

Figure C.4: Protein domains enriched in target proteins. The heatmap depicts negative log(FDR) values at a 1% cutoff.

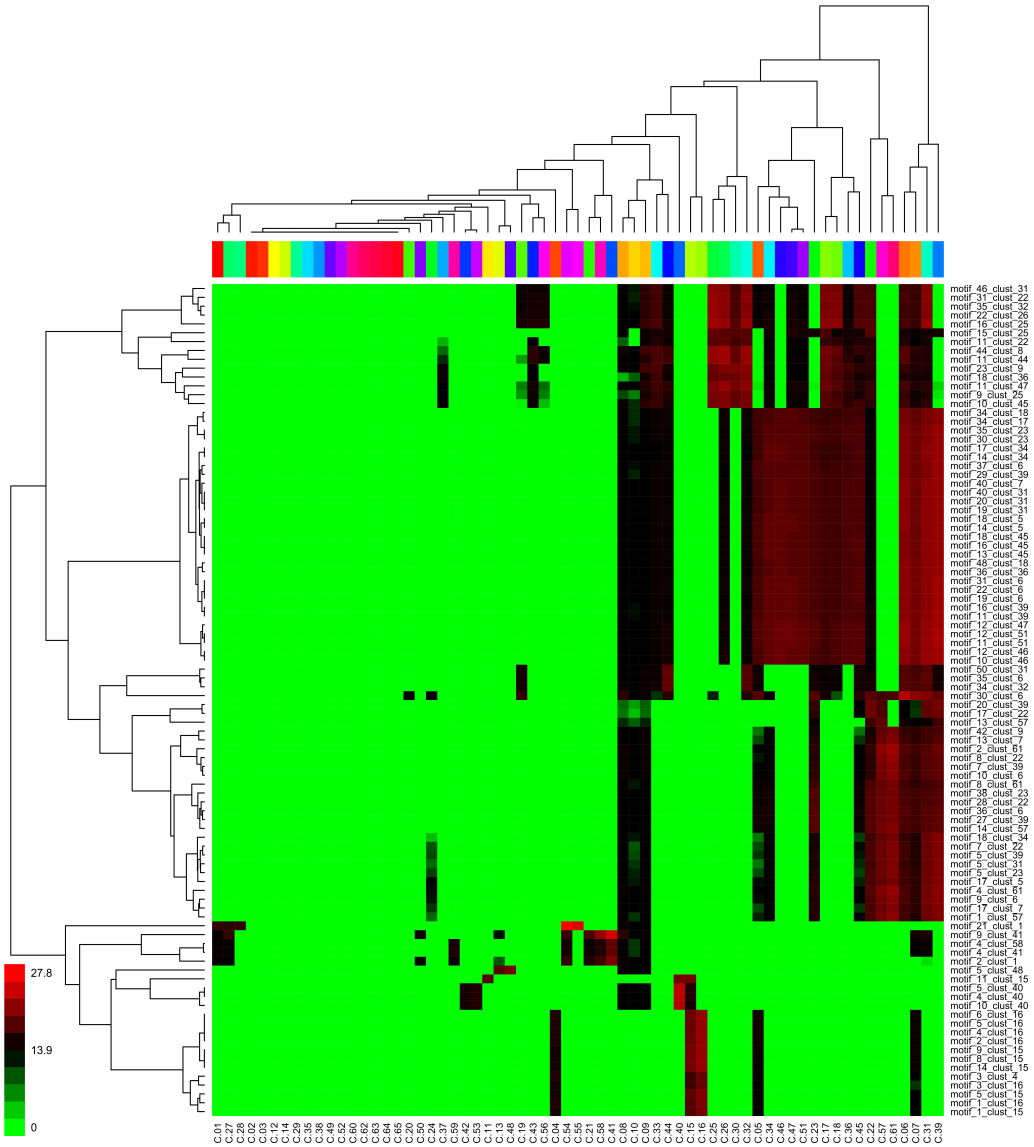

Figure C.5: Sequence motifs enriched in target proteins. The heatmap depicts negative log E-values at a 1% cutoff. Individual sequence motifs are accessible in Excel file: Compounds-BES-Clusters.Sequence-Motifs-Enrichment.

C.2 Supplementary Tables

Table C.1: Summary table of BES clustering with number of compounds per cluster and associated targets. The number of targets linked to HIV, cancer or both is reported. Clusters with targets associated to both diseases and or have silhouette width are called high quality *repurposing clusters* (these are 15 clusters; cluster 2-7, 11, 13, 15-21; orange-shaded in the table). The number of enriched KEGG pathways, GO terms, protein domains and sequence motifs per cluster is shown at a 1% FDR cutoff. Categories being enriched in the overall dataset have been removed. The last column reports the silhouette width of each cluster: 61 out of the 65 resulting clusters are none-singleton.

Cluster	Compd	Target	Target HIV	Target Cancer	Target Both	KEGG Path	GO (bp) Term	GO (mf) Term	Protein Domain	Motif	Silh. Width
c01	228	20	2	19	1	0	0	1	2	22	0.39
c02	60	6	3	4	1	0	13	5	5	0	0.89
c03	34	4	3	2	1	0	8	4	2	0	0.91
c04	20	5	4	2	1	0	1	0	4	9	0.95
c05	72	22	4	19	1	0	26	4	0	30	0.78
c06	63	31	3	28	0	3	50	6	5	83	0.46
c07	98	72	8	66	2	3	63	7	2	71	0.37
c08	324	105	8	98	1	0	29	2	0	10	0.05
c09	267	98	9	90	1	1	30	1	2	12	-0.17
c10	1944	158	16	146	4	0	17	0	0	0	0.00
c11	36	14	3	13	2	0	4	2	4	6	0.63
c12	27	3	1	2	0	0	0	3	7	0	-0.07
c13	77	30	1	29	0	0	1	3	0	2	0.32
c14	9	4	1	3	0	0	0	2	3	0	-0.78
c15	48	9	3	8	2	0	15	4	8	22	0.84
c16	22	2	2	2	2	1	17	3	4	14	0.27
c17	11	47	1	46	0	3	84	8	6	53	0.53
c18	17	42	1	41	0	1	68	9	5	46	0.60

Continued on next page

Table C.1 – *Continued from previous page*

Cluster	Compd	Target	Target HIV	Target Cancer	Target Both	KEGG Path	GO (bp) Term	GO (mf) Term	Protein Domain	Motif	Silh. Width
c19	8	8	1	7	0	0	0	4	1	3	1.00
c20	12	13	1	12	0	1	21	5	1	0	0.98
c21	82	28	1	28	1	2	1	0	0	9	0.46
c22	141	20	0	20	0	3	52	7	16	70	0.65
c23	94	23	0	23	0	5	90	6	0	65	0.70
c24	33	6	0	6	0	0	1	3	0	0	0.52
c25	16	19	0	19	0	0	40	5	4	21	0.98
c26	6	26	0	26	0	0	58	6	7	21	0.95
c27	75	13	0	13	0	0	0	2	0	9	0.43
c28	43	7	0	7	0	1	4	0	8	10	0.49
C29	33	4	0	4	0	0	4	3	2	0	0.97
c30	6	26	0	26	0	1	62	8	6	22	0.98
c31	43	28	0	28	0	0	55	10	12	94	0.58
c32	15	30	0	30	0	0	96	7	6	29	0.46
c33	16	33	0	33	0	1	84	7	5	21	-0.17
c34	25	9	0	9	0	0	32	8	3	32	-0.11
c35	9	5	0	5	0	0	49	2	1	0	0.89
c36	6	25	0	25	0	1	52	10	3	44	0.54
c37	3	20	0	20	0	0	45	4	2	7	0.79
c38	2	5	0	5	0	0	2	4	1	0	0.64
c39	60	17	0	17	0	7	50	5	12	75	0.48
C40	100	7	0	7	0	0	14	12	7	8	0.95
c41	43	5	0	5	0	1	4	1	7	9	0.76
c42	13	4	0	4	0	0	0	0	2	0	0.11
c43	1	8	0	8	0	0	31	7	6	3	0.00
c44	3	15	0	15	0	0	45	8	6	18	0.67
c45	49	16	0	16	0	0	60	3	11	49	0.86
c46	13	8	0	8	0	0	14	11	8	32	0.81

Continued on next page

Table C.1 – *Continued from previous page*

Cluster	Compd	Target	Target HIV	Target Cancer	Target Both	KEGG Path	GO (bp) Term	GO (mf) Term	Protein Domain	Motif	Silh. Width
c47	4	9	0	9	0	1	23	13	4	35	0.51
C48	102	8	0	8	0	0	2	10	9	2	0.73
c49	22	7	0	7	0	0	6	7	6	0	0.46
c50	39	3	0	3	0	0	0	1	2	0	0.63
c51	2	10	0	10	0	0	26	7	4	33	1.00
c52	1	4	0	4	0	0	0	4	1	0	0.00
c53	25	2	0	2	0	0	0	1	2	3	0.35
c54	80	10	0	10	0	0	1	3	5	4	-0.01
c55	35	5	0	5	0	0	1	3	1	1	0.48
c56	2	6	0	6	0	1	9	1	6	1	0.53
c57	2	5	0	5	0	7	8	5	8	46	1.00
c58	2	4	0	4	0	1	0	1	5	5	0.82
c59	5	2	0	2	0	0	0	2	4	4	-0.47
C60	1	3	0	3	0	1	13	5	3	0	0.00
c61	2	3	0	3	0	11	4	0	3	35	1.00
c62	12	2	0	2	0	0	0	0	0	0	0.62
c63	7	5	0	5	0	0	13	2	3	0	0.98
c64	5	5	0	5	0	0	13	2	3	0	0.87
c65	1	3	0	3	0	0	6	2	1	0	0.00

Table C.2: Top enriched biological vocabularies. Top three significantly enriched KEGG pathways (KEGG:), GO terms (GO:) and protein domains (IPR:) in 15 *repurposing clusters*.

