Genomic Data Integration
&
Gene Expression Analysis of Single Cells

Dissertation zur Erlangung
des Doktorgrades der Naturwissenschaften (Dr. rer. nat.)
der Fakultät für Biologie und Vorklinische Medizin
der Universität Regensburg

vorgelegt von
Herrn Matthias Maneck M.Sc.
aus
Berlin

im Jahr 2011

Das Promotionsgesuch wurde eingereicht am: 22.11.2011

Die Arbeit wurde angeleitet von:

1. Betreuer: Prof. Dr. Wolfram Gronwald
2. Betreuer: Prof. Dr. Rainer Spang

Unterschrift:

For my parents

Contents

List of Figures

List of Tables

Abbreviations and Notation

ABC	-	activated B-cell lymphoma
AUC	-	area under the curve
BCL6	-	B-cell CLL/lymphoma 6
BP	-	biological process
CC	-	cellular component
cDNA	-	complementary DNA
CGH	-	comparative genomic hybridization
ChIP	-	chromatin immunoprecipitation
CV	-	cross validation
Cy3	-	Cyanine-3-dNTP, fluorescently labeled DNA (green)
Cy5	-	Cyanine-5-dNTP, fluorescently labeled DNA (red)
DLBCL	-	diffuse large B-cell lymphoma
DNA	-	deoxyribonucleic acid
DTC	-	disseminated tumor cell
FPR	-	false positive rate
GC	-	germinal center
GCB	-	germinal center like B-cell lymphoma
GO	-	gene ontology
HL	-	Hodgkin lymphoma
HMEC	-	human mammary epithelial cells
KDE	-	kernel density estimation
LPS	-	lipopolysaccharide
MF	-	molecular function
mRNA	-	messenger RNA
n	-	number of samples in a gene expression data set
NHL	-	non-Hodgkin lymphoma

NSC	-	nearest shrunken centroids
p	-	number of genes in a gene expression data set
P	-	a P-value
P_{adj}	-	a P-value adjusted for multiple testing
PAM	-	the cluster algorithm partitioning around medoids
PAMr	-	prediction analysis for microarrays
Pb_2	-	the second polar body
PCR	-	polymerase chain reaction
qPCR	-	quantitative PCR
r	-	a correlation coefficient
rmsd	-	root mean square deviation
RNA	-	ribonucleic acid
ROC	-	receiver operator characteristic
TPR	-	true positive rate
tRNA	-	transport RNA
X, T, G	-	expression matrices of microarray data sets
Y	-	vector containing sample labels

Preface

Acknowledgement

This work was accomplished in the Computational Diagnostics Group of the Institute of Functional Genomics of the University of Regensburg in cooperation with the Chair of Experimental Medicine and Therapy Research of the department of pathology of the University Hospital of Regensburg. First and foremost, I thank my supervisors Rainer Spang and Christoph Klein for giving me the opportunity to carry out my research projects in their groups. This gave me the opportunity to work in both, a theoretical bioinformatics group as well as in a clinically oriented lab.

I thank all past and present colleagues for the pleasant working atmosphere. Especially I thank my office mates Benedict, Christian K., Claudio, and Mohammed, as well as Stefan, Juby, Christian H. and Tully from the Computational Diagnostics group and Isabell, Anya, Daniel and Sophie from the Chair of Experimental Medicine and Therapy Research for fruitful (non-) scientific discussions.

Above all, I am exceptionally grateful to my girlfriend Corinna, my parents and my brother for their relentless support and love.

Publications

Parts of this thesis have been published in peer-review journals. The *guided clustering* algorithm described in chapter 4 has been published in the *Bioinformatics* journal (Maneck *et al.*, 2011). Further, the content of chapter 6 has been published in *The EMBO Journal* (Vermilyea *et al.*, 2011).

Figures

Figures 1.1 and 6.5 have been taken from Jares (2006) and Vermilyea *et al.* (2011) respectively.

Motivation & outline

Cancer is one of the leading causes of death world wide. According to the World Health Organization (WHO) 7.6 million deaths, which is about 13% of all deaths, were the result of cancer in 2008[1]. This frightening number is projected to continue to rise to over 30 million in 2030. In response the WHO launched its Noncommunicable Diseases Action Plan that focusses on preventing and controlling cancer. Prevention of diseases is of course better than treatment or control. However, cancer treatment is one of the major challenges for present health systems. Efficiency and effect of present cancer treatments are mainly linked to an early detection and precise diagnosis. Early detection always involves the patients awareness of early signs and symptoms. A precise diagnosis of cancer is only possible based on tissue biopsies, which are judged by a pathologist according to morphological and molecular properties.

Another way to investigate molecular characteristics of tumor biopsies, which has been subject to intensive research in recent years, is gene expression profiling. In fact, it has been shown by Hummel *et al.* (2006) that microarray gene expression analysis is able to improve the molecular stratification and classify former unknown cases of lymphoma. We believe that information about expression states of thousands of genes measured simultaneously holds even greater potential. This potential can be unlocked if gene expression data is combined with additional data from other experiments that measure specific perturbations on the transcriptome. Such perturbations might be stimulation of cell lines with certain compounds, knock-out or knock-in of specific genes or binding affinities of transcription factors to the DNA. The integration of tumor biopsies with additional experimental data allows the detection of gene clusters within the expression profiles that respond to perturbations. In this thesis the prospects of data integration in this context are subject of chapters 3 and 4. In chapter 3 we will review the data integration literature with respect to gene expression analysis. This is followed by a detailed description of a novel data integration method developed during this thesis in chapter 4.

Recently the analysis of single cells has moved into the focus of cancer research. The present model of disseminated tumor cell (DTC) and early metastatic spread suggests that single tumor cells disseminate from the primary tumor in a very early

[1]GLOBOCAN 2008, International Agency for Research on Cancer

stage of disease and may cause metastasis, even if the primary tumor is removed early as reported by Hüsemann *et al.* (2008). According to Scher and Pantel (2009) the input of DTCs into cancer research is manifold. For example, the analysis of DTCs could establish new prognostic biomarkers that predict disease recurrence after surgery or identify patients in need of antiproliferative therapy. However, the analysis of single cells raises new technological and analytical challenges as the amount of available sample material is low. Usually clinical samples consist of thousands of cells. In this thesis we investigate single cell microarray gene expression profiling in chapter 5. We compare the performance of several normalization procedures with respect to different applications of gene expression analysis, namely differential gene expression analysis and classification. An application of single cell microarray analysis is given in chapter 6. There we leave the field of cancer research and enter reproductive medicine and embryonic research. We analyze samples that naturally consist of only one cell, the murine zygote. We investigate the very first steps of embryogenesis by comparing the single cells of 2− and 3−cell blastomeres. The aim of this analysis is to answer the question whether there exist conserved differences within the transcriptomes of couplets or triplets derived from 2− or 3−cell blastomeres.

In the introductory part of this thesis, namely chapter 1 and 2, the biological terms and concepts used throughout this thesis are described. This is followed by an introduction into the machine learning techniques used in data analysis, namely classification, clustering and kernel density estimation.
The thesis closes with a summary and final conclusions.

Part I

Introduction

We start by introducing the terminological basics and background for the content of this thesis.

In chapter 1 the most important molecular biological terms used are described. This is followed by a description of molecular biological techniques and methods that were used to generate the data sets analyzed in this work. Several textbooks were used for this part of the introduction namely: "Basiswissen Biochemie: mit Pathobiochemie" (Löffler, 2009), "Biochemie" (Stryer, 1996), "Lehrbuch der Molekularen Zellbiologie" (Alberts et al.*, 2005) and "Bioinformatics - Sequence and Genome Analysis" (Mount, 2004) .*

In chapter 2 a brief introduction into the concepts of machine learning methodology used is given. This comprises classification, kernel density estimation and clustering. In this chapter the textbooks "The Elements of Statistical Learning" (Hastie et al.*, 2001), "Elements of Computational Statistics" (Gentle, 2002) and "Semi-Supervised Learning" (Chapelle* et al.*, 2006) as well as the original publications of Kaufman and Rousseeuw (1990), Rousseeuw (1987) and Tibshirani* et al. *(2002) were used.*

Chapter 1

Biological Content

1.1 Molecular Biology

Gene Regulation Each eukaryotic cell contains the whole genome of its specific organism and has therefore the potential to express all genes. However, only a fraction of genes is expressed within a cell in a certain condition and time point. Between different tissues the set of expressed genes differs. This is called differential gene expression. It is of crucial importance for all cells to regulate their gene expression in response to variable environmental conditions. In eukaryotic cells the regulatory mechanisms vary from the inactivation of genes by methylation via mRNA editing or degradation to transcription factors. Within an organism the cells interact and gene regulation is controlled and triggered by a complex network of signals.