ID	Term	P.BY
	Clust02	
KEGG:4062	Chemokine signaling pathway	3.49E-06
KEGG:4060	Cytokine-cytokine receptor interaction	9.48E-06
GO:0006874	cellular calcium ion homeostasis	5.54E-08
GO:0072507	divalent inorganic cation homeostasis	5.54E-08
GO:0070098	chemokine-mediated signaling pathway	9.74E-08
IPR000355	Chemokine receptor family	3.69E-13
IPR002236	CC chemokine receptor 1	4.42E-10
IPR000276	G protein-coupled receptor; rhodopsin-like	2.28E-08
	Clust03	
KEGG:4062	Chemokine signaling pathway	8.68E-06
KEGG:4060	Cytokine-cytokine receptor interaction	1.69E-05
GO:0070098	chemokine-mediated signaling pathway	1.32E-08
GO:0002407	dendritic cell chemotaxis	3.89E-06
GO:0006874	cellular calcium ion homeostasis	4.37E-05
IPR000355	Chemokine receptor family	4.09E-11
IPR002236	CC chemokine receptor 1	5.43E-11
IPR002240	CC chemokine receptor 5	3.52E-07
	Clust04	
GO:0070098	chemokine-mediated signaling pathway	1.12E-04
GO:0090026	positive regulation of monocyte chemotaxis	3.82E-03
GO:0002407	dendritic cell chemotaxis	8.46E-03
IPR000355	Chemokine receptor family	8.47E-07
IPR002236	CC chemokine receptor 1	5.35E-06
	Clust05	
KEGG:5200	Pathways in cancer	1.13E-06
KEGG:4062	Chemokine signaling pathway	2.85E-05
KEGG:4012	ErbB signaling pathway	5.44E-05
GO:0018193	peptidyl-amino acid modification	9.80E-08
GO:0016310	phosphorylation	9.80E-08
GO:0043549	regulation of kinase activity	1.31E-06
IPR001245	Serine-threonine/tyrosine-protein kinase catalytic dom	1.08E-19
IPR000719	Protein kinase domain	1.08E-19
IPR002290	Serine/threonine/dual specificity prot kinase; catalytic	1.08E-19

Continued on next page

Table C.2 – *Continued from previous page*

ID	Term	P.BY
	Clust06	
KEGG:5200	Pathways in cancer	1.66E-11
KEGG:4012	ErbB signaling pathway	2.19E-09
KEGG:4062	Chemokine signaling pathway	1.02E-08
GO:0018193	peptidyl-amino acid modification	6.31E-14
GO:0036211	protein modification process	1.69E-13
GO:0009893	positive regulation of metabolic process	3.76E-11
IPR011009	Protein kinase-like domain	4.03E-36
IPR020635	Tyrosine-protein kinase; catalytic domain	1.10E-35
IPR001245	Serine-threonine/tyrosine-protein kinase catalytic dom	1.50E-35
	Clust07	
KEGG:5200	Pathways in cancer	1.81E-21
KEGG:4012	ErbB signaling pathway	3.31E-16
KEGG:5215	Prostate cancer	3.42E-16
GO:0040011	locomotion	7.11E-25
GO:0046777	protein autophosphorylation	7.11E-25
GO:0006928	cellular component movement	3.75E-23
IPR020635	Tyrosine-protein kinase; catalytic domain	5.99E-58
IPR001245	Serine-threonine/tyrosine-protein kinase catalytic dom	1.32E-57
IPR002290	Serine/threonine/dual specificity prot kinase; catalytic	1.32E-57
	Clust11	
KEGG:140	Steroid hormone biosynthesis	8.98E-03
KEGG:5220	Chronic myeloid leukemia	9.92E-03
GO:0010870	positive regulation of receptor biosynthetic process	4.76E-04
GO:0061198	fungiform papilla formation	3.86E-03
GO:0008209	androgen metabolic process	3.86E-03
IPR023801	Histone deacetylase domain	5.02E-09
IPR000286	Histone deacetylase superfamily	5.02E-09
IPR002397	Cytochrome P450; B-class	9.69E-05
	Clust13	
KEGG:4080	Neuroactive ligand-receptor interaction	9.58E-04
GO:0010870	positive regulation of receptor biosynthetic process	4.87E-03
GO:0043112	receptor metabolic process	4.87E-03
GO:0006367	transcription initiat from RNA polymerase II promoter	4.87E-03
IPR023801	Histone deacetylase domain	3.48E-07
IPR000286	Histone deacetylase superfamily	3.48E-07
IPR003074	Peroxisome proliferator-activated receptor	5.10E-07

Continued on next page

Table C.2 – *Continued from previous page*

ID	Term	P.BY
	Clust15	
KEGG:5200	Pathways in cancer	7.06E-04
KEGG:4621	NOD-like receptor signaling pathway	1.12E-03
GO:0006367	transcription initiation from RNA polymerase II promoter	1.57E-04
GO:0030522	intracellular receptor signaling pathway	6.31E-04
GO:2000116	regulation of cysteine-type endopeptidase activity	1.69E-03
IPR001628	Zinc finger; nuclear hormone receptor-type	2.21E-10
IPR000536	Nuclear hormone receptor; ligand-binding; core	2.21E-10
IPR008946	Nuclear hormone receptor; ligand-binding	2.21E-10
	Clust16	
KEGG:4621	NOD-like receptor signaling pathway	1.31E-03
KEGG:5222	Small cell lung cancer	1.31E-03
KEGG:4210	Apoptosis	1.31E-03
GO:0070424	reg. nucleotide-bind oligomerization dom contain SP	4.10E-05
GO:0039535	regulation of RIG-I signaling pathway	4.30E-05
GO:0039528	cytoplasmic pattern recognit recept SP respo to virus	1.07E-04
IPR001370	Baculoviral inhibition of apoptosis protein repeat	1.24E-06
IPR001315	CARD domain	9.60E-06
IPR011029	Death-like domain	7.28E-05
	Clust17	
KEGG:5200	Pathways in cancer	3.05E-14
KEGG:4012	ErbB signaling pathway	1.15E-13
KEGG:5215	Prostate cancer	1.15E-13
GO:0046777	protein autophosphorylation	2.71E-35
GO:0036211	protein modification process	1.46E-28
GO:0018193	peptidyl-amino acid modification	3.41E-24
IPR001245	Serine-threonine/tyrosine-protein kinase catalytic dom	5.44E-75
IPR002290	Serine/threonine/dual specificity prot kinase; catalytic	5.44E-75
	Clust18	
KEGG:5200	Pathways in cancer	4.05E-13
KEGG:4012	ErbB signaling pathway	6.87E-13
KEGG:5215	Prostate cancer	7.15E-10
GO:0046777	protein autophosphorylation	3.16E-34
GO:0036211	protein modification process	7.91E-25
GO:0040011	locomotion	5.39E-20
IPR020635	Tyrosine-protein kinase; catalytic domain	4.64E-67
IPR001245	Serine-threonine/tyrosine-protein kinase catalytic dom	8.15E-67
IPR002290	Serine/threonine/dual specificity prot kinase; catalytic	8.15E-67

Continued on next page

Table C.2 – *Continued from previous page*

ID	Term	P.BY
	Clust19	
GO:0002764	immune response-regulating signaling pathway	3.16E-03
GO:0038095	Fc-epsilon receptor signaling pathway	3.16E-03
IPR000719	Protein kinase domain	2.72E-12
IPR001245	Serine-threonine/tyrosine-protein kinase catalytic dom	2.72E-12
IPR002290	Serine/threonine/dual specificity prot kinase; catalytic	2.72E-12
	Clust20	
KEGG:5222	Small cell lung cancer	1.98E-06
KEGG:4062	Chemokine signaling pathway	2.83E-06
KEGG:4660	T cell receptor signaling pathway	2.83E-06
GO:0036211	protein modification process	4.06E-05
GO:0046777	protein autophosphorylation	9.14E-04
GO:0010604	posit regulation of macromolecule metabolic process	2.00E-03
IPR000719	Protein kinase domain	4.01E-20
IPR001245	Serine-threonine/tyrosine-protein kinase catalytic dom	4.01E-20
IPR002290	Serine/threonine/dual specificity prot kinase; catalytic	4.01E-20
	Culst21	
KEGG:4080	Neuroactive ligand-receptor interaction	1.11E-03
KEGG:140	Steroid hormone biosynthesis	5.15E-05
KEGG:4610	Complement and coagulation cascades	6.18E-03
GO:0040007	growth	6.14E-03
IPR003074	Peroxisome proliferator-activated receptor	1.22E-06
IPR022321	Insulin-like growth factor-bind prot family 1-6; chordata	1.33E-05

C.2.1 Supplementary Excel Files

Additional tables in excel files containing multiple sheets. The excel file can be found in the attached compact disc

Excel file 1 Gene Ontology (Biological Processes) enrichment analysis for targets of compounds in BES clusters.

Excel file 2 Gene Ontology enrichment (Molecular Function) analysis for targets of compounds in BES clusters.

Excel file 3 KEGG pathway enrichment analysis for targets of compounds in BES clusters.

Excel file 4 Enrichment analysis of protein domains in targets of compounds in BES clusters.

Excel file 5 Enrichment of sequence motifs in targets of compounds in BES clusters.

Excel file 6 Top enriched Gene Ontology terms (biological processes) in different clusters.

Excel file 7 Literature mining results for individual target proteins.

Bibliography

Abnaof, K., Mallela, N., Walenda, G., Meurer, S. K., Seré, K., Lin, Q., Smeets, B. et al. (2014). TGF-β stimulation in human and murine cells reveals commonly affected biological processes and pathways at transcription level. *BMC systems biology*, volume 8 (1): 55.

Abu-Jamous, B., Fa, R., Roberts, D. J. and Nandi, A. K. (2013a). Paradigm of Tunable Clustering Using Binarization of Consensus Partition Matrices (Bi-CoPaM) for Gene Discovery. *PLoS ONE*, volume 8 (2).

Abu-Jamous, B., Fa, R., Roberts, D. J. and Nandi, A. K. (2013b). Yeast gene CMR1/YDL156W is consistently co-expressed with genes participating in DNA-metabolic processes in a variety of stringent clustering experiments. *Journal of the Royal Society, Interface / the Royal Society*, volume 10 (81): 20120990.

Aden, D. P., Fogel, A., Plotkin, S., Damjanov, I. and Knowles, B. B. (1979). Controlled synthesis of HBsAg in a differentiated human liver carcinoma-derived cell line. *Nature*, volume 282 (5739): 615–616.

Agresti, A. (1996). An Introduction to Categorical Data Analysis. *Pharmaceutical Statistics*, volume 7 (4): 307–307.

Agresti, A. (2002). Categorical Data Analysis. *Wiley series in probability and statistics*, volume 45 (1): xv, 710 p. ST – Categorical data analysis.

Akobeng, A. K. (2007). Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta paediatrica*, volume 96 (3): 338–341.

Aldrich, J. (1997). R.A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, volume 12 (3): 162–176.

Alexa, A., Rahnenführer, J. and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, volume 22 (13): 1600–1607.

Alexandrow, M. G. and Moses, H. L. (1995). Transforming growth factor beta 1 inhibits mouse keratinocytes late in G1 independent of effects on gene transcription. *Cancer Research*, volume 55 (17): 3928–3932.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, volume 215 (3): 403–410.

Altshuler, B. (1981). Modeling of dose-response relationships.

Andersen, E. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, volume 32 (2): 283–301.

Aravin, A. and Tuschl, T. (2005). Identification and characterization of small RNAs involved in RNA silencing.

Arthur, L. (2010a). hugene10sttranscriptcluster.db: Affymetrix Mouse Gene 1.0-ST Array Transcriptcluster Revision 8 annotation data (chip hugene10sttranscriptcluster). R package version 8.0.1. *www.bioconductor.org*.

Arthur, L. (2010b). mogene10sttranscriptcluster.db: Affymetrix Mouse Gene 1.0-ST Array Transcriptcluster Revision 8 annotation data (chip mogene10sttranscriptcluster). R package version 8.0.1. *www.bioconductor.org*.

Aryee, M. J., Gutiérrez-Pabello, J. A., Kramnik, I., Maiti, T. and Quackenbush, J. (2009). An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation). *BMC Bioinformatics*, volume 10 (1): 409.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P. et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, volume 25 (1): 25–29.

Azmi, A. S. (2012). *Systems Biology in Cancer Research and Drug Discovery*. Springer Science & Business Media.

Backman, T. W. H., Cao, Y. and Girke, T. (2011). ChemMine tools: An online service for analyzing and clustering small molecules. *Nucleic Acids Research*, volume 39 (SUPPL. 2).

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J. et al. (2009). MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Research*, volume 37 (SUPPL. 2).

Bajorath, J. (2014). Evolution of the activity cliff concept for structure–activity relationship analysis and drug discovery. *Future Medicinal Chemistry*, volume 6 (14): 1545–1549.

Bakheet, T. M. and Doig, A. J. (2009). Properties and identification of human protein drug targets. *Bioinformatics (Oxford, England)*, volume 25 (4): 451–457.

Baldi, P. and Long, A. D. (2001a). A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics (Oxford, England)*, volume 17 (6): 509–519.

Baldi, P. and Long, A. D. (2001b). Microarray Expression Data : Regularized T-Test. *Bioinformatics*, volume 17 (6): 509–519.

Baltimore, B. D., Berg, P., Botchan, M., Carroll, D., Charo, R. A., Church, G., Corn, J. E. et al. (2015). A prudent path forward for genomic engineering and germline gene modification. *Science*, volume 348 (6230): 36–38.

Bar-Joseph, Z., Gerber, G., Simon, I., Gifford, D. K. and Jaakkola, T. S. (2003). Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proceedings of the National Academy of Sciences of the United States of America*, volume 100 (18): 10146–10151.

Bauer, S., Koller, M., Cepok, S., Todorova-Rudolph, A., Nowak, M., Nockher, W. A., Lorenz, R. et al. (2008). NK and CD4+ T cell changes in blood after seizures in temporal lobe epilepsy. *Exp Neurol*, volume 211 (2): 370–377.

Bauer-Mehren, A., Bundschus, M., Rautschka, M., Mayer, M. A., Sanz, F. and Furlong, L. I. (2011). Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PLoS ONE*, volume 6 (6).

Beaumont, M. A. and Rannala, B. (2004). The Bayesian revolution in genetics. *Nature reviews. Genetics*, volume 5 (4): 251–61.

Ben-Hur, A., Elisseeff, A. and Guyon, I. (2002). A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 6–17.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological*, volume 57 (1): 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple hypothesis testing under dependency. *Annals of Statistics*, volume 29 (4): 1165–1188.

Berger, E. A., Murphy, P. M. and Farber, J. M. (1999). Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. *Annual review of immunology*, volume 17: 657–700.

Berger, J. A., Hautaniemi, S., Järvinen, A.-K., Edgren, H., Mitra, S. K. and Astola, J. (2004). Optimized LOWESS normalization parameter selection for DNA microarray data. *BMC bioinformatics*, volume 5: 194.

Berry, M. W. and Castellanos, M. (2007). *Survey of Text Mining II: Clustering, Classification, and Retrieval*. Springer Science & Business Media.

Birbrair, A., Zhang, T., Wang, Z.-M., Messi, M. L., Mintz, A. and Delbono, O. (2013). Type-1 pericytes participate in fibrous tissue deposition in aged skeletal muscle. *American journal of physiology. Cell physiology*, volume 305 (11): C1098–113.

Bisswanger, H. (2002). *Enzyme Kinetics Principles and Methods*. Wiley-Blackwell.

Bleakley, K. and Yamanishi, Y. (2009). Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics*, volume 25 (18): 2397–2403.

Boes, T. and Neuhäuser, M. (2005). Normalization for Affymetrix GeneChips.

Bolstad, B. M., Irizarry, R. A., Astrand, M. and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, volume 19 (2): 185–193.

Bonferroni, C. (1935). Il calcolo delle assicurazioni su gruppi di teste. *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13 – 60.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Bork, P. and Koonin, E. V. (1996). Protein sequence motifs. *Current opinion in structural biology*, volume 6 (3): 366–376.

Borkowski, T. A., Letterio, J. J., Farr, A. G. and Udey, M. C. (1996). A role for endogenous transforming growth factor beta 1 in Langerhans cell biology: the skin of transforming growth factor beta 1 null mice is devoid of epidermal Langerhans cells. *The Journal of experimental medicine*, volume 184 (6): 2417–22.

Brader, M. L., Kaarsholm, N. C., Harnung, S. E. and Dunn, M. F. (1997). Ligand perturbation effects on a pseudotetrahedral Co(II)(His)3-ligand site: A magnetic circular dichroism study of the Co(II)-substituted insulin hexamer. *Journal of Biological Chemistry*, volume 272 (2): 1088–1094.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J. et al. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature genetics*, volume 29 (4): 365–371.

Breitling, R., Amtmann, A. and Herzyk, P. (2004). Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC bioinformatics*, volume 5: 34.

Brouwers, L., Iskar, M., Zeller, G., van Noort, V. and Bork, P. (2011). Network neighbors of drug targets contribute to drug side-effect similarity. *PLoS ONE*, volume 6 (7).

Buenemann, C. L., Willy, C., Buchmann, A., Schmiechen, A. and Schwarz, M. (2001). Transforming growth factor-beta1-induced Smad signaling, cell-cycle arrest and apoptosis in hepatoma cells. *Carcinogenesis*, volume 22 (3): 447–452.

Buse, A. (1982). The likelihood ratio, wald, and lagrange multiplier tests: an expository note. *The American Statistician*, volume 36 (3): 153–157.

Busschaert, P., Geeraerd, a. H., Uyttendaele, M. and Van Impe, J. F. (2010). Estimating distributions out of qualitative and (semi)quantitative microbiological contamination data for use in risk assessment. *International Journal of Food Microbiology*, volume 138 (3): 260–269.

Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R. et al. (2009). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical chemistry*, volume 55 (4): 611–22.

Butina, D., Segall, M. D. and Frankcombe, K. (2002). Predicting ADME properties in silico: Methods and models.

Byvatov, E., Fechner, U., Sadowski, J. and Schneider, G. (2003). Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. *Journal of Chemical Information and Computer Sciences*, volume 43 (6): 1882–1889.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D. et al. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic acids research*, volume 32 (Database issue): D262–D266.

Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J. and Bork, P. (2008). Drug target identification using side-effect similarity. *Science (New York, N.Y.)*, volume 321 (5886): 263–266.

Casella, G. (1985). An Introduction to Empirical Bayes Data Analysis. *American Statistician*, volume 39 (2): 83–87.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, z., Anwar, N., Schultz, N. et al. (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, volume 39 (SUPPL. 1).

Chang, T. W. (1983). Binding of cells to matrixes of distinct antibodies coated on solid surface. *Journal of immunological methods*, volume 65 (1-2): 217–223.

Chen, S.-J., Ning, H., Ishida, W., Sodin-Semrl, S., Takagawa, S., Mori, Y. and Varga, J. (2006). The early-immediate gene EGR-1 is induced by transforming growth factor-beta and mediates stimulation of collagen gene expression. *The Journal of Biological Chemistry*, volume 281 (30): 21183–21197.

Chen, X. (2009). Curve-based clustering of time course gene expression data using self-organizing maps. *Journal of bioinformatics and computational biology*, volume 7 (4): 645–661.

Chen, X., Liang, H., Zhang, J., Zen, K. and Zhang, C. Y. (2012). Secreted microRNAs: A new form of intercellular communication.

Chen, Y.-J., Kodell, R., Sistare, F., Thompson, K. L., Morris, S. and Chen, J. J. (2003). Normalization methods for analysis of microarray gene-expression data. *Journal of biopharmaceutical statistics*, volume 13 (1): 57–74.

Cheng, A. C., Coleman, R. G., Smyth, K. T., Cao, Q., Soulard, P., Caffrey, D. R., Salzberg, A. C. et al. (2007). Structure-based maximal affinity model predicts small-molecule druggability. *Nature biotechnology*, volume 25 (1): 71–75.

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W. et al. (2012). Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Computational Biology*, volume 8 (5).

Cheng, Y. and Prusoff, W. H. (1973). Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochemical pharmacology*, volume 22 (23): 3099–108.

Cho, C. H. (2011). Frontier of epilepsy research - mTOR signaling pathway. *Experimental & molecular medicine*, volume 43 (5): 231–274.

Chu, A. S., Diaz, R., Hui, J.-J., Yanger, K., Zong, Y., Alpini, G., Stanger, B. Z. et al. (2011). Lineage tracing demonstrates no evidence of cholangiocyte epithelial-to-mesenchymal transition in murine models of hepatic fibrosis. *Hepatology*, volume 53 (5): 1685–1695.

Chu, L.-H. and Chen, B.-S. (2008). Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets. *BMC systems biology*, volume 2 (1): 56.

Chu, T. M., Weir, B. and Wolfinger, R. (2002). A systematic statistical linear modeling approach to oligonucleotide array experiments. *Mathematical Biosciences*, volume 176 (1): 35–51.

Clark, R. A., McCoy, G. A., Folkvord, J. M. and McPherson, J. M. (1997). TGF-beta 1 stimulates cultured human fibroblasts to proliferate and produce tissue-like fibroplasia: a fibronectin matrix-dependent event. *Journal of cellular physiology*, volume 170 (1): 69–80.

Coffey, N., Hinde, J. and Holian, E. (2014). Clustering longitudinal profiles using P-splines and mixed effects models applied to time-course gene expression data. *Computational Statistics and Data Analysis*, volume 71: 14–19.

Cohen, S. B., Zheng, G., Heyman, H. C. and Stavnezer, E. (1999). Heterodimers of the SnoN and Ski oncoproteins form preferentially over homodimers and are more potent transforming agents. *Nucleic Acids Research*, volume 27 (4): 1006–1014.

Commeau, N., Parent, E., Delignette-Muller, M. L. and Cornu, M. (2012). Fitting a lognormal distribution to enumeration and absence/presence data. *International Journal of Food Microbiology*, volume 155 (3): 146–152.

Costa, I. G., de Carvalho, F. d. A. T. and de Souto, M. C. P. (2004). Comparative analysis of clustering methods for gene expression time course data. *Genetics and Molecular Biology*, volume 27 (4): 623–631.

Cover, T. L. and Blanke, S. R. (2005). Helicobacter pylori VacA, a paradigm for toxin multifunctionality. *Nature reviews. Microbiology*, volume 3 (4): 320–332.

Cramer, G. M., Ford, R. A. and Hall, R. L. (1978). Estimation of toxic hazard–a decision tree approach. *Food and cosmetics toxicology*, volume 16 (3): 255–276.

Cramer, R. D., Patterson, D. E. and Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*, volume 110 (18): 5959–5967.

Cramer, R. D., Patterson, D. E. and Bunce, J. D. (1989). Recent advances in comparative molecular field analysis (CoMFA). *Progress in clinical and biological research*, volume 291: 161–165.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions - Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, volume 31 (4): 377–403.

Creemers, E. E., Tijsen, A. J. and Pinto, Y. M. (2012). Circulating MicroRNAs: Novel biomarkers and extracellular communicators in cardiovascular disease?

Crump, K. S., Hoel, D. G., Langley, C. H. and Peto, R. (1976). Fundamental carcinogenic processes and their implications for low dose risk assessment. *Cancer Research*, volume 36 (9).

David Andersson, C., Chen, B. Y. and Linusson, A. (2011). Erratum: Mapping of ligand-binding cavities in proteins.

de Gruyter, W. (1996). *Multivariate statistische Verfahren*. Walter de Gruyter.

De Jong, L. A. A., Uges, D. R. A., Franke, J. P. and Bischoff, R. (2005). Receptor-ligand binding assays: Technologies and applications.

de Jonge, H. J. M., Fehrmann, R. S. N., de Bont, E. S. J. M., Hofstra, R. M. W., Gerbens, F., Kamps, W. A., de Vries, E. G. E. et al. (2007). Evidence based selection of housekeeping genes. *PLoS ONE*, volume 2 (9).

Dehmer, M., Emmert-Streib, F., Graber, A. and Salvador, A. (Editors) (2011). *Applied Statistics for Network Biology: Methods in Systems Biology*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

Delignette-muller, M. L. and Dutang, C. (2015). fitdistrplus : An R Package for Fitting Distributions. *Journal of Statistical Software*, volume 64 (4): 1–34.

Dembélé, D. (2013). A Flexible Microarray Data Simulation Model. *Microarrays*, volume 2 (2): 115–130.

Deng, Y., He, Z., Van Nostrand, J. D. and Zhou, J. (2008). Design and analysis of mismatch probes for long oligonucleotide microarrays. *BMC genomics*, volume 9: 491.