Transcription Factors Transcription factors are molecules that can bind to specific DNA sequences. They can either activate or repress the process of transcription and are part of the gene regulation machinery.

Differentiation & Proliferation Highly specialized tissues are responsible for different requirements of an organism. The lungs are responsible for the oxygen absorption, the muscles for movement and the intestine for the absorption of nutrients. However, all the cells of an organism develop out of a single fertilized egg. As an organism develops many cells are needed and the number of cells increases via cell division. This increase of cell numbers is called proliferation. In chapter 6 we investigate the very first cell divisions by analyzing single cells from 2- and 3-cell mice blastomeres and zygotes. In contrast, differentiation is the process that

specialize a cell to a certain biological task.

Cancer Cancer is a group of diseases characterized by uncontrolled cell proliferation. The disease is based on the accumulation of mutations within a single cell. These mutations result in a breakdown of the normal regulatory control mechanisms. Mutations can be caused by errors during DNA replication, contact with harmful chemicals, exposure to radiation or be inherited. Tumors can be developed in all types of tissues and are usually named based on their cellular origin. A tumor can spread to other parts of the body. This process is called metastasis.

Oncogenes & Tumor Suppressor Genes Genes can be directly involved in the formation of cancer. One distinguishes between oncogenes and tumor suppressor genes. Oncogenes are mutated genes that contribute to the development of a cancer. They often code for signaling proteins that are involved in the regulation of cell division. In their mutated form they produce a constant stimulus that drives the cell to proliferate. Tumor suppressor genes work in the opposite direction of oncogenes. They code for proteins that keep the cell division under control. However, if a tumor suppressor gene is mutated the control mechanism can not be maintained. This loss of control over cell division may contribute to the development of a cancer. Identification of onco- and tumor suppressor genes is the key to understanding cancer. Their regulation and protein products are interesting targets for drug development and therapies.

Lymphoma Lymphoma are a type of blood or hematopoietic cancer that can be developed in many parts of the body, including the lymph nodes, spleen, bone marrow and blood. They occur when lymphocytes, a type of white blood cells, proliferate abnormally. The collective term lymphoma contains many subtypes of this disease. There are two main types of lymphoma, the Hodgkin lymphoma (HL) and the non-Hodgkin lymphoma (NHL). The NHL are further classified by their cell of origin. This cell is either a B- or T-lymphocyte, thus the respective lymphoma is referred to as B-cell or T-cell lymphoma. B-cell lymphomas comprise about 95% of all lymphomas. In chapter 4 we analyze data from two aggressive forms, the diffuse large B-cell lymphoma (DLBCL) and the Burkitt lymphoma. While the data set contains samples from both subtypes the analysis focusses on the DLBCL. We investigate the influence of the transcription factor BCL6 on the transcriptome, since BCL6 is a key oncogene in the development of DLBCL.

Cell Lines Cells can be grown separated from the original tissue under controlled conditions. This technique is called cell culture. Permanently established cell cultures that can proliferate indefinitely are called cell lines.

In cancer research cell lines derived from tumors are widely used as model organisms. They are exposed to newly developed drugs or other compounds to test their effects. Another application is the investigation of transcriptome changes given a certain perturbation. This perturbation can be a silenced or activated gene or the application of a specific compound that will activate or block receptors. Then the transcriptome of perturbed cells is compared with control cells without the perturbation using gene expression microarrays. Such experiments give new insights into the regulatory mechanism triggered by the perturbation. In this work data from cell line experiments is integrated with gene expression profiles of lymphoma patients in chapter 4. Further we used cell lines to generate the expression data employed in the evaluation of singe cell gene expression analysis in chapter 5.

1.2 Measuring Techniques

Microarray Today exists a large variety of microarrays technologies to measure different kinds of biological data like gene expression, protein abundance, protein binding or genomic aberrations. Their advantage compared to other methods is, that they measure not only one gene or protein at a time, but thousands. Therefore, microarray data are snapshots of an entire proteome or transcriptome within a sample at a certain time point. In this work we focus mainly on gene expression microarray data, which was also the first application of this technology. There are two primary types of gene expression microarrays: (i) The cDNA microarray which was invented by the Pat Brown laboratory in 1995 (Schena *et al.*, 1995) and (ii) the high-density oligonucleotide array invented by Affymetrix in 1996 (Lockhart *et al.*, 1996).

The basic concept of any gene expression microarray experiment is to extract mRNA from a tissue and reverse transcribe it into cDNA. As a prerequisite for measuring, cDNA is amplified in a way that the relative molecular abundances of different mRNAs are preserved. The cDNA is then labeled with a fluorescent dye and used as a target to bind to complementary DNA sequences. The targets are detected by single-stranded cDNA or oligonucleotide probes that are attached as fixed spots on a support surface like a glass slide. The abundance of a particular mRNA can be measured indirectly by quantifying the fluorescence intensity of the corresponding probe spot. A quantification of the measured mRNA abundance can be done by comparing the spot intensities of a sample to the intensities of a control experiment. This quantification is not absolute but relative to the control. Figure 1.1 illustrates the generation and application of cDNA and high-density oligonucleotide arrays.

Until today microarray technologies have been developed to be routine high throughput methods that have been proven to be a valuable tool of modern molecular biology. Their versatility is highly responsible for their success. They can be employed to detect differentially expressed genes between samples of different biological origin (Chee *et al.*, 1996), or to identify a set of genes that discriminates between different types of samples for diagnosis or prognosis (van 't Veer *et al.*, 2002). They can also be used to define novel subgroups within a certain sample population (Alizadeh *et al.*, 2000). Other microarray technologies can be used for polymorphism analysis (Wang *et al.*, 1998), sequencing (Pease *et al.*, 1994) and de-

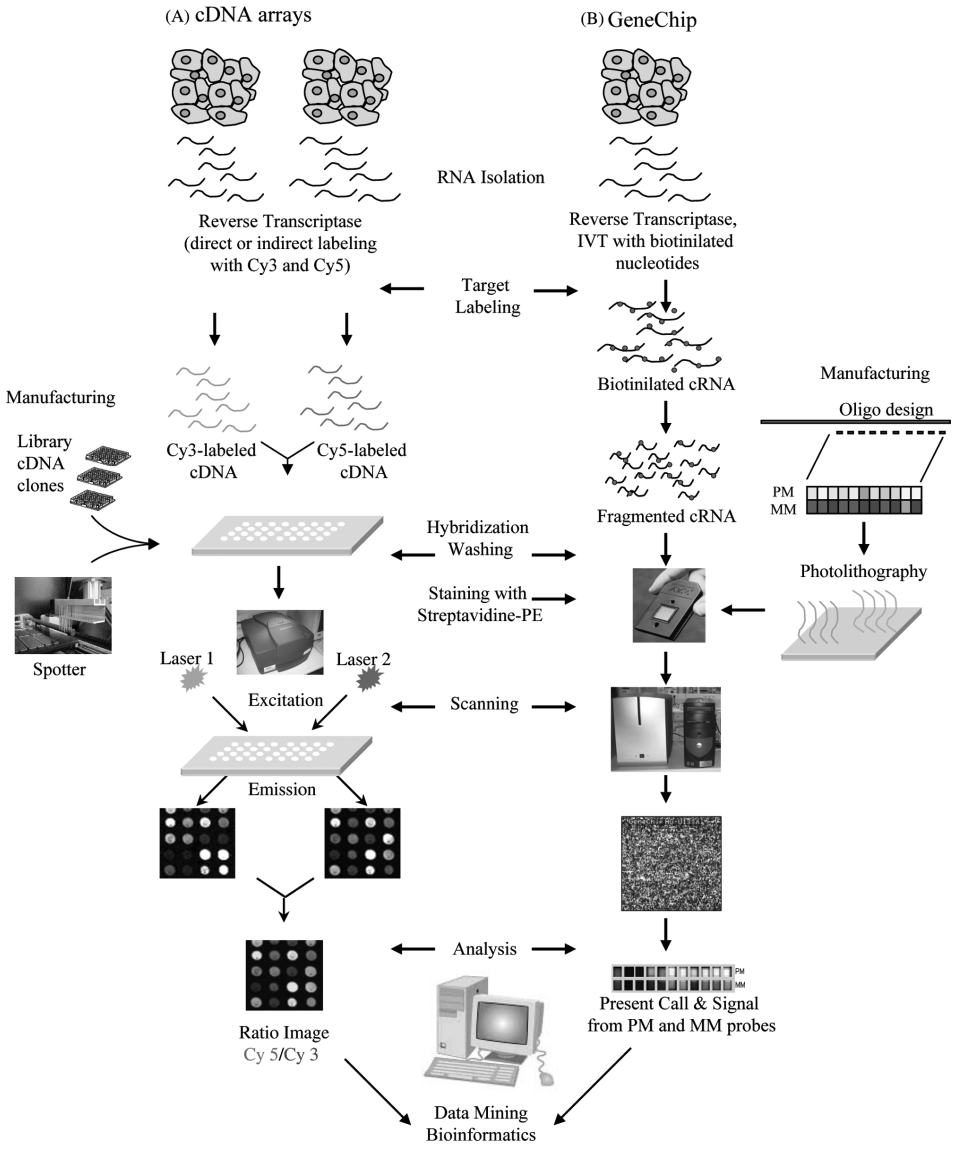


Figure 1.1: Comparison of the production and hybridization steps in (A) cDNA and (B) high density oligonucleotide arrays.

tection of protein DNA interactions (Ren *et al.*, 2000). Data analyzed in this thesis primarily originates from gene expression microarray experiments. In chapter 4 we use additional data from ChIP-on-chip experiments.

qPCR (quantitative Polymerase Chain Reaction) Polymerase chain reaction is a technique for DNA amplification. In theory, PCR can produce an arbitrary number of copies out of a single DNA segment. It was developed in the 1980s by the American biochemist Kary Mullis, who was rewarded with the Nobel Price for

chemistry in 1993.