Di Stefano, V., Zaccagnini, G., Capogrossi, M. C. and Martelli, F. (2011). MicroRNAs as peripheral blood biomarkers of cardiovascular disease.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., Einecke, G. et al. (2009). Gene-set analysis and reduction. *Briefings in Bioinformatics*, volume 10 (1): 24–34.

Do, J. H. and Choi, D.-K. (2006). Normalization of microarray data: single-labeled and dual-labeled arrays. *Molecules and cells*, volume 22 (3): 254–261.

Dominici, M., Le Blanc, K., Mueller, I., Slaper-Cortenbach, I., Marini, F., Krause, D., Deans, R. et al. (2006). Minimal criteria for defining multipotent mesenchymal stromal cells. The International Society for Cellular Therapy position statement. *Cytotherapy*, volume 8 (4): 315–317.

Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome biology*, volume 3 (7): RESEARCH0036.

Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, volume 12 (August): 111–139.

Dufva, M. (2009). Introduction to microarray technology. *Methods in molecular biology (Clifton, N.J.)*, volume 529: 1–22.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A. and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, volume 21 (16): 3439–40.

Durinck, S., Spellman, P. T., Birney, E. and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, volume 4 (8): 1184–1191.

Dvinge, H. and Bertone, P. (2009). HTqPCR: high-throughput analysis and visualization of quantitative real-time PCR data in R. *Bioinformatics (Oxford, England)*, volume 25 (24): 3325–3326.

Eckel, J. E., Gennings, C., Chinchilli, V. M., Burgoon, L. D. and Zacharewski, T. R. (2004). Empirical bayes gene screening tool for time-course or dose-response microarray data. *Journal of biopharmaceutical statistics*, volume 14 (3): 647–670.

Edwards, D. (2003). Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics*, volume 19 (7): 825–833.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM - CBMS-NSF Regional Conference Series in Applied Mathematics.

Efron, B. (2003). Robbins, empirical Bayes and microarrays.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, volume 23 (1): 70–86.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment.

Ehrenreich, A. (2006). DNA microarray technology for the microbiologist: An overview.

Eisenberg, E. and Levanon, E. Y. (2003). Human housekeeping genes are compact.

Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011). *Cluster Analysis*. John Wiley & Sons.

Falcon, S. and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, volume 23 (2): 257–258.

Felker, P., Seré, K., Lin, Q., Becker, C., Hristov, M., Hieronymus, T. and Zenke, M. (2010). TGF-beta1 accelerates dendritic cell differentiation from common dendritic cell progenitors and directs subset specification toward conventional dendritic cells. *The Journal of Immunology*, volume 185 (9): 5326–5335.

Ferhat, L. (2012). Potential role of drebrin A, an F-actin binding protein, in reactive synaptic plasticity after pilocarpine-induced seizures: Functional implications in epilepsy.

Filkov, V. and Skiena, S. (2003). Integrating microarray data by consensus clustering. *Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence*.

Findlay, J. W. A. and Dillard, R. F. (2007). Appropriate calibration curve fitting in ligand binding assays. *The AAPS journal*, volume 9 (2): E260–E267.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P. et al. (2011). Ensembl 2011. *Nucleic Acids Research*, volume 39 (Database issue): D800–D806.

Fraley, C. and Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation.

Fraley, C. and Raftery, A. E. (2003). Enhanced Model-Based Clustering, Density Estimation, and Discriminant Analysis Software: MCLUST.

Frey, B. J. and Dueck, D. (2007). Clustering by Passing Messages Between Data Points. *Science*, volume 315 (5814): 972–976.

Froehlich, H., Fellmann, M., Sueltmann, H., Poustka, A. and Beissbarth, T. (2007). Large scale statistical inference of signaling pathways from RNAi and microarray data. *BMC bioinformatics*, volume 8 (1): 386.

Fröhlich, H., Speer, N., Poustka, A. and Beissbarth, T. (2007). GOSim–an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC bioinformatics*, volume 8: 166.

Gao, L., McBeath, R. and Chen, C. S. (2010). Stem cell shape regulates a chondrogenic versus myogenic fate through Rac1 and N-cadherin. *Stem Cells*, volume 28 (3): 564–572.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y. et al. (2012). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, volume 40 (D1).

Gaydos, B., Krams, M., Perevozskaya, I., Bretz, F., Liu, Q., Gallo, P., Berry, D. et al. (2006). Adaptive Dose-Response Studies. *Drug Information Journal*, volume 40: 451–461.

Geistlinger, L., Csaba, G., Küffner, R., Mulder, N. and Zimmer, R. (2011). From sets to graphs: Towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics*, volume 27 (13).

GeneExpression (2015). Gene Expression Omnibus (Repository).

Ghosh, A. K., Bhattacharyya, S., Lakos, G., Chen, S.-J., Mori, Y. and Varga, J. (2004). Disruption of transforming growth factor beta signaling and profibrotic responses in normal skin fibroblasts by peroxisome proliferator-activated receptor gamma. *Arthritis and rheumatism*, volume 50 (4): 1305–1318.

Giacomini, D., Páez-Pereda, M., Theodoropoulou, M., Labeur, M., Refojo, D., Gerez, J., Chervin, A. et al. (2006). Bone morphogenetic protein-4 inhibits corticotroph tumor cells: involvement in the retinoic acid inhibitory action. *Endocrinology*, volume 147 (1): 247–256.

Ginsburg, G. S. and Willard, H. F. (2012). *Genomic and Personalized Medicine, Volumes 1-2*. Academic Press.

Git, A., Dvinge, H., Salmon-Divon, M., Osborne, M., Kutter, C., Hadfield, J., Bertone, P. et al. (2010). Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA (New York, N.Y.)*, volume 16 (5): 991–1006.

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R. and Valencia, A. (2012). EnrichNet: network-based gene set enrichment analysis. *Bioinformatics (Oxford, England)*, volume 28 (18): i451–i457.

Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics*, volume 23 (8): 980–987.

Goldmann, T. and Gonzalez, J. S. (2000). DNA-printing: Utilization of a standard inkjet printer for the transfer of nucleic acids to solid supports. *Journal of Biochemical and Biophysical Methods*, volume 42 (3): 105–110.

Gollub, J. and Sherlock, G. (2006). Clustering microarray data. *Methods in enzymology*, volume 411: 194–213.

Gottlieb, A., Stein, G. Y., Ruppin, E. and Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, volume 7: 496.

Grant, C. E., Bailey, T. L. and Noble, W. S. (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, volume 27 (7): 1017–1018.

Greene, W. (2005). Censored Data and Truncated Distributions. *The Handbook of Econometrics*, pages 695–734.

Griffiths-Jones, S. (2004). The microRNA Registry. *Nucleic Acids Research*, volume 32 (Database issue): D109–D111.

Griffiths-Jones, S. (2010). miRBase: microRNA sequences and annotation. *Current protocols in bioinformatics editoral board Andreas D Baxevanis et al*, volume Chapter 12: Unit 12.9.1–10.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer Science & Business Media.

Gu, C. (2004). Model diagnostics for smoothing spline ANOVA mdoels. *The Canadian Journal of Statistics*, volume 32 (4): 347–358.

Gu, C. and Ma, P. (2005). Optimal smoothing in nonparametric mixed-effect models. *Annals of Statistics*, volume 33 (3): 1357–1379.

Gudrun, W., Abnaof, K., Joussen, S., Meurer, S., Smeets, H., Rath, B., Hoffmann, K. et al. (2013). TGF-beta1 Does Not Induce Senescence of Multipotent Mesenchymal Stromal Cells and Has Similar Effects in Early and Late Passages. *PLoS ONE*, volume 8 (10).

Guo, D., Arnspiger, S., Rensing, N. R. and Wong, M. (2012). Brief seizures cause dendritic injury. *Neurobiology of Disease*, volume 45 (1): 348–355.

Guo, X., Qi, H., Verfaillie, C. M. and Pan, W. (2003). Statistical significance analysis of longitudinal gene expression data. *Bioinformatics*, volume 19 (13): 1628–1635.

Guo, Z., Guilfoyle, R. A., Thiel, A. J., Wang, R. and Smith, L. M. (1994). Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. *Nucleic acids research*, volume 22 (24): 5456–5465.

Hamalainen, H. K., Tubman, J. C., Vikman, S., Kyrölä, T., Ylikoski, E., Warrington, J. A. and Lahesmaa, R. (2001). Identification and validation of endogenous reference genes for expression profiling of T helper cell differentiation by quantitative real-time RT-PCR. *Analytical biochemistry*, volume 299 (1): 63–70.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K. et al. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*, volume 32 (Database issue): D258–D261.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Heller, M. J. (2002). DNA microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, volume 4: 129–153.

Hentges, S. and Sarkar, D. K. (2001). Transforming growth factor-beta regulation of estradiol-induced prolactinomas. *Frontiers in Neuroendocrinology*, volume 22 (4): 340–363.

Heyer, L. J., Kruglyak, S. and Yooseph, S. (1999). Exploring expression data identification and analysis of coexpressed genes. *Genome Research*, volume 9 (11): 1106–1115.

Higuchi, R., Dollinger, G., Walsh, P. S. and Griffith, R. (1992). Simultaneous amplification and detection of specific DNA sequences. *Bio/technology (Nature Publishing Company)*, volume 10 (4): 413–417.

Higuchi, R., Fockler, C., Dollinger, G. and Watson, R. (1993). Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Bio/technology (Nature Publishing Company)*, volume 11 (9): 1026–1030.

Hinck, A. P. (2012). Structural studies of the TGF-βs and their receptors - insights into evolution of the TGF-β superfamily. *FEBS letters*, volume 586 (14): 1860–70.

Hochreiter, S., Clevert, D.-A. and Obermayer, K. (2006). A new summarization method for Affymetrix probe level data. *Bioinformatics*, volume 22 (8): 943–949.

Hopkins, A. L. and Groom, C. R. (2002). The druggable genome. *Nature reviews. Drug discovery*, volume 1 (9): 727–730.

Horwitz, E. M., Le Blanc, K., Dominici, M., Mueller, I., Slaper-Cortenbach, I., Marini, F. C., Deans, R. J. et al. (2005). Clarification of the nomenclature for MSC: The International Society for Cellular Therapy position statement. *Cytotherapy*, volume 7 (5): 393–395.

Hu, G. and Agarwal, P. (2009). Human disease-drug network based on genomic expression profiles. *PLoS ONE*, volume 4 (8).

Hua, Y. J., Tu, K., Tang, Z. Y., Li, Y. X. and Xiao, H. S. (2008). Comparison of normalization methods with microRNA microarray. *Genomics*, volume 92 (2): 122–128.

Huang, S., Yeo, A. A., Gelbert, L., Lin, X., Nisenbaum, L. and Bemis, K. G. (2004). At What Scale Should Microarray Data Be Analyzed? *American Journal of PharmacoGenomics*, volume 4 (2): 129–139.

Huang, X., Zhang, H., Yang, J., Wu, J., McMahon, J., Lin, Y., Cao, Z. et al. (2010). Pharmacological inhibition of the mammalian target of rapamycin pathway suppresses acquired epilepsy. *Neurobiology of Disease*, volume 40 (1): 193–199.

Huber, W., von Heydebreck, A., Sueltmann, H., Poustka, A. and Vingron, M. (2003). Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical applications in genetics and molecular biology*, volume 2: Article3.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics (Oxford, England)*, volume 18 Suppl 1: S96–S104.

Huggett, J., Dheda, K., Bustin, S. and Zumla, A. (2005). Real-time RT-PCR normalisation; strategies and considerations. *Genes and immunity*, volume 6 (4): 279–284.