The PCR process runs in cycles, which each consist of three steps, namely denaturation, primer annealing and elongation. In each cycle the amount of DNA molecules is doubled. In the denaturation step the two DNA strains are separated by heating. Then follows the annealing step in which the temperature is lowered to allow the primer sequences to bind to the single stranded DNA. In the elongation step the DNA polymerase replicates the second strain of each single stranded DNA molecule. It is possible to amplify only specific DNA segments by using only primers that flank the targeted sequence. Hence PCR allows both, the amplification of the whole genome or specific DNA segments. PCR can also be used to amplify RNA. To do so RNA molecules have to be transcribed into cDNA molecules before the amplification process is started. This transformation is called reverse transcription.

Apart from amplification PCR can also be used to quantify the initial amount of targeted transcript. This method is called quantitative PCR (qPCR). For the quantification fluorescent dyes are used that are build into the replicated sequence during the elongation step. With each cycle the dye intensity increases as the number of transcript copies increase. The number of cycles is counted until a specific intensity threshold is reached. This number is compared to a reference PCR using a target sequence with known abundance (Wang *et al.*, 1989). This technique can be used to detect chromosomal aberrations like copy number changes or to measure gene expression if qPCR was done on reverse transcribed cDNA.

Apart from being an integral part of experimental protocols generating microarray data, PCR was used to validate results from microarray gene expression analysis in chapter 4.

ChIP-on-chip ChIP-on-chip combines chromatin immunoprecipitation (ChIP) with the microarray technology (chip). The ChIP technique isolates and identifies DNA sequences that were be bound by specific DNA binding proteins, for example transcription factors (Ren *et al.*, 2000). In principal, DNA is cross linked with the protein of interest and sheared. Then the DNA fragments with a bound protein are separated from all other DNA fragments using antibodies specific to the protein of interest. Depending on the specific method these antibodies are attached to a solid surface or have a magnetic bead that allows the fixation of the protein-DNA complexes, such that the other DNA fragments can be washed away. The bound DNA segments are then separated from the protein of interest and purified. After

amplification the DNA segments are denatured into single stranded DNA fragments. These are labeled with a fluorescent dye. The labeled fragments are then hybridized to a DNA microarray. As each spot of the microarray represents a specific location in the genome one can identify the DNA binding positions by identifying spots with high fluorescence intensities. In chapter 4 we use data from ChIP-on-chip experiments to identify gene clusters that are targets of the transcription factor BCL6.

Chapter 2

Statistical Context

2.1 Machine Learning

Notation This thesis deals with gene expression microarray data. Here we give the notation that will be used throughout the thesis. We will denote data sets consisting of n samples and p genes by a matrix $X \in \mathbb{R}^{p \times n}$. Each row $x_{i\cdot}$ of X contains the expression of a single gene across all samples, while each column $x_{\cdot j}$ contains the expression profile of a single sample. Denoting a gene by an index letter, like any gene g, is equivalent to $x_{g\cdot}$. If we work on more that one data set simultaneously, we will introduce notional conventions when needed. In some cases additional information is available that assigns samples into groups or classes. We will call this information labels and store them in a vector Y with $y_j \in [1, \ldots, k]$ and $j = 1, \ldots, n$, where k is the number of classes.

Un-, Semi- & Supervised Learning Machine learning problems are categorized into three different groups, namely un-, semi- and supervised problems. Unsupervised learning seeks to determine how data is structured. Well known examples of unsupervised methods are clustering and density estimation. Cluster algorithms aim to separate the data objects into a given number of groups. Density estimation analyzes how the data objects are distributed in the data space.

In contrast, supervised learning problems arise from data sets that consist of inputs and outputs. Inputs could be expression profiles $x_{\cdot j}$ of patient samples and the corresponding outputs could be class labels y_j. Supervised methods aim to predict the output from the inputs. Applications of supervised learning are for

example classification and regression methods.

Semi-supervised problems are halfway between supervised and unsupervised problems. Here the output information is given only for a part of the inputs, meaning that labels are only available for some expression profiles. Similar to supervised methods the goal is to predict the outputs of all inputs. Semi-supervised learning methods extend supervised approaches by incorporating structural or distribution properties learned from the unlabeled inputs.

2.1.1 Classification

Classification is one of the major applications of supervised learning. Given a gene expression data set X with corresponding class labels Y, we denote a classifier or prediction model that has been estimated from training data as $f_{class}(X)$. For a certain expression profile $x_{\cdot j}$ the predicted label is denoted as $f_{class}(x_{\cdot j})$. Differences between $f_{class}(x_{\cdot j})$ and the corresponding label y_j are called errors. The number of errors made by a classifier is measured by a loss function

$$L(y_j, f_{class}(x_{\cdot j})) = \begin{cases} 1 & \text{if } y_j = f_{class}(x_{\cdot j}) \\ 0 & \text{if } y_j \neq f_{class}(x_{\cdot j}) \end{cases} \tag{2.1}$$

One distinguishes between two types of errors, the *training* and the *test error*. The *training error* is the average loss over the training samples

$$\overline{err} = \frac{1}{n} \sum_{j=1}^{n} L(y_j, f_{class}(x_{\cdot j})) \; . \tag{2.2}$$

However, to assess the quality of a classifier one uses the *test error*, which is estimated by the average loss over an independent test data set X' containing n' samples and known class labels Y'

$$\overline{Err} = \frac{1}{n'} \sum_{j=1}^{n'} L(y'_j, f_{class}(x'_{\cdot j})) \; . \tag{2.3}$$

In the next part follows a detailed explanation of the *nearest shrunken centroids* (NSC) classification algorithm.

Nearest Shrunken Centroids The NSC method was proposed by Tibshirani *et al.* (2002) and is designed for the analysis of microarray expression data. In this context the task is to classify and predict the diagnostic category of a sample

based on its gene expression profile. This classification problem is challenging since one is faced with a large number of genes from which to predict classes for a small number of samples. It is important to identify the genes that contribute most to the classification, so that non-contributing genes can be discarded.

Given a gene expression data set X and a vector of corresponding class labels $Y \in 1, 2, \ldots, K$, let C_k be a vector of indices of the n_k samples belonging to class k. The centroid $\bar{x}_{.k}$ of a class k is defined as the average expression per gene across all samples of that class: $\bar{x}_{ik} = \sum_{j \in C_k} x_{ij}/n_k$. Similar, the overall centroid is defined as the average expression per gene across all samples $\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n$. In order to identify genes that contribute most to the classification the class centroids are shrunken stepwise towards the overall centroid. This is done by comparing the centroid of each class k to the overall centroid. We define a difference d_{ik} for each gene as

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k \cdot (s_i + s_0)} \tag{2.4}$$

where each gene is standardized by the pooled within-class standard deviation

$$s_i^2 = \frac{1}{n - K} \sum_k \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2 \tag{2.5}$$

and $m_k = \sqrt{1/n_k + 1/n}$, so that $m_k \cdot s_i$ is equal to the estimated standard error of the numerator in d_{ik}. The positive constant s_0 guards against large d_{ik} values arising from genes with low overall variance. Shrinkage is done by reducing the values of d_{ik} stepwise towards zero. This process is tuned by a parameter Δ

$$d'_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+ \, , \tag{2.6}$$

where $\text{sign}(x)$ equals 1 or -1 if x is positive or negative and $(|d_{ik}| - \Delta)_+$ denotes the maximum of $|d_{ik}| - \Delta$ and zero. Genes are discarded if d'_{ik} becomes zero. A shrunken class centroid $\bar{x}'_{ik}$ can then be defined by rewriting equation 2.4 as

$$\bar{x}'_{ik} = \bar{x}_i + m_k(s_i + s_0)d'_{ik} \, . \tag{2.7}$$

A sample $x_{.j}$ is classified by determining the closest shrunken class centroid. The distance between a class centroid and a sample is measured by the discriminant score

$$\delta_k(x_{.j}) = \sum_{i=1}^{p} \frac{(x_{ij} - \bar{x}'_{ik})^2}{(s_i + s_0)^2} - 2 \log \pi_k \tag{2.8}$$

where π_k is the prior probability of class k, that is defined as the overall frequency of class k in the population, hence $\sum_{k=1}^{K} \pi_k = 1$. Usually π_k is set to n_k/n. The prediction model is then defined as

$$f_{class}(x_{\cdot j}, \Delta) = r \text{ where } \delta_r(x_{\cdot j}) = \min_k \delta_k(x_{\cdot j}) \; . \tag{2.9}$$

The result of the NSC classifier depends on the chosen shrinkage parameter Δ. We chose Δ so that the *test error* is minimized by employing a cross validation procedure as described below.