Hulme, E. C. and Trevethick, M. A. (2010). Ligand binding assays at equilibrium: Validation and interpretation.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T. et al. (2012). InterPro in 2011: New developments in the family and domain prediction database. *Nucleic Acids Research*, volume 40 (D1).

Hurle, M. R., Yang, L., Xie, Q., Rajpal, D. K., Sanseau, P. and Agarwal, P. (2013). Computational drug repositioning: from data to therapeutics. *Clinical pharmacology and therapeutics*, volume 93 (4): 335–41.

Hutvágner, G. and Zamore, P. D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. *Science (New York, N.Y.)*, volume 297 (5589): 2056–2060.

Im, H.-I. and Kenny, P. J. (2012). MicroRNAs in neuronal function and dysfunction.

Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., Murino, L. et al. (2010). 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*, volume 107 (33): 14621–14626.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, volume 4 (2): 249–264.

Iskar, M., Campillos, M., Kuhn, M., Jensen, L. J., van Noort, V. and Bork, P. (2010). Drug-induced regulation of target expression. *PLoS Computational Biology*, volume 6 (9).

Ito, T., Sawada, R., Fujiwara, Y., Seyama, Y. and Tsuchiya, T. (2007). FGF-2 suppresses cellular senescence of human mesenchymal stem cells by down-regulation of TGF-beta2. *Biochemical and Biophysical Research Communications*, volume 359 (1): 108–114.

Iwata, H., Sawada, R., Mizutani, S. and Yamanishi, Y. (2015). Systematic drug repositioning for a wide range of diseases with integrative analyses of phenotypic and molecular data. *Journal of chemical information and modeling*, volume 55 (2): 446–59.

Jain, A. and Moreau, J. (1987). Bootstrap technique in cluster analysis.

James, G. M. and Sugar, C. A. (2003). Clustering for Sparsely Sampled Functional Data.

Jansen, R. C. (2003). Studying complex biological systems using multifactorial perturbation. *Nature Reviews Genetics*, volume 4 (2): 145–151.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T. et al. (2009). STRING 8 - A global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, volume 37 (SUPPL. 1).

Jimenez-Mateos, E. M., Bray, I., Sanz-Rodriguez, A., Engel, T., McKiernan, R. C., Mouri, G., Tanaka, K. et al. (2011). MiRNA expression profile after status epilepticus and hippocampal neuroprotection by targeting miR-132. *American Journal of Pathology*, volume 179 (5): 2519–2532.

Jimenez-Mateos, E. M., Engel, T., Merino-Serrais, P., McKiernan, R. C., Tanaka, K., Mouri, G., Sano, T. et al. (2012). Silencing microRNA-134 produces neuroprotective and prolonged seizure-suppressive effects.

Jimenez-Mateos, E. M. and Henshall, D. C. (2013). Epilepsy and microRNA.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, volume 32 (3): 241–254.

Johnson, W. E., Li, C. and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics Oxford England*, volume 8 (1): 118–127.

Jonnalagadda, S. and Srinivasan, R. (2008). Principal components analysis based methodology to identify differentially expressed genes in time-course microarray data. *BMC bioinformatics*, volume 9: 267.

Jung, S., Bullis, J. B., Lau, I. H., Jones, T. D., Warner, L. N. and Poolos, N. P. (2010). Downregulation of dendritic HCN channel gating in epilepsy is mediated by altered phosphorylation signaling. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, volume 30 (19): 6678–6688.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, volume 28 (1): 27–30.

Karp, R. M. (2010). Reducibility among combinatorial problems. *In 50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*, pages 219–241. Springer Berlin Heidelberg.

Kaufman, L. and Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. *In Finding Groups in Data*, pages 1–67. John Wiley & Sons, Inc.

Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J. and Shoichet, B. K. (2007). Relating protein pharmacology by ligand chemistry. *Nature biotechnology*, volume 25 (2): 197–206.

Keller, T. H., Pichota, A. and Yin, Z. (2006). A practical view of 'druggability'.

Kerr, M. K. and Churchill, G. A. (2001). Statistical design and the analysis of gene expression microarray data. *Genet Res*, volume 77 (2): 123–128.

Kerr, M. K., Martin, M. and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of computational biology : a journal of computational molecular cell biology*, volume 7 (6): 819–837.

Khatri, P. and Draghici, S. (2005). Ontological analysis of gene expression data: Current tools, limitations, and open problems.

Kim, J. K., Gabel, H. W., Kamath, R. S., Tewari, M., Pasquinelli, A., Rual, J.-F., Kennedy, S. et al. (2005). Functional genomic analysis of RNA interference in C. elegans. *Science (New York, N.Y.)*, volume 308 (5725): 1164–1167.

Knowles, B. B., Howe, C. C. and Aden, D. P. (1980). Human hepatocellular carcinoma cell lines secrete the major plasma proteins and hepatitis B surface antigen. *Science*, volume 209 (4455): 497–499.

Kobow, K. and Blümcke, I. (2011). The methylation hypothesis: Do epigenetic chromatin modifications play a role in epileptogenesis? *Epilepsia*, volume 52 (SUPPL. 4): 15–19.

Koch, C. M., Reck, K., Shao, K., Lin, Q., Joussen, S., Ziegler, P., Walenda, G. et al. (2012). Pluripotent stem cells escape from senescence-associated DNA methylation changes. *Genome Research*.

Kosik, K. S. (2006). The neuronal microRNA system. *Nature reviews. Neuroscience*, volume 7 (12): 911–920.

Kozomara, A. and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, volume 39 (Database issue): D152–7.

Kravchik, M., Sunkar, R., Damodharan, S., Stav, R., Zohar, M., Isaacson, T. and Arazi, T. (2014). Global and local perturbation of the tomato microRNA pathway by a trans -activated DICER-LIKE 1 mutant. *Journal of Experimental Botany*, volume 65 (2): 725–739.

Kretschmann, A., Danis, B., Andonovic, L., Abnaof, K., van Rikxoort, M., Siegel, F., Mazzuferi, M. et al. (2015a). Different microRNA profiles in chronic epilepsy versus acute seizure mouse models. *Journal of molecular neuroscience : MN*, volume 55 (2): 466–79.

Kretschmann, A., Danis, B., Andonovic, L., Abnaof, K., van Rikxoort, M., Siegel, F., Mazzuferi, M. et al. (2015b). Epilepsy Acute 6-Hertz Mouse Model Data Set.

Kretschmann, A., Danis, B., Andonovic, L., Abnaof, K., van Rikxoort, M., Siegel, F., Mazzuferi, M. et al. (2015c). Epilepsy Chronic Pilocarpine Mouse Model Data Set.

Kretschmann, A., Danis, B., Andonovic, L., Abnaof, K., van Rikxoort, M., Siegel, F., Mazzuferi, M. et al. (2015d). Epilepsy Chronic SSSE Mouse Model Data Set.

Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J. and Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, volume 6: 343.

Kuo, F. C., Lee, M.-L. T., Whitmore, G. A. and Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences of the United States of America*, volume 97 (18): 9834–9839.

Lachgar, A., Jaureguiberry, G., Le Buenac, H., Bizzini, B., Zagury, J. F., Rappaport, J. and Zagury, D. (1998). Binding of HIV-1 to RBCs involves the Duffy Antigen Receptors for Chemokines (DARC). *Biomedicine and Pharmacotherapy*, volume 52 (10): 436–439.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J. et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science (New York, N.Y.)*, volume 313 (5795): 1929–1935.

Landry, Y. and Gies, J. P. (2008). Drugs and their molecular targets: An updated overview.

Lanphier, E., Urnov, F., Haecker, S. E., Werner, M. and Smolenski, J. (2015). Don't edit the human germ line. *Nature*, volume 519 (7544): 410–1.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., MacIejewski, A. et al. (2014). DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Research*, volume 42 (D1).

Lee, M.-L. T., Lu, W., Whitmore, G. A. and Beier, D. (2002). Models for microarray gene expression data. *Journal of biopharmaceutical statistics*, volume 12 (1): 1–19.

Lee, S., Lee, K. H., Song, M. and Lee, D. (2011). Building the process-drug-side effect network to discover the relationship between biological processes and side effects. *BMC bioinformatics*, volume 12 Suppl 2: S2.

Leek, J. T., Monsen, E., Dabney, A. R. and Storey, J. D. (2006). EDGE: extraction and analysis of differential gene expression. *Bioinformatics (Oxford, England)*, volume 22 (4): 507–508.

Leha, A., Beissbarth, T. and Jung, K. (2011). Sequential interim analyses of survival data in DNA microarray experiments. *BMC bioinformatics*, volume 12: 127.

Levine, E. and Domany, E. (2001). Resampling method for unsupervised estimation of cluster validity. *Neural computation*, volume 13 (11): 2573–2593.

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

Li, K.-X., Lu, Y.-M., Xu, Z.-H., Zhang, J., Zhu, J.-M., Zhang, J.-M., Cao, S.-X. et al. (2011). Neuregulin 1 regulates excitability of fast-spiking neurons through Kv1.1 and acts in epilepsy.

Li, Q. and Lai, L. (2007). Prediction of potential drug targets based on simple sequence properties. *BMC bioinformatics*, volume 8: 353.

Liang, M., Briggs, A. G., Rute, E., Greene, A. S. and Cowley, A. W. (2003). Quantitative assessment of the importance of dye switching and biological replication in cDNA microarray studies. *Physiological genomics*, volume 14 (3): 199–207.

Liang, P., Xu, Y., Zhang, X., Ding, C., Huang, R., Zhang, Z., Lv, J. et al. (2015). CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes. *Protein & cell*, volume 6 (5): 363–72.

Liberali, P., Snijder, B. and Pelkmans, L. (2014). Single-cell and multivariate approaches in genetic perturbation screens. *Nature Reviews Genetics*, volume 16 (1): 18–32.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. *In Proceedings of ICML*, pages 296–304.

Lin, H., Sun, L. and Crooks, R. M. (2005). Replication of a DNA microarray. *Journal of the American Chemical Society*, volume 127 (32): 11210–1.

Liu, R.-M. and Gaston Pravia, K. A. (2010). Oxidative stress and glutathione in TGF-beta-mediated fibrogenesis. *Free radical biology & medicine*, volume 48 (1): 1–15.

Livak, K. J. and Schmittgen, T. D. (2001). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2-DDCT Method. *Methods*, volume 25 (4): 402–408.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M. et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology*, volume 14 (13): 1675–1680.

Lohmann, M., Walenda, G., Hemeda, H., Joussen, S., Drescher, W., Jockenhoevel, S., Hutschenreuter, G. et al. (2012). Donor age of human platelet lysate affects proliferation and differentiation of mesenchymal stem cells. *PloS one*, volume 7 (5): e37839.

Loinger, A., Shemla, Y., Simon, I., Margalit, H. and Biham, O. (2012). Competition between small RNAs: A quantitative view. *Biophysical Journal*, volume 102 (8): 1712–1721.

Lonnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statistical Sinica*, volume 12 (12): 31–46.

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, volume 19 (4): 474–482.

Luan, Y. and Li, H. (2004). Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics (Oxford, England)*, volume 20 (3): 332–339.

Luehr, S., Hartmann, H. and Söding, J. (2012). The XXmotif web server for eXhaustive, weight matriX-based motif discovery in nucleotide sequences. *Nucleic acids research*, volume 40 (Web Server issue): W104–9.

M. Johnson, G. M. (1990). *Concepts and Applcations of Molecular Similarity*. Wiley, New York.

Ma, P., Castillo-Davis, C. I., Zhong, W. and Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research*, volume 34 (4): 1261–1269.

Macarron, R. (2006). Critical review of the role of HTS in drug discovery. *Drug Discovery Today*, volume 11 (7-8): 277–279.

Maeda, H., Fujimoto, C., Haruki, Y., Maeda, T., Kokeguchi, S., Petelin, M., Arai, H. et al. (2003). Quantitative real-time PCR using TaqMan and SYBR Green for Actinobacillus actinomycetemcomitans, Porphyromonas gingivalis, Prevotella intermedia, tetQ gene and total bacteria. *FEMS Immunology and Medical Microbiology*, volume 39 (1): 81–86.

Mahony, S. and Benos, P. V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Research*, volume 35 (Web Server issue): W253–W258.

Mar, J. C., Kimura, Y., Schroder, K., Irvine, K. M., Hayashizaki, Y., Suzuki, H., Hume, D. et al. (2009). Data-driven normalization strategies for high-throughput quantitative RT-PCR. *BMC bioinformatics*, volume 10 (1): 110.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R. et al. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, volume 9 (8): 796–804.

Margis, R., Margis, R. and Rieder, C. R. M. (2011). Identification of blood microRNAs associated to Parkinsonós disease. *Journal of Biotechnology*, volume 152 (3): 96–101.

Martinez Molina, D., Jafari, R., Ignatushchenko, M., Seki, T., Larsson, E. A., Dan, C., Sreekumar, L. et al. (2013). Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay. *Science (New York, N.Y.)*, volume 341 (6141): 84–7.

Maskos, U. and Southern, E. M. (1992). Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. *Nucleic acids research*, volume 20 (7): 1679–1684.

Massagué, J. (1990). The transforming growth factor-beta family. *Annula Review of Cell Biology*, volume 6: 597–641.

Massagué, J. (1998). TGF-β signal transduction. *Annual Review of Biochemistry*, volume 67 (1): 753–791.

Massanet-Vila, R., Albert, F. F., Caminal, P. and Perera, A. (2012). Network-based enrichment analysis of gene expression through protein-protein interaction data. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, volume 2012: 6317–20.

Mayhoub, A. S., Marler, L., Kondratyuk, T. P., Park, E. J., Pezzuto, J. M. and Cushman, M. (2012). Optimization of the aromatase inhibitory activities of pyridylthiazole analogues of resveratrol. *Bioorganic and Medicinal Chemistry*, volume 20 (7): 2427–2434.

Mazzuferi, M., Kumar, G., Rospo, C. and Kaminski, R. M. (2012). Rapid epileptogenesis in the mouse pilocarpine model: Video-EEG, pharmacokinetic and histopathological characterization. *Experimental Neurology*, volume 238 (2): 156–167.

Mazzuferi, M., Kumar, G., Van Eyll, J., Danis, B., Foerch, P. and Kaminski, R. M. (2013). Nrf2 defense pathway: Experimental evidence for its protective role in epilepsy. *Annals of Neurology*, volume 74 (4): 560–568.

McClelland, S., Flynn, C., Dubé, C., Richichi, C., Zha, Q., Ghestem, A., Esclapez, M. et al. (2011). Neuron-restrictive silencer factor-mediated hyperpolarization-activated cyclic nucleotide gated channelopathy in experimental temporal lobe epilepsy. *Annals of Neurology*, volume 70 (3): 454–464.

McGoldrick, A., Bensaude, E., Ibata, G., Sharp, G. and Paton, D. J. (1999). Closed one-tube reverse transcription nested polymerase chain reaction for the detection of pestiviral RNA with fluorescent probes. *Journal of Virological Methods*, volume 79 (1): 85–95.

McKiernan, R. C., Jimenez-Mateos, E. M., Sano, T., Bray, I., Stallings, R. L., Simon, R. P. and Henshall, D. C. (2012). Expression profiling the microRNA response to epileptic preconditioning identifies miR-184 as a modulator of seizure-induced neuronal death. *Experimental Neurology*, volume 237 (2): 346–354.

McNeill, E. and Van Vactor, D. (2012). MicroRNAs Shape the Neuronal Landscape.

Mehio, W., Kemp, G. J. L., Taylor, P. and Walkinshaw, M. D. (2010). Identification of protein binding surfaces using surface triplet propensities. *Bioinformatics (Oxford, England)*, volume 26 (20): 2549–2555.

Mellado, M., Rodríguez-Frade, J. M., Mañes, S. and Martínez-A, C. (2001). Chemokine signaling and functional responses: the role of receptor dimerization and TK pathway activation. *Annual review of immunology*, volume 19: 397–421.

Mestdagh, P., Van Vlierberghe, P., De Weer, A., Muth, D., Westermann, F., Speleman, F. and Vandesompele, J. (2009). A novel and universal method for microRNA RT-qPCR data normalization. *Genome biology*, volume 10 (6): R64.

Meurer SK, Walenda G, Abnaof K, Joussen S, Lin Q, Zenke M, Hoffmann K, Wagner W, Fröhlich H, W. R. (2013a). TGF-beta1-Induced Gene Expression Changes in HepG2 Cells (Dataset).

Meurer SK, Walenda G, Abnaof K, Joussen S, Lin Q, Zenke M, Hoffmann K, Wagner W, Fröhlich H, W. R. (2013b). TGF-beta1-Induced Gene Expression Changes in Primary Murine Hepatocytes (Dataset).

Meurer SK, Walenda G, Abnaof K, Joussen S, Lin Q, Zenke M, Hoffmann K, Wagner W, Fröhlich H, W. R. (2013c). TGF-beta1 Stimulation in Human Mesenchymal Stromal Cells, Gene Expression (Dataset).

Meurer SK, Walenda G, Abnaof K, Joussen S, Lin Q, Zenke M, Hoffmann K, Wagner W, Fröhlich H, W. R. (2013d). TGF-beta1 stimulation in Mouse Hematopoietic Progenitor Cells (Multipotent Progenitor & Common Dendritic Progenitor) - Time-Course (Dataset).

Miller, R. G. (1974). The Jackknife–A Review. *Biometrika*, volume 61 (1): 1–15.

Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., Peterson, A. et al. (2008). Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences of the United States of America*, volume 105 (30): 10513–10518.

Montaner, D. and Dopazo, J. (2010). Multidimensional gene set analysis of genomic data. *PLoS ONE*, volume 5 (4): e10348.

Monti, S., Tamayo, P., Mesirov, J. and Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, volume 52 (1-2): 91–118.

Moore, a. (2001). K-means and Hierarchical Clustering. *Statistical Data Mining Tutorials*, pages 1–24.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Biswas, M. et al. (2002). InterPro: an integrated documentation resource for protein families, domains and functional sites. *Briefings in bioinformatics*, volume 3 (3): 225–235.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P. et al. (2007). New developments in the InterPro database. *Nucleic Acids Research*, volume 35 (SUPPL. 1).

Munson, P. J. and Rodbard, D. (1988). An exact correction to the "Cheng-Prusoff" correction. *Journal of receptor research*, volume 8 (1-4): 533–546.

Naeger, D. M., Kohi, M. P., Webb, E. M., Phelps, A., Ordovas, K. G. and Newman, T. B. (2013). Correctly using sensitivity, specificity, and predictive values in clinical practice: How to avoid three common pitfalls. *American Journal of Roentgenology*, volume 200 (6): 566–570.

Nayal, M. and Honig, B. (2006). On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. *Proteins: Structure, Function and Genetics*, volume 63 (4): 892–906.

Nelander, S., Wang, W., Nilsson, B., She, Q.-B., Pratilas, C., Rosen, N., Gennemark, P. et al. (2008). Models from experiments: combinatorial drug perturbations of cancer cells. *Molecular systems biology*, volume 4: 216.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R. and Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of computational biology : a journal of computational molecular cell biology*, volume 8 (1): 37–52.

Nguyen, N. and Caruana, R. (2007). Consensus clusterings. *In Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 607–612.

Nomura, N., Sasamoto, S., Ishii, S., Date, T., Matsui, M. and Ishizaki, R. (1989). Isolation of human cDNA clones of ski and the ski-related gene, sno. *Nucleic Acids Research*, volume 17 (14): 5489–5500.

O'Connell, M. R., Oakes, B. L., Sternberg, S. H., East-Seletsky, A., Kaplan, M. and Doudna, J. A. (2014). Programmable RNA recognition and cleavage by CRISPR/Cas9. *Nature*, volume 516 (7530): 263–266.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes.

Oliver, B. and Leblanc, B. (2003). How many genes in a genome? *Genome biology*, volume 5 (1): 204.

Oomizu, S., Chaturvedi, K. and Sarkar, D. K. (2004). Folliculostellate cells determine the susceptibility of lactotropes to estradiol's mitogenic action. *Endocrinology*, volume 145 (3): 1473–1480.

Orlova, V. V., Liu, Z., Goumans, M.-J. and Ten Dijke, P. (2011). Controlling angiogenesis by two unique TGF-β type I receptor signaling pathways. *Histology and histopathology*, volume 26 (9): 1219–1230.

Orth, A. P., Batalov, S., Perrone, M. and Chanda, S. K. (2004). The promise of genomics to identify novel therapeutic targets. *Expert opinion on therapeutic targets*, volume 8 (6): 587–596.

Overington, J. P., Al-Lazikani, B. and Hopkins, A. L. (2006). How many drug targets are there? *Nature reviews. Drug discovery*, volume 5 (12): 993–996.

Pabinger, S., Rödiger, S., Kriegner, A., Vierlinger, K. and Weinhäusel, A. (2014). A survey of tools for the analysis of quantitative PCR (qPCR) data. *Biomolecular Detection and Quantification*, volume 1 (1): 23–33.

Pan, K.-H., Lih, C.-J. and Cohen, S. N. (2005). Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, volume 102 (25): 8961–8965.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, volume 18 (4): 546–554.

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G. and Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, volume 56 (1): 45–50.

Park, T., Yi, S. G., Lee, S., Lee, S. Y., Yoo, D. H., Ahn, J. I. and Lee, Y. S. (2003). Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics*, volume 19 (6): 694–703.

Parmigiani, G., Garett, E. S., Irizarry, R. A. and Zeger, S. L. (2003). *The Analysis of Gene Expression Data: Methods and Software*. Springer Science & Business Media.

Pearson, K. (1897). Mathematical Contributions to the Theory of Evolution.–On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs on JSTOR. *Proceedings of the Royal Society of London*, volume Vol. 60 (1896 - 1897): 489–498.

Pease, J. E. and Horuk, R. (2014). Recent progress in the development of antagonists to the chemokine receptors CCR3 and CCR4. *Expert opinion on drug discovery*, volume 9 (5): 467–83.

Peltier, H. J. and Latham, G. J. (2008). Normalization of microRNA expression levels in quantitative RT-PCR assays: identification of suitable reference RNA targets in normal and cancerous human solid tissues. *RNA (New York, N.Y.)*, volume 14 (5): 844–852.

Pesquita, C., Faria, D., Falcão, A. O., Lord, P. and Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS computational biology*, volume 5 (7): e1000443.

Petrov, V. V., Fagard, R. H. and Lijnen, P. J. (2002). Stimulation of collagen production by transforming growth factor-beta1 during differentiation of cardiac fibroblasts to myofibroblasts. *Hypertension*, volume 39 (2): 258–63.

Pfaffl, M. W., Tichopad, A., Prgomet, C. and Neuvians, T. P. (2004). Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–Excel-based tool using pair-wise correlations. *Biotechnology letters*, volume 26 (6): 509–515.

Phillips, P. C. B. (1986). The Exact Distribution of the Wald Statistic. *Econometrica*, volume 54 (4): 881–895.