Cross Validation Cross validation (CV) is a method widely used for model assessment and selection. Here we will consider CV in the context of classification. CV is one of the simplest ways for estimating the *test error* of a classifier or prediction model. Given a gene expression data set X with corresponding class labels Y and a prediction model $f_{class}(X)$ one can assess the performance of the $f_{class}(X)$ using an independent test set. Unfortunately, this is often not possible since data is scarce. To circumvent this problem CV uses one part of the available data to train the model and another part to test it. The data is split into K subsets (or folds) of equal size (K-fold CV). Each of the K subsets is predicted by a model trained on the other $K-1$ sets. Figure 2.1(a) illustrates the arrangement of a data set into K subsets and the K training / prediction runs. The *test error* is then calculated by combining the errors in the K predicted subsets. The CV estimate of the *test error* is defined as

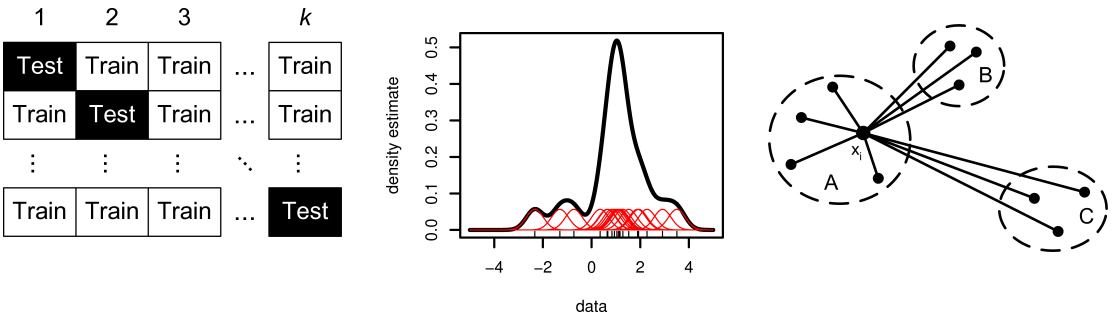
$$\overline{Err}_{CV} = \frac{1}{n} \sum_{j=1}^{n} L(y_j, f_{class}^{-k(j)}(x_{\cdot j})) \tag{2.10}$$

where $f_{class}^{-k(j)}(x_{\cdot j})$ is the prediction model trained without the subset k that contains sample j. Common choices of K are 5 or 10. If K equals n one speaks of *leave-one-out* cross validation. In this work 10-fold CV was used if not stated otherwise.

If the training procedure of a prediction model depends on a tuning parameter Δ, one can use CV to determine the optimal choice of Δ. In case of the NSC classifier Δ is the amount of shrinkage. We index the set of models by $f_{class}(x_{\cdot j}, \Delta)$. For each Δ the CV error is defined as

$$\overline{Err}_{CV}(\Delta) = \frac{1}{n} \sum_{j=1}^{n} L(y_j, f_{class}^{-k(j)}(x_{\cdot j}, \Delta)) \; . \tag{2.11}$$

$\overline{Err}_{CV}$ is an estimate of the *test error* as a function of Δ. The optimal choice of the tuning parameter Δ is Δ' that minimizes this function. The optimal prediction



(a) (b) (c)

Figure 2.1: **(a)** Cross validation scheme: Each row represents an individual classification run, while the columns indicate the division of the samples into k sub sets. During each run, the classifier is learned on the training set and used to predict the test set. **(b)** Density estimation for a small data set consisting of 20 elements. The density estimate at each point (black line) is the average contribution from each of the Gaussian kernels (red lines). The kernels were scaled down by the factor of 20 to fit in the graph. The data points are indicated by vertical bars at the bottom. **(c)** Arrangement of elements involved to calculate the silhouette $s(i)$ of element x_i belonging to cluster A. B is the nearest neighborhood cluster of x_i and C is an other, more distant cluster.

model is then $f_{class}(X, \Delta')$. In case of the NSC classifier Δ' is determined by trying different choices of Δ.

2.1.2 Kernel Density Estimation

Given the gene expression data set X we assume that the p genes $x_{i\cdot}$ are sampled independently from an unknown distribution. Kernel density estimation (KDE) is an unsupervised learning procedure that aims to estimate the probability density function $f_{dens}(x)$ from the given data set X, for any $x \in \mathbb{R}^p$. A local estimate of the probability density at x has the form

$$\hat{f}_{dens}(x) = \frac{\#\{x_{i\cdot} \in \mathcal{N}(x)\}}{p\,\sigma} \tag{2.12}$$

where $\mathcal{N}(x)$ is a small metric neighborhood around x of width σ and $\#\{\ldots\}$ counts the number of genes in $\mathcal{N}(x)$. This function tends to be rough, since genes can either be inside or outside a neighborhood $\mathcal{N}(x)$. A smoother function can be achieved by using the *Parzen* estimate, which weights the counting as a function of distance to

$x_{i\cdot}$

$$\hat{f}_{dens}(x) = \frac{1}{p\,\sigma} \sum_{l=1}^{p} K_\sigma(x_{l\cdot}, x) \tag{2.13}$$

where $K_\sigma(x_{l\cdot}, x)$ is a function with bandwidth σ that gives decreasing weights to genes $x_{l\cdot}$ as the distance to x increases. K_σ can be seen as a function that measures the similarity between two genes. The closer the genes are, the higher is their similarity. Such functions are also called kernels. The Gaussian kernel is a popular choice for K_σ. Assuming mean zero and standard deviation σ equation 2.13 has the form

$$\hat{f}_{dens}(x) = \frac{1}{p\,\sqrt{2\sigma^2\pi}} \sum_{l=1}^{p} \exp\left(\frac{-||x_{l\cdot} - x||^2}{2\sigma^2}\right) . \tag{2.14}$$

An example for density estimation using a Gaussian kernel is shown in Figure 2.1(b).

The bandwidth σ of the kernel is a parameter that has a strong influence on the resulting density estimate. Small σ lead to under-smoothing resulting in spurious data artifacts. On the other hand large values lead to over-smoothing which obscures much of the underlying data structure. A popular measure to assess the accuracy of the estimated density function is the mean integrated squared error (MISE) suggested by Rosenblatt (1971):

$$\text{MISE} = E|f_{dens}(x) - \hat{f}_{dens}(x,\sigma)|^2 = \frac{1}{p}\sum_{i=1}^{p}\left(f_{dens}(x_{i\cdot}) - \hat{f}_{dens}(x_{i\cdot},\sigma)\right)^2 \tag{2.15}$$

where $\hat{f}_{dens}(x,\sigma)$ is a set of estimated density functions indexed by σ. However, this measure assumes that the true density is known which is usually not true. This led to numerous data driven bandwidth selection methods of which the cross validation and plug-in methods suggested by Rudemo (1982), Bowman (1984), Park and Marron (1990) and Sheather and Jones (1991) are most popular and well established for low dimensional data. Unfortunately, when working with high dimensional data like gene expression measurements such methods are not applicable as they are computational complex and time intensive. A careful selection by hand is more practicable.

2.1.3 Clustering

Clustering can be defined as the segmentation of data based on a (dis-) similarity measure. This segmentation results in a grouping of data objects into subsets or 'clusters', such that data objects in the same cluster are more similar than data

objects in different clusters. Importantly, each object needs to be assigned to exactly one cluster. Additionally, some cluster methods arrange the clusters in a natural hierarchy. This is done by successively grouping the clusters themselves so that at each level of hierarchy clusters within the same group are more similar as those in different groups.