Pillai, R. S. (2005). MicroRNA function: multiple mechanisms for a tiny RNA? *RNA (New York, N.Y.)*, volume 11 (12): 1753–1761.

Poirel, C. L., Owens, C. C. and Murali, T. M. (2011). Network-based functional enrichment.

Ponchel, F., Toomes, C., Bransfield, K., Leong, F. T., Douglas, S. H., Field, S. L., Bell, S. M. et al. (2003). Real-time PCR based on SYBR-Green I fluorescence: an alternative to the TaqMan assay for a relative quantification of gene rearrangements, gene amplifications and micro gene deletions. *BMC biotechnology*, volume 3 (1): 18.

Post, S., Abdallah, B. M., Bentzon, J. F. and Kassem, M. (2008). Latent TGF-beta binding proteins (LTBPs)-1 and -3 coordinate proliferation and osteogenic differentiation of human mesenchymal stem cells. *Bone*, volume 43 (1): 679–688.

Pozhitkov, A., Noble, P. A., Domazet-Lošo, T., Nolte, A. W., Sonnenberg, R., Staehler, P., Beier, M. et al. (2006). Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted. *Nucleic Acids Research*, volume 34 (9).

Pradervand, S., Weber, J., Thomas, J., Bueno, M., Wirapati, P., Lefort, K., Dotto, G. P. et al. (2009). Impact of normalization on miRNA microarray expression profiling. *RNA (New York, N.Y.)*, volume 15 (3): 493–501.

Praveen, P. and Fröhlich, H. (2013). Boosting Probabilistic Graphical Model Inference by Incorporating Prior Knowledge from Multiple Sources. *PLoS ONE*, volume 8 (6).

Pritchard, C. C., Kroh, E., Wood, B., Arroyo, J. D., Dougherty, K. J., Miyaji, M. M., Tait, J. F. et al. (2012). Blood cell origin of circulating microRNAs: A cautionary note for cancer biomarker studies. *Cancer Prevention Research*, volume 5 (3): 492–497.

Qabaja, A., Alshalalfa, M., Alanazi, E. and Alhajj, R. (2014). Prediction of novel drug indications using network driven biological data prioritization and integration. *Journal of cheminformatics*, volume 6 (1): 1.

Quackenbush, J. (2001). Computational analysis of microarray data. *Nature reviews. Genetics*, volume 2 (6): 418–427.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature genetics*, volume 32 Suppl: 496–501.

Rao, Y., Lee, Y., Jarjoura, D., Ruppert, A. S., Liu, C.-G., Hsu, J. C. and Hagan, J. P. (2008). A comparison of normalization techniques for microRNA microarray data. *Statistical Applications in Genetics and Molecular Biology*, volume 7 (1): Article22.

Renner, U., Paez-Pereda, M., Arzt, E. and Stalla, G. K. (2004). Growth factors and cytokines: function and molecular regulation in pituitary adenomas. *Frontiers of Hormone Research*, volume 32: 96–109.

Rishton, G. M. (1997). Reactive compounds and in vitro false positives in HTS.

Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A. and Smyth, G. K. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, volume 23 (20): 2700–2707.

Robbins, H. (1956). An Empirical Bayes Approach to Statistics. *In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, x, pages 157–163. University of California Press.

Roberts AB, S. M. (1993). Physiological actions and clinical applications of transforming growth factor-beta (TGF-beta). *Growth Factors*, volume 8 (1): 1–9.

Roberts AB, Heine UI, Flanders KC, S. M. (1990). Transforming growth factor-beta. Major role in regulation of extracellular matrix. *Annals of the New York Academy of Sciences*, volume 580: 225:32.

Rocke, D. M. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of computational biology : a journal of computational molecular cell biology*, volume 8 (6): 557–569.

Roider, H. G., Pavlova, N., Kirov, I., Slavov, S., Slavov, T., Uzunov, Z. and Weiss, B. (2014). Drug2Gene: an exhaustive resource to explore effectively the drug-target relation network. *BMC bioinformatics*, volume 15 (1): 68.

Ross, S. A. and Davis, C. D. (2011). MicroRNA, nutrition, and cancer prevention. *Advances in nutrition (Bethesda, Md.)*, volume 2 (6): 472–85.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, volume 20: 53–65.

Ruppert, D., Wand, M. and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press, Cambridge.

Russ, A. P. and Lampel, S. (2005). The druggable genome: An update.

Sartor, M. A., Leikauf, G. D. and Medvedovic, M. (2009). LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, volume 25 (2): 211–217.

Sartor, M. A., Mahavisno, V., Keshamouni, V. G., Cavalcoli, J., Wright, Z., Karnovsky, A., Kuick, R. et al. (2010). ConceptGen: a gene set enrichment and gene set relation mapping tool. *Bioinformatics*, volume 26 (4): 456–463.

Schaefer, A., O'Carroll, D., Tan, C. L., Hillman, D., Sugimori, M., Llinas, R. and Greengard, P. (2007). Cerebellar neurodegeneration in the absence of microRNAs. *The Journal of experimental medicine*, volume 204 (7): 1553–8.

Scharl, T. and Leisch, F. (2006). Jackknife distances for clustering time–course gene expression data. *In 2006 JSM Proceedings*, pages 346–353. American Statistical Association, Alexandria, USA.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray.

Schmittgen, T. D., Lee, E. J. and Jiang, J. (2008). High-throughput real-time PCR. *Methods in molecular biology (Clifton, N.J.)*, volume 429: 89–98.

Schmittgen, T. D. and Livak, K. J. (2008). Analyzing real-time PCR data by the comparative C(T) method. *Nature protocols*, volume 3 (6): 1101–1108.

Schmittgen, T. D. and Zakrajsek, B. A. (2000). Effect of experimental treatment on housekeeping gene expression: validation by real-time, quantitative RT-PCR. *Journal of biochemical and biophysical methods*, volume 46 (1-2): 69–81.

Schneider, H.-C. and Klabunde, T. (2013). Understanding drugs and diseases by systems biology? *Bioorganic & medicinal chemistry letters*, volume 23 (5): 1168–76.

Schöler, N., Langer, C. and Kuchenbauer, F. (2011). Circulating microRNAs as biomarkers - True Blood?

Schützenmeister, A. and Piepho, H.-P. (2010). Background correction of two-colour cDNA microarray data using spatial smoothing methods. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, volume 120 (2): 475–490.

Seglen, P. (1976). Preparation of isolated rat liver cells. *Methods Cell Biology*, volume 13: 29–83.

Senbabaoglu, Y., Michailidis, G. and Li, J. Z. (2014a). A reassessment of consensus clustering for class discovery. Technical report, Department of Computational Medicine & Bioinformatics, University of Michiga.

Senbabaoglu, Y., Michailidis, G. and Li, J. Z. (2014b). Critical limitations of consensus clustering in class discovery. *Scientific reports*, volume 4: 6207.

Seré, K., Baek, J.-H., Ober-Blöbaum, J., Müller-Newen, G., Tacke, F., Yokota, Y., Zenke, M. et al. (2012). Two distinct types of Langerhans cells populate the skin during steady state and inflammation. *Immunity*, volume 37 (5): 905–16.

Shalem, O., Sanjana, N. E. and Zhang, F. (2015). High-throughput functional genomics using CRISPR-Cas9. *Nature reviews. Genetics*, volume 16 (5): 299–311.

Shannon, W., Culverhouse, R. and Duncan, J. (2003). Analyzing microarray data using cluster analysis. *Pharmacogenomics*, volume 4 (1): 41–52.

Sharma, M. B., Limaye, L. S. and Kale, V. P. (2012). Mimicking the functional hematopoietic stem cell niche in vitro: Recapitulation of marrow physiology by hydrogel-based three-dimensional cultures of mesenchymal stromal cells. *Haematologica*, volume 97 (5): 651–660.

Shoichet, B. K. (2006). Interpreting steep dose-response curves in early inhibitor discovery. *Journal of Medicinal Chemistry*, volume 49 (25): 7274–7277.

Shojaie, A. and Michailidis, G. (2010). Network enrichment analysis in complex experiments. *Statistical applications in genetics and molecular biology*, volume 9 (1): Article22.

Silver, J. D., Ritchie, M. E. and Smyth, G. K. (2009). Microarray background correction: maximum likelihood estimation for the normal–exponential convolution. *Biostatistics Oxford England*, volume 10 (2): 352–363.

Silver, N., Best, S., Jiang, J. and Thein, S. L. (2006). Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC molecular biology*, volume 7: 33.

Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., Sweet-Cordero, A., Sage, J. et al. (2011). Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science translational medicine*, volume 3 (96): 96ra77.

Skehan, P., Storeng, R., Scudiero, D., Monks, A., McMahon, J., Vistica, D., Warren, J. T. et al. (1990). New colorimetric cytotoxicity assay for anticancer-drug screening. *Journal of the National Cancer Institute*, volume 82 (13): 1107–1112.

Smyth, G. K. (2004a). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, volume 3 (1): Article3.

Smyth, G. K. (2004b). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, volume 3: Article3.

Smyth, G. K. (2005). Limma : Linear Models for Microarray Data. *Bioinformatics*, volume pages (2005): 397–420.

Smyth, G. K. and Speed, T. (2003). Normalization of cDNA microarray data. *Methods San Diego Calif*, volume 31 (4): 265–273.

Son, Y. S. and Baek, J. (2008). A modified correlation coefficient based similarity measure for clustering time-course gene expression data. *Pattern Recognition Letters*, volume 29 (3): 232–242.

Song, J. J., Lee, H. J., Morris, J. S. and Kang, S. H. (2007). Clustering of time-course gene expression data using functional data analysis. *Computational Biology and Chemistry*, volume 31 (4): 265–274.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic acids research*, volume 34 (Database issue): D535–D539.

Stingele, S., Stoehr, G., Peplowska, K., Cox, J., Mann, M. and Storchova, Z. (2012). Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells.

Storey, J. and Tibshirani, R. (2003a). SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays. *In The Analysis of Gene Expression Data SE - 12*, pages 272–290. Springer Science & Business Media.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, volume 64 (3): 479–498.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, volume 31 (6): 2013–2035.

Storey, J. D. (2010). False Discovery Rates. *Princeton University, Princeton, USA*, volume 1 (January): 1–7.

Storey, J. D., Dai, J. Y. and Leek, J. T. (2007). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics*, volume 8 (2): 414–432.

Storey, J. D. and Tibshirani, R. (2003b). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, volume 100 (16): 9440–9445.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G. and Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, volume 102 (36): 12837–12842.

Strehl, A. and Ghosh, J. (2003). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, volume 3: 583–617.

Strobl, H., Bello-Fernandez, C., Riedl, E., Pickl, W. F., Majdic, O., Lyman, S. D. and Knapp, W. (1997). flt3 ligand in cooperation with transforming growth factor-beta1 potentiates in vitro development of Langerhans-type dendritic cells and allows single-cell dendritic cell cluster formation under serum-free conditions. *Blood*, volume 90 (4): 1425–1434.

Strobl, H., Riedl, E., Scheinecker, C., Bello-Fernandez, C., Pickl, W. F., Rappersberger, K., Majdic, O. et al. (1996). TGF-beta 1 promotes in vitro development of dendritic cells from CD34+ hemopoietic progenitors. *The Journal of Immunology*, volume 157 (4): 1499–1507.

Stuhle, L. and Wold, S. (1989). Analysis of variance (ANOVA).

Sturn, A., Quackenbush, J. and Trajanoski, Z. (2002). Genesis: cluster analysis of microarray data. *Bioinformatics (Oxford, England)*, volume 18 (1): 207–8.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. a., Paulovich, A. et al. (2005). 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*, volume 102 (43): 15545–50.

Sühnel, J. (1998). Parallel dose-response curves in combination experiments. *Bulletin of mathematical biology*, volume 60 (2): 197–213.

Szanto, A., Narkar, V., Shen, Q., Uray, I. P., Davies, P. J. A. and Nagy, L. (2004). Retinoid X receptors: X-ploring their (patho)physiological functions. *Cell death and differentiation*, volume 11 Suppl 2 (S2): S126–43.

Tai, Y. C. and Speed, T. P. (2006). A multivariate empirical Bayes statistic for replicated microarray time course data. *Annals of Statistics*, volume 34 (5): 2387–2412.

Tamayo, P., Steinhardt, G., Liberzon, A. and Mesirov, J. P. (2012). The limitations of simple gene set enrichment analysis assuming gene independence.

Tibshirani, R. and Walther, G. (2005). Cluster Validation by Prediction Strength.

Tilstone, C. (2003). DNA Microarrays: Vital Statistics. *Nature*, volume 424: 610–612.

Topchy, a. and Jain, A. (2004). A mixture model for clustering ensembles. *Proc. SIAM Int. Conf. Data Mining*, pages 379–390.

Topchy, A., Jain, A. and Punch, W. (2003). Combining multiple weak clusterings. *Third IEEE International Conference on Data Mining*.

Turnbull, B. W. (1975). The Empirical Distribution Function with Arbitrarily Grouped , Censored and Truncated Data point in time is to be incorporated. *Journal of the Royal Statistical Society*, volume 38 (3): 290–295.

Turner, N. and Grose, R. (2010). Fibroblast growth factor signalling: from development to cancer. *Nature reviews. Cancer*, volume 10 (2): 116–29.

Turski, W. A., Cavalheiro, E. A., Schwarz, M., Czuczwar, S. J., Kleinrok, Z. and Turski, L. (1983). Limbic seizures produced by pilocarpine in rats: behavioural, electroencephalographic and neuropathological study. *Behavioural brain research*, volume 9 (3): 315–335.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, volume 98 (9): 5116–5121.

Uehara, T., Ono, A., Maruyama, T., Kato, I., Yamada, H., Ohno, Y. and Urushidani, T. (2010). The Japanese toxicogenomics project: Application of toxicogenomics. *Molecular Nutrition and Food Research*, volume 54 (2): 218–227.

Ultsch, A. (2003). *Is Log Radio a Good Value for Identifying Differential Expressed Genes in Microarray Experiments?* Fachbereich Mathematik und Informatik.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A. and Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome biology*, volume 3 (7): RESEARCH0034.

VanGuilder, H. D., Vrana, K. E. and Freeman, W. M. (2008). Twenty-five years of quantitative PCR for gene expression analysis.

Vickers, K. C., Palmisano, B. T., Shoucri, B. M., Shamburek, R. D. and Remaley, A. T. (2011). MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins. *Nature Cell Biology*, volume 13 (4): 423–433.

Volvert, M.-L., Rogister, F., Moonen, G., Malgrange, B. and Nguyen, L. (2012). MicroRNAs tune cerebral cortical neurogenesis.

Waage, P. and Gulberg, C. M. (1986). Studies concerning affinity. *J. Chem. Educ.*, volume 63 (12): 1044.

Wagner, W. and Ho, A. D. (2007). Mesenchymal stem cell preparations–comparing apples and oranges. *Stem cell reviews*, volume 3 (4): 239–48.

Walenda, G., Hemeda, H., Schneider, R. K., Merkel, R., Hoffmann, B. and Wagner, W. (2012). Human Platelet Lysate Gel Provides a Novel 3D-Matrix for Enhanced Culture Expansion of Mesenchymal Stromal Cells. *Tissue engineering Part C Methods*, volume 1 (2): 1–32.

Wang, D., Park, J. S., Chu, J. S. F., Krakowski, A., Luo, K., Chen, D. J. and Li, S. (2002). Proteomic profiling of bone marrow mesenchymal stem cells upon transforming growth factor beta1 stimulation. *The Journal of Biological Chemistry*, volume 277 (42): 36045–36051.

Wang, J., Tan, L., Tan, L., Tian, Y., Ma, J., Tan, C.-C., Wang, H.-F. et al. (2015). Circulating microRNAs are promising novel biomarkers for drug-resistant epilepsy. *Scientific reports*, volume 5: 10201.

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. and Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics*, volume 23 (10): 1274–1281.

Wang, K., Sun, J., Zhou, S., Wan, C., Qin, S., Li, C., He, L. et al. (2013a). Prediction of Drug-Target Interactions for Drug Repositioning Only Based on Genomic Expression Similarity. *PLoS Computational Biology*, volume 9 (11).

Wang, Y., Chen, S., Deng, N. and Wang, Y. (2013b). Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PLoS ONE*, volume 8 (11).

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, volume 58 (301): 236–244.

Weiss, S. R., Post, R. M., Gold, P. W., Chrousos, G., Sullivan, T. L., Walker, D. and Pert, A. (1986). CRF-induced seizures and behavior: interaction with amygdala kindling. *Brain research*, volume 372 (2): 345–351.

Williamson, M. P. (2013). Using chemical shift perturbation to characterise ligand binding.

Wilson, D. L., Buckley, M. J., Helliwell, C. A. and Wilson, I. W. (2003). New normalization methods for cDNA microarray data. *Bioinformatics*, volume 19 (11): 1325–1332.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. et al. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of computational biology : a journal of computational molecular cell biology*, volume 8 (6): 625–637.

Workman, C., Jensen, L. J., Jarmer, H., Berka, R., Gautier, L., Nielser, H. B. r., Saxild, H.-H. et al. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome biology*, volume 3 (9): research0048.

Wu, H., Kerr, M. K., Cui, X. and Churchill, G. a. (2003). MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments. *In Parmigiani G Garrett E S Irizarry R A and Zeger S L*, pages 313–341.

Wu, Y. (2010). Chemokine control of HIV-1 infection: beyond a binding competition. *Retrovirology*, volume 7: 86.

Xia, Z., Wu, L.-Y., Zhou, X. and Wong, S. T. C. (2010). Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC systems biology*, volume 4 Suppl 2: S6.

Xu, L. L., Shanmugam, N., Segawa, T., Sesterhenn, I. A., McLeod, D. G., Moul, J. W. and Srivastava, S. (2000). A novel androgen-regulated gene, PMEPA1, located on chromosome 20q13 exhibits high level expression in prostate. *Genomics*, volume 66 (3): 257–263.

Xu, L. L., Shi, Y., Petrovics, G., Sun, C., Makarem, M., Zhang, W., Sesterhenn, I. A. et al. (2003). PMEPA1, an androgen-regulated NEDD4-binding protein, exhibits cell growth inhibitory function and decreased expression during prostate cancer progression. *Cancer Research*, volume 63 (15): 4299–4304.

Xu XM, Yuan GJ, Li QW, Shan SL, J. S. (2012). Hyperthermia Inhibits Transforming Growth Factor Beta-Induced Epithelial-Mesenchymal Transition (EMT) in HepG2 Hepatocellular Carcinoma Cells. *Hepatogastroenterology*, volume 59(119).

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W. and Kanehisa, M. (2008). Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, volume 24 (13).

Yamanishi, Y., Kotera, M., Kanehisa, M. and Goto, S. (2010). Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, volume 26 (12).

Yan, X., Liu, Z. and Chen, Y. (2009). Regulation of TGF-b signaling by Smad7 Overview of TGF-b Signaling Pathways. *Acta Biochimica et Biophysica Hungarica*, volume 41 (4): 263–272.

Yanagisawa, K., Osada, H., Masuda, A., Kondo, M., Saito, T., Yatabe, Y., Takagi, K. et al. (1998). Induction of apoptosis by Smad3 and down-regulation of Smad3 expression in response to TGF-beta in human normal lung epithelial cells. *Oncogene*, volume 17 (13): 1743–1747.

Yang, E. Y. and Moses, H. L. (1990). Transforming growth factor beta 1-induced changes in cell migration, proliferation, and angiogenesis in the chicken chorioallantoic membrane. *The Journal of cell biology*, volume 111 (2): 731–741.

Yang, I. V., Chen, E., Hasseman, J. P., Liang, W., Frank, B. C., Wang, S., Sharov, V. et al. (2002). Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome biology*, volume 3 (11): research0062.

Yang, L. and Agarwal, P. (2011). Systematic drug repositioning based on clinical side-effects. *PLoS ONE*, volume 6 (12).

Yang, Y. H., Dudoit, S., Luu, P. and Speed, T. P. (2001). Normalization for cDNA microarray data. *In* Bittner, M. L., Chen, Y., Dorsel, A. N. and Dougherty, E. R. (Editors), *Microarrays Optical Technologies and Informatics*, volume 4266 of *Proceedings of SPIE*, pages 141–152. Department of Statistics, University of California, Berkeley.

Yang, Y. H. and Speed, T. (2002). Design issues for cDNA microarray experiments. *Nature reviews. Genetics*, volume 3 (8): 579–588.

Yang, Y. H. and Thome, N. P. (2003). Normalization for Two-color cDNA Microarray Data. *Lecture Notes–Monograph Series*, volume 40: 403–418.

Ye, H., Liu, Q. and Wei, J. (2014). Construction of drug network based on side effects and its application for drug repositioning. *PLoS ONE*, volume 9 (2).

Yellaboina, S., Tasneem, A., Zaykin, D. V., Raghavachari, B. and Jothi, R. (2011). DOMINE: A comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Research*, volume 39 (SUPPL. 1).

Yi, S.-G., Joo, Y.-J. and Park, T. (2009). Rank-based clustering analysis for the time-course microarray data. *Journal of bioinformatics and computational biology*, volume 7 (1): 75–91.

Yoon, D., Yi, S.-G., Kim, J.-H. and Park, T. (2004). Two-stage normalization using background intensities in cDNA microarray data. *BMC bioinformatics*, volume 5: 97.

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

Yu, L., Gulati, P., Fernandez, S., Pennell, M., Kirschner, L. and Jarjoura, D. (2011). Fully Moderated T-statistic for Small Sample Size Gene Expression Arrays.

Yuan, J. S., Reed, A., Chen, F. and Stewart, C. N. (2006). Statistical analysis of real-time PCR data. *BMC bioinformatics*, volume 7: 85.

Zacher, B., Khalid, A., Gade, S., Younesi, E., Tresch, A. and Fröhlich, H. (2012). Joint bayesian inference of condition-specific miRNA and transcription factor activities from combined gene and microRNA expression data. *Bioinformatics*, volume 28 (13): 1714–1720.

Zeng, L.-H., Xu, L., Rensing, N. R., Sinatra, P. M., Rothman, S. M. and Wong, M. (2007). Kainate seizures cause acute dendritic injury and actin depolymerization in vivo. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, volume 27 (43): 11604–11613.

Zenke, M. and Hieronymus, T. (2006). Towards an understanding of the transcription factor network of dendritic cell development.

Zhang, J. D. and Wiemann, S. (2009). KEGGgraph: A graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics*, volume 25 (11): 1470–1471.

Zhang, L., Hurek, T. and Reinhold-Hurek, B. (2005). Position of the fluorescent label is a crucial factor determining signal intensity in microarray hybridizations. *Nucleic Acids Research*, volume 33 (19): 1–8.

Zhao, J., Zhang, X.-S. and Zhang, S. (2014). Predicting cooperative drug effects through the quantitative cellular profiling of response to individual drugs. *CPT: pharmacometrics & systems pharmacology*, volume 3 (November 2013): e102.