There exist different types of cluster algorithms. A very popular type is combinatorial clustering. Such algorithms directly assign data objects to clusters without referring to probabilistic data models. Given a gene expression data set X, the data objects that are to be clustered can be genes as well as samples. We define the dissimilarities between each pair of data objects, here samples, by a distance function $d(x_{\cdot j}, x_{\cdot l})$. The pairwise distances are stored in a symmetric matrix D. A popular distance measure is the Euclidean distance

$$D_{jl} = d(x_{\cdot j}, x_{\cdot l}) = \sqrt{\sum_{i=1}^{p} (x_{ij} - x_{il})^2} \; . \tag{2.16}$$

Clustering aims to separate the samples into K clusters C_i, with $i = 1, \ldots, K$ where each cluster C_i stores the sample indices of cluster i, so that samples within the same cluster are more similar than samples within different clusters. The within cluster distances, can be measured by a loss function

$$D_{within} = \sum_{k=1}^{K} \sum_{j \in C_k} \sum_{l \in C_k} D_{jl} \; . \tag{2.17}$$

A clustering that minimizes D_{within} will be optimal, as it automatically maximizes the between cluster distances

$$D_{between} = \sum_{k=1}^{K} \sum_{j \in C_k} \sum_{l \notin C_k} D_{jl} \tag{2.18}$$

since

$$D_{total} = D_{within} + D_{between} = \sum_{j=1}^{n} \sum_{l=1}^{n} D_{jl} \; . \tag{2.19}$$

The optimal clustering can be found by enumerating all possible partitions into K clusters. However, this is only feasible for small data sets as the number of combinations growth rapidly with increasing data size. To circumvent this problem practical combinatorial cluster algorithms try to examine only a small fraction of all possible partitions. The goal is to identify a small subset that is likely to contain the

optimal one, or at least a good suboptimal partition. The algorithm that was used in the thesis is *partitioning around medoids* that is described in the next section.

Partitioning around Medoids The *partitioning around medoids* (PAM) algorithm was proposed by Kaufman and Rousseeuw (1990). Given a gene expression data set X and the number of clusters K, the method aims to find a set M of K representative data objects (medoids). This set defines a clustering by dividing the data objects into K groups, so that each object belongs to its nearest medoid. The set of medoids is chosen such that the sum of distances of each data object to its closest medoid is minimized. The PAM algorithm consists of two phases, the *build* phase and the *swap* phase. During the *build* phase K medoids are added consecutively to the set of medoids M, while in the *swap* phase this set is improved. In the following we assume that we cluster the samples $x_{\cdot j}$ of the gene expression data set X.

The *build* phase is initialized by selecting the first medoid m_1. This is the sample that minimizes the sum of dissimilarities between itself and all other samples. Hence, m_1 is the sample that is most centrally located in the data set. All other medoids m_i with $i = 2, \ldots, k$ are chosen iteratively by maximizing the loss function

$$L_{build}(x_{\cdot j}) = \sum_{l=1}^{n} \left(\min_{m \in M} [d(m, x_{\cdot l})] - d(x_{\cdot j}, x_{\cdot l}) \right) \tag{2.20}$$

where $\min_{m \in M}[d(m, x_{\cdot l})]$ is the minimal distance between a sample $x_{\cdot l}$ and the already chosen medoids m, and $d(x_{\cdot j}, x_{\cdot l})$ is the distance between two samples $x_{\cdot j}$ and $x_{\cdot l}$. $L_{build}(x_{\cdot j})$ is the number, by that the sum of distances between samples and their closest medoids would be reduced if $x_{\cdot j}$ was added to the set of medoids. In each iteration step L_{build} is calculated for all samples $x_{\cdot j}$ and the sample that maximizes L_{build} is added to the set of medoids M. This procedure is repeated until K medoids are found.

In the *swap* phase the algorithm aims to improve the set of medoids M and therefore also the clustering implied by M. This is done by iteratively searching the pair of objects $(m_i, x_{\cdot j})$ that if *swapped*, reduces sum of distances between samples and their closest medoid most. A *swap* $(m_i, x_{\cdot j})$ means that m_i is removed from and $x_{\cdot j}$ is added to the set of medoids. We denote the new set of medoids generated by a *swap* as M'. The loss function L_{swap} measures the change of minimal distances

between samples and their nearest medoids induced by the *swap* $(m_i, x_{\cdot j})$.

$$L_{swap}(m_i, x_{\cdot j}) = \sum_{l=1}^{n} \min_{m' \in M'} d(m', x_{\cdot l}) - \sum_{l=1}^{n} \min_{m \in M} d(m, x_{\cdot l}) \tag{2.21}$$

Details of an efficient implementation of this loss function can be found in Kaufman and Rousseeuw (1990). $L_{swap}(m_i, x_{\cdot j})$ can take negative or positive values depending on whether the *swap* was beneficial or not. Negative values indicate beneficial *swaps* as sum of minimal distances between samples and their closest medoids is reduced. The optimal *swap* is the one that minimizes L_{swap}. This iteration is repeated until no beneficial *swap* remains.

Following this heuristic procedure the resulting partition will not necessarily be the global optimum with minimal sum of the within cluster distances D_{within} (see equation 2.17) but a good approximation. The number of clusters K in which the data set is separated is a parameter and has to be specified. To select the optimal K a measure is needed that assesses the quality of a given clustering. However, sum of distances between data objects and their closest medoid, which is a result of the PAM procedure, or D_{within} can not be used, as it decreases naturally with increasing K. Because of this an external cluster validation score has to be used. In this thesis the silhouette score is employed.

Silhouettes Various methods are available that segment a given data set into a set of K clusters, like $K-$means, $K-$medians or the PAM algorithm that is used throughout this thesis. All of these methods result in a set of clusters, where each cluster contains a certain number of data objects. However, it remains unclear if these clusters reflect a structure that is truly present in the data, or if the data was just segmented into some artificial groups. In the end, cluster methods always segment the data into K groups, regardless whether this is supplied by data structure or not. A method is needed that assess the quality of a given clustering. Different methods have been proposed in the literature and some of them were compared by Smolkin and Ghosh (2003). The various methods mainly differ in their definition of cluster quality. In this thesis we employ the method proposed by Rousseeuw (1987) and calculate *silhouettes* to assess cluster quality. Using *silhouettes*, clusters are of high quality if data objects within the same cluster have low dissimilarities or distances compared elements in different clusters. This definition is particularly justified with respect to the PAM algorithm, since PAM aims to segment data, such that the within cluster distances are minimized.

Given a gene expression data set X, a matrix of pairwise distances D and a set of clusters, then the silhouette s_i of a sample $x_{\cdot i}$ belonging to cluster A is defined as

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{2.22}$$

where

$$a_i = \frac{1}{N_A} \sum_{x_{\cdot j} \in A} d(x_{\cdot i}, x_{\cdot j}) \tag{2.23}$$

is the average distance of $x_{\cdot i}$ to all other samples of A, with N_A being the number of samples in A, and

$$b_i = \min_{C \neq A} \frac{1}{N_C} \sum_{x_{\cdot j} \in C} d(x_{\cdot i}, x_{\cdot j}) \tag{2.24}$$

is the minimal average distance of $x_{\cdot i}$ and all samples of any cluster C different from A. We denote the cluster for which the minimum b_i is achieved as neighbor B of x_i. Figure 2.1(c) illustrates the arrangement of the clusters A, B and C. If a cluster contains only one element $s_i = 0$.

From the definition above follows that

$$-1 \leq s_i \leq 1 \tag{2.25}$$

for each sample $x_{\cdot i}$. Large s_i, that are close to 1, imply that the within cluster dissimilarity a_i is much smaller than any between cluster dissimilarity b_i. Hence, $x_{\cdot i}$ can be considered as well clustered. Vice versa small values of s_i mean that a_i is much larger than b_i, so that on average $x_{\cdot i}$ is situated much closer to its neighbor cluster B than to A. Therefore it seems to be more natural to assign $x_{\cdot i}$ to its neighbor B, meaning that $x_{\cdot i}$ has been misclassified. If s_i is about zero, then a_i and b_i are approximately equal and it is not clear if $x_{\cdot i}$ belongs to cluster A or B.

The silhouettes define a measure how well each sample $x_{\cdot i}$ matches the given clustering. To assess the quality of a given clustering we calculate the average silhouette

$$\bar{s} = \sum_{i=1}^{N} s_i \ . \tag{2.26}$$

One can compare a set of different clusterings by comparing their average silhouettes. The clustering that obtains the maximal average silhouette fits best to the data structure. Following this, the parameter K of the PAM algorithm can be tuned by maximizing the average silhouette.

2.1.4 Differential Gene Expression

Given a gene expression data set X consisting of two different types of samples, for example tissue biopsies of healthy (group A) and diseased patients (group B), then differentially gene expression analysis describes a set of methods that answer the question whether the expression levels of a gene $x_{i\cdot}$ differs systematically between the two groups. This difference is usually defined as the difference between the mean expression within group A and B.

While the details may differ depending on the method, the general concept is to assume that a gene is not differentially expressed between the groups and then test this null-hypothesis using an appropriate statistical test. Formally we define the null-hypothesis H_0 for each gene g:

$$H_0 : \text{The gene is not differentially expressed.}$$

If H_0 is rejected by the test, then g is considered to be differentially expressed between the two classes. The test that is most commonly used is Student's *t-test*, which combines the mean difference between the groups with the variance within the groups. Assuming equal variance in both groups the t-score is defined for each gene i as:

$$t_i = \frac{\bar{x}_{iB} - \bar{x}_{iA}}{s_i} \ , \tag{2.27}$$

where $x_{iA} = \frac{1}{n_A} \sum_{j \in A} x_{ij}$, the value of x_{iB} is defined similar, and n_A and n_B are the number of samples in the groups. The pooled standard deviation s_i is defined as

$$s_i^2 = \frac{1/n_A + 1/n_B}{n_A + n_B - 2} \sum_{k=\{A,B\}} \sum_{j \in k} (x_{ij} - \bar{x}_{ik})^2 \ . \tag{2.28}$$

A look-up of t_i in the t-distribution delivers the probability (p-value) of obtaining a test statistic with absolute value of at least $|t_i|$ by chance. The p-value is then used to decide whether H_0 is rejected or not. If the p-value is below a certain significance threshold, usually 0.05, H_0 is rejected and a gene is said to be differentially expressed. However, in the context of gene expression analysis this *raw* p-value can be misleading since several thousand genes are tested simultaneously. A large number of simultaneous tests leads to an unnatural high rejection rate of H_0. The percentage of genes equal to the selected significance threshold will be reject H_0 by chance. This observation is called multiple-testing problem. Several statistical methods have been developed to correct the *raw* p-values for multiple testing. Throughout this

thesis the method proposed by Benjamini and Hochberg (1995) was used. The resulting adjusted p-values (P_{adj}) reflect the false discovery rate (FDR), which is the expected proportion of incorrectly rejected null hypotheses.

Part II

Genomic Data Integration

This part focusses on genomic data integration in the context of gene expression experiments using microarray technology.

We begin in chapter 3 with reviewing the literature on data integration methods and concepts. Since data integration is a broad field with application ranging across various methods, data sources and types, we will focus on literature that is closely connected to the topic of this thesis.

We continue in chapter 4 with the introduction of a novel data integration method for the simultaneous analysis of clinical microarray gene expression profiles and experimental data.

Chapter 3

A review on genomic data integration

In recent years, the number of available molecular biological methods has immensely increased. Today, modern technologies allow an almost complete characterization of biologic samples on the cellular level. Complementary properties include determination of genotypes using PCR or DNA sequencing and investigation of the DNA methylation status as well as transcriptome, proteome and metabolome snapshots. Further, chromatin immunoprecipitation assay like ChIP-seq allow protein-DNA interaction studies. Additionally, protein-protein interaction and protein localization screens are available that allow deeper insights into the biomolecular interplay within cells. Many more methods exist, but not all can be mentioned here. However, even if partly complementary, each profiling method sheds different light on the functioning and malfunctioning of cells. Their joint full potential can only be realized when different sources are combined. Furthermore, research is done on and with different species and model organisms. New challenges arise in cross-species analyzes since genomes and biomolecular interplay are similar but not identical. However, the value of model organism in biological and medical research has been shown in many applications. The development of new drugs and therapies always involves animal tests and cell culture experiments. Thus, integrating data from different sources is an important part of modern biomedical research. An important aspect of data integration is the development of tailored statistical methods that are able to leverage knowledge contained within a diverse range of data sources and at the same time, being able to provide evidence to answer the types of question being posed

by the research community as a whole. While the concept of statistical data integration is self evident, its realization in genomics is challenging. Obstacles include the heterogeneity of experimental setups, study designs, profiling platforms, sample handling, and data management. Furthermore, missing meta-data and insufficient documentation of heuristic and complex multistep analysis procedures complicate the endeavor.

According to Pavlidis *et al.* (2002) data integration methods can be divided into three different types depending on the analysis step at which the integration is done. They distinguish between early, intermediate and late integration. Early integration takes place on the level of input data by simply unifying different data sets to a single one. The analysis is then done by applying standard methods to the unified data. This approach demands that the different data sets are of similar type or have a common feature space and scale. For example one can combine different gene expression data sets if the same platform was used, or if the platforms differ the measured genes overlap. In contrast late integration takes place on the level of analysis results. Here data from different sources is analyzed individually and the final statistical results are combined. Following this approach heterogeneity in data type and analysis procedure can be overcome. Intermediate integration describes the simultaneous analysis of several data sources. The integration step is implemented within the analysis procedure. Such methods are more complex than the analysis of a single source and are often tailored towards a specific problem. However, simultaneous analysis allows a more flexible manipulation of the integration process. In their work Pavlidis *et al.* (2002) shows that the intermediate integration is often superior to early and late approaches but requires sophisticated weighting of each data type.

In biomedical research late data integration enjoys great popularity as it allows the combination of heterogeneous data types. The prerequisite for late data integration is converting data from different sources into a common format. One format that is widely used are gene sets or ranked lists of genes as they can represent a variety of different informations like expression differences, binding affinities and molecular functions or processes. Different methods have been proposed by Beissbarth and Speed (2004), Subramanian *et al.* (2005) and Lottaz *et al.* (2006) tailored towards specific problem settings. In order to gain biological insights from a gene list it is necessary to analyze the functional annotations of all genes in the list.

These annotations were experimentally determined and validated. Each annotation describes properties of a set of genes or their products. Each gene may have several annotations. Beissbarth and Speed (2004) identify annotations that are statistically significant within a given gene list by comparing the annotation present in the list with a control set of genes. While the gene list of interest is usually a handful of genes showing strongest response to the experiment, the control set can be a database of annotated genes or all annotated genes represented on a microarray. Annotations significantly over represented in the gene list compared to the control set are descriptive for the list and experiment. A major drawback of this over-representation analysis is that the gene list was usually produced by a single-gene analysis like differential expression analysis. This may miss important effects on pathways, since cellular processes often affect sets of genes. For instance, a relative small increase in the expression of all genes encoding members of a metabolic pathway may dramatically influence the biologic processes regulated by that pathway more dramatically than a high increase of expression in a single gene. To overcome this problem Subramanian *et al.* (2005) suggest to measure the enrichment of gene sets in ranked lists that include all genes measured in the experiment. Consequently they call their approach gene set enrichment analysis (GSEA). Given multiple ranked gene lists Lottaz *et al.* (2006) detects considerable overlaps among the top-ranking genes. Gene sets conserved across ranked lists emerging from different experiments (different species) may hold information about conserved biological processes.
Integration on the level of gene lists avoids the need for joint quantitive data models that describe the dependencies between individual profiles.

A quantitative early integration approach is the concatenation of feature vectors from different platforms in the context of classification problems. If several types of high dimensional readouts are available for the same group of samples, predictive signatures can be constructed by combining selected features across all data types thus exploring potential complementary information. Somewhat surprisingly, several authors observed only marginal improvements in classification accuracy resulting from early data integration. Boulesteix *et al.* (2008) report on several cancer types where the integration of microarray data with standard clinical predictors, like age or sex is not beneficial. The authors give two possible explanations for this observation. First, the microarray data might be simply not relevant for the classification problem and therefore not improve any classifier nor be able to achieve high

accuracy alone. Second, the microarray is relevant for the prediction problem, but redundant or weaker than the already available clinical parameters. In this scenario classification on the microarray data alone would have good performance. However, such redundancy does not imply a causal relationship between clinical parameters and gene expression. Edén *et al.* (2004) already observed that 'good old' clinical markers have similar prognostic power as microarray gene expression data. Similar results were reported by Lu *et al.* (2005) in the context of prediction protein-protein interactions from genomic features like mRNA co-expression, functional similarity or phylogenetic profiles. The authors extended the original feature list of Jansen *et al.* (2003) by additional features. However, even if those new features showed high univariate prediction strength and were largely conditionally independent no improvement of classifier performance was observed. Lu *et al.* (2005) reason, that integrating a few good features already approaches the maximal predictive power, or limit, of current genomic data integration.
Both examples show that simply piling up more and more data sources will not necessarily lead an improvement of existing analysis results or new biological insights.

In contrast to integrating several data sources within the same sample population data integration methods can also be used to aggregate information across different sample populations. A prominent example for this application is the integration of clinical and experimental sample population in the field of gene expression analysis. In this context Bild *et al.* (2006) combined expression data generated experimentally by overexpression of active oncogenes in quiescent primary human mammary epithelial cells (HMECs) with tumor samples of various carcinomas. Oncogene specific gene signatures were identified by comparing the oncogene transfected HMECs with a control group employing a Bayesian binary regression procedure proposed by West *et al.* (2001). These signatures were then used to predict oncogenic activation probabilities of the tumor samples by applying the regression model. The authors showed that combinations of activation probabilities of different oncogenes can predict outcome and treatment efficiency of the cancer samples. This analysis strategy is sequential in that predictive gene sets are identified and combined to predictive signatures in the HMECs only. In a second independent integration step they are applied to clinical data. Note that this procedure is fully supervised as signatures are not affected by properties of the clinical data.
The exact opposite sequential analysis strategy was described by Läuter *et al.* (2009).

The authors start their analysis on the clinical data. For each gene they generate a gene set by selecting all genes that exceed user a defined correlation threshold with the respective gene. Those gene sets are then tested for joint differential expression between the different experimental conditions within the HMECs. The authors reject all gene sets that do exceed a certain threshold of significance. Although the authors ensure correlation between genes within the tumor samples and joint differential expression in the HMECs, the HMECs do not influence the formation of gene clusters.
The first analysis approach combining both data sets from the onset was described by Bentink *et al.* (2008). This new approach is based on an unsupervised class discovery procedure proposed by von Heydebreck *et al.* (2001) that searches for bipartitions within a gene expression data. Bentink *et al.* (2008) refined this method to meet a semi-supervised scenario in which the tumor data is the unlabeled and the HMECs the labeled data. This semi-supervised approach allows the authors to find classification of tumor samples based on coherently expressed genes, that simultaneously separate experimental conditions.

In the flowing chapter of this thesis we will complement and extend the approach of Bentink *et al.* (2008). We will develop a novel data integration strategy named *guided clustering* that combines experimental and clinical high throughput data of possibly different genomic data types. *Guided clustering* is tailored to analysis scenarios, where the construction of a diagnostic signature is not driven by class labels on the clinical data, for instance disease types or clinical outcomes, but by a biological focus for example the activity of a transcription factor or an entire pathway. The biological focus of the signature is established by a complementing experimental study on model organisms like cell lines or mice. Examples for those experimental studies are cell line perturbation experiments as performed by Bild *et al.* (2006) or profiling of CpG methylation status as done by Gebhard *et al.* (2006). *Guided clustering* uses a density estimation approach to identify sets of genes that show strong response to the experimental condition while at the same time display coherent expression across the clinical data set. In contrast to the semi-supervised approach of Bentink *et al.* (2008) *guided clustering* operates fully unsupervised. Nevertheless the feature selection procedure that drives ordering of patient samples is guided by the complementing experimental data. Furthermore *guided clustering* extends the framework of previous methods in that it allows the

integration of data from different genomic platforms. Instead of separating clinical samples into classes our approach provides quantitative predictions of experimental conditions like pathway activation.

In principal, *guided clustering* can be applied to any data integration problem that links a sample clustering problem to a feature selection problem driven by a second data set. In the next chapter we will introduce our method in detail and compare its performance with other approaches. Further, we will report on two exemplary applications: (i) The prediction of transcription factor activity in clinical samples guided by a chromatin immunoprecipitation experiment and (ii) the prediction of pathway activity guided by cell culture perturbation experiments.

Chapter 4

Guided Clustering

4.1 Problem Setting

Gene expression analysis of cancer biopsies has been subject of intensive research in recent years. A detailed characterization of patient samples is vital for diagnosis and treatment. There exists a multitude of techniques that allow determination of different characteristics, for example staining of protein markers, comparative genomic hybridization (CGH) to determine genomic aberrations and other microarray technologies that create snapshots of the transcriptome or proteome. In practice not every technique can be applied to every sample since modern biomolecular methods are often cost intensive. However, due to the interaction between all cellular components and processes malfunctions can be observed with several measurements. For example genomic aberrations can be directly detected by CGH, but they will also effect the transcriptome and proteome, so that they can often be detected by microarrays too as shown by Mullighan *et al.* (2009). Gene expression microarrays have been proven to be a valuable tool to distinguish between cancer types and subtypes (Golub *et al.* (1999), Alizadeh *et al.* (1999), Bea *et al.* (2005) and Dave *et al.* (2006)) and also defined new disease subtypes (Alizadeh *et al.* (2000) and Sotiriou *et al.* (2003)). This refinement of diagnosis can greatly improve the treatment efficiency of patients, since different cancer subtypes show different drug resistances (Staunton *et al.*, 2001). Obviously, it would be extremely helpful, if one could specifically design experiments to search for subgroups of interest. Experiments necessary to define such groups are for example knock-down through RNAi, the transfection of constitutively active forms of genes, or application of drugs. However, such ex-

periments can not be conducted directly on the patient population. This lead to the approach to perform experiments of interest on model organisms, observe the responses and search similar patterns in patient data. For example, gene expression profiles of a model organism can measure the gene-wise response to certain perturbations. The knowledge of those responses have two mayor benefits: (i) It may guide to potentially interesting sets of correlated genes, since they react to perturbation, and (ii) delivers a possible explanation for the coherent expression among patient samples. This setting leads to the data integration problem of identifying sets of genes that are coherently expressed across patient samples and showing perturbation response in a model organism. In the following we will refer to the perturbation experiment as the guiding data.

Let T_{ij} be a matrix of tumor expression profiles and G_{ij} the guiding data, where rows i denotes genes while columns j denotes samples. For simplicity, we assume that the same profiling platform is used and hence the same genes are monitored in both data sets. The guiding data set G consists of two types of samples: Those which were subject to perturbation and the corresponding unperturbed control group. Since we know which profiles were experimentally perturbed and which not the guiding data set is labeled. The labels are stored in a binary vector Y, where a perturbed profile is indicated by a one and a control sample by a zero. Like experimental perturbations, somatic mutations in cancers can effect gene regulation and leave similar traces in expression profiles. Hence, we assume that the effect modeled by the perturbation experiment is also present within the tumors samples and varies across T due to different somatic mutations. However, a priori it is unknown in which tumor samples a perturbation effect is present and to what extend. Thus the data set T is unlabeled.

In the following sections we describe an algorithm called *guided clustering* that aims to detect and quantify the presence of a perturbation within tumor data samples by identifying sets of genes that are coherently expressed within T and show strong perturbation response in G. The findings of this work are published in the peer-reviewed journal *Bioinformatics* (Maneck *et al.*, 2011).

4.2 Algorithm

For clarity, we first describe guided clustering in the application context of oncogenic pathway activation in tumor samples, and later describe a series of modifications that adapt the method to different applications.

In general, perturbation of a pathway will lead to either induction or repression of its target genes. Consequently, the guiding data set G will contain genes that are up and genes that are down regulated within the perturbed group relative to the control group. Hence, depending on the particular gene, pathway activation is indicated by high or low expression. To simplify computation, we multiply the expression values of pathway repressed genes by -1 in both datasets, such that numerically all targets display 'high' expression upon pathway activation.

4.2.1 Definition and fusion of similarity matrices

In order to find sets of genes that are coherently expressed within T we need to define a measure of coherence between genes in the tumor data T. We use a correlation based gene distance metric as suggested by Eisen *et al.* (1998). For any two genes g and h the distance between g and h is defined as:

$$d(g,h) = 1 - \max\left(\rho(g,h), 0\right) \ , \tag{4.1}$$

where $\rho(g,h)$ is the Spearman correlation between the expression vectors of genes g and h. Spearman's correlation coefficient guards against the disturbing influence of single outlier expression values as it converts the expression values to ranks before the correlation is calculated. Note that anti-correlated pairs of genes are set to zero, as we are only interested in correlated sets of genes. Anti-correlated genes have a conflicting interpretation in terms of pathway activation, since we corrected all genes such that 'high' expression values represent pathway activation as described above.

Concerning the guiding data G there is no need to calculate all pairwise distances between genes since we are only interested in whether a gene is responding to the pathway activation or not. Pathway responding genes are supposed to show a difference in expression between the group samples subject to perturbation and the control group. Like with the pairwise gene distances in T we measure pathway response by the correlation between a gene g and the label vector Y:

$$d(g,Y) = 1 - \max\left(\varrho(g,Y), 0\right) \tag{4.2}$$

where $\varrho(g, Y)$ is the Pearson correlation between the expression vector of gene g and the label vector Y. Here Pearson's correlation coefficient is used since we want to preserve the influence of the within class variance on the distance.

Together both distance functions implement the two objectives we are aiming at: (i) Genes that are coherently expressed across the tumor samples will have low pairwise distances in within T. (ii) Genes that show response to pathway activation have low distances towards the label vector Y. We need to develop an approach that select genes based on those two conditions simultaneously.

Based on the fact that coherently expressed genes in T are situated in close proximity of each other, one can rank genes according to their proximity to neighboring genes using kernel density estimation (KDE). An introduction into KDE can be found in section 2.1.2. We follow this approach by applying a Gaussian function to the pairwise distances within T. The result is a matrix A_T of pairwise gene similarities. For two genes g and h A_T is defined as

$$A_T(g, h) = \exp\left(\frac{-d(g, h)^2}{2\sigma^2}\right) \tag{4.3}$$

where σ is a smoothing parameter that allows adjustment of the smoothing bandwidth as shown in Figure 4.1(a). Parameter calibration will be discussed in section 4.2.4. High row sums of A_T indicate that the corresponding gene is situated in a dense area or gene cluster.

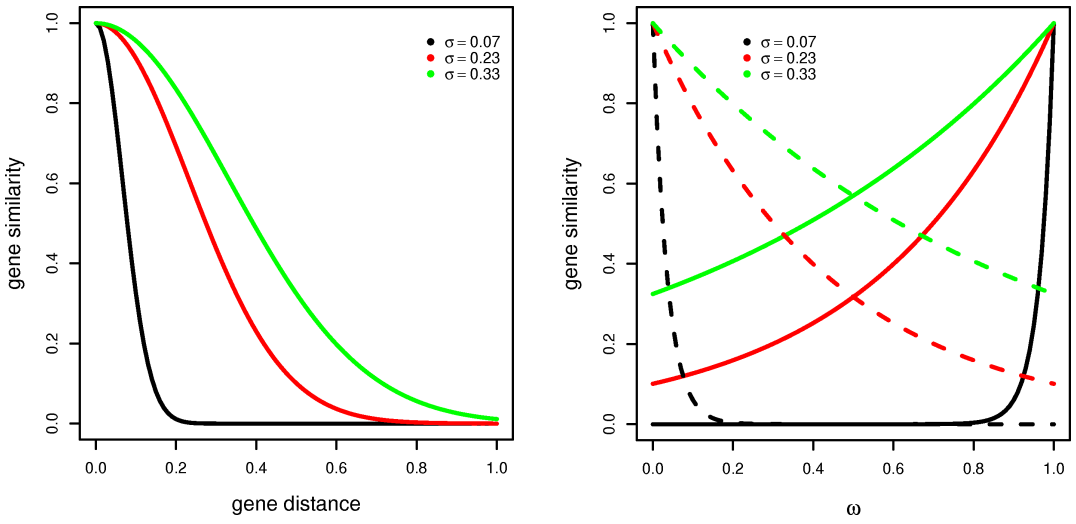
To integrate gene response to pathway activation in the density estimation process we transform the distances between gene expression across G and the label vector Y similar to the pairwise distances of T. We define a diagonal similarity matrix A_G as

$$A_G(g, g) = \exp\left(\frac{-d(g, Y)^2}{2\sigma^2}\right) \tag{4.4}$$

where all entries $A_G(g, h)$ with $g \neq h$ are set to zero. High values of $A_G(g, g)$ correspond to genes that respond to pathway activation, while genes that do not respond strongly approach zero. Again the parameter σ specifies the bandwidth of the smoothing function.

The integration step is done by reweighting A_T with A_G using matrix multiplication. We call this procedure matrix fusion:

$$W = A_G^{1/2} A_T A_G^{1/2} \,, \tag{4.5}$$

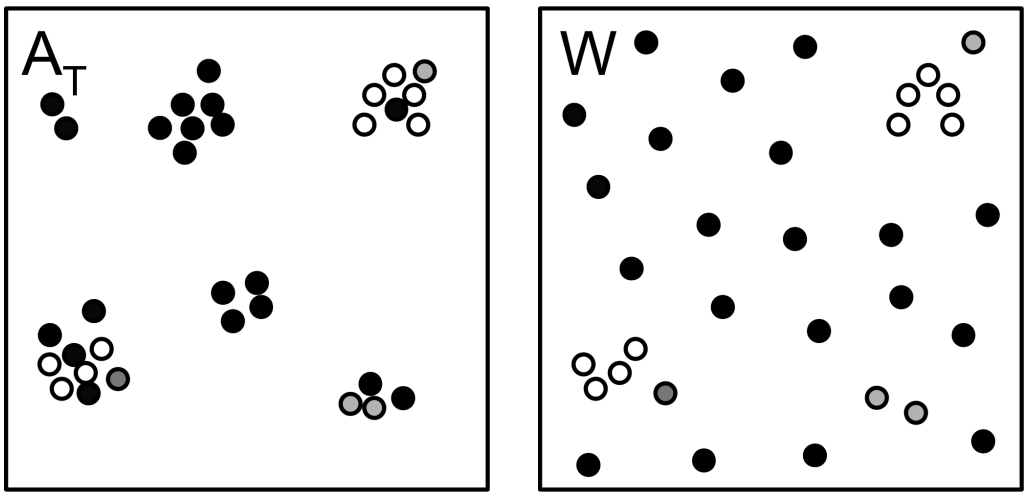


(a) (b)

Figure 4.1: Gaussian smoothing functions transform gene distances into gene similarities: **(a)** Parameter σ influences the smoothing bandwidth or slope of the Gaussian function. **(b)** The two data sets T and G are weighted against each other by the parameter ω. The graphs show the resulting gene similarity depending on ω for different choices of σ for the gene distance of 0.5. Solid lines show similarity values of A_T and dashed lines values of A_G.

where left and right hand side multiplication ensures symmetry of the resulting matrix W that holds high values only for pairs of genes that show consistent expression in T and simultaneously a common response to pathway activation in G.

Figure 4.2 schematically explains the effect of matrix fusion. The points in the left panel show a set of genes embedded in the 2-dimensional plane. Their distance reflects similarity according to A_T. Coherently expressed genes are situated closer to each other. The gray tone encodes information from the guiding data. Black points do not show differential expression in G while gray points are targets, and the brighter the point, the stronger the gene responds to pathway perturbation. Note that all genes fall into clusters. This is typical for expression data. Barely any one gene is regulated independently from any other genes. The right panel shows the same genes again. However this time, the distances are based on the fused similarity matrix. The dark points moved out of the clusters and distribute uniformly. The only remaining dense areas consist of bright genes that were already close to each other in the left panel. The effect of matrix fusion can be viewed as a



(a) (b)

Figure 4.2: Matrix fusion induces a magnetic repulsion between genes: Each point represents one gene. The distance between points reflects the similarity of genes while gray shades represent the genes response to pathway perturbation. The brighter a point the stronger the gene responds to the perturbation. **(a)** Gene similarities based on coexpression in the tumor data only. All genes fall in clusters, since genes are regulated in concert. **(b)** The same genes as in (a), but distances are based on the fused similarity matrix W. Genes that do not respond to the pathway moved out of the clusters and distribute uniformly across the plane.

magnetic repulsion between genes. The less a gene responds to pathway activation, the stronger is its repulsion from all other genes. Genes remaining in clusters are potential pathway target genes that are consistently regulated in tumors. We will use their consensus expression as a surrogate for pathway activity in tumors. Figure 4.2 shows that the matrix fusion induces dissimilarities that makes clustering of genes a hard problem, since many genes are no longer in clusters but on their own.

Most available cluster methods are problematic in situation where data includes large numbers of scattered points, as they are aiming to assign all data points to clusters. Applying standard algorithms to such data usually results in clusters that are skewed or misleading. Tseng and Wong (2005) suggested a resampling based approach that partially overcomes this problem, but we will not follow this approach here. Instead we aim to detect the top most dense modules of genes, thus leaving the majority of genes unassigned to any cluster. A simple procedure to extract dense sets of genes is described in the next section.

4.2.2 Extraction of tight expression modules

Given the fused similarity matrix W and following the KDE approach, we define the neighborhood density $K(g)$ for each gene as the row sum of W

$$K(g) = \sum_{i=1}^{p} W_{g,i} \tag{4.6}$$

where p is the total number of genes in the data set. A gene g with a high value $K(g)$ is located in a large and dense cluster of genes.

The module extraction procedure of *guided clustering* starts by selecting the gene g_0 that maximizes $K(g)$ as a seed gene. Next a module of genes C is grown around g_0 using average linkage by iteratively adding genes g_k that maximize

$$\gamma(g_0, g_1, \ldots, g_{k-1}, g_k) = \frac{\sum_{i,j \leq k} W_{g_i,g_j}}{|C| + 1} \tag{4.7}$$

where $|C|$ is the number of genes in C. The iteration is terminated, if no gene g_k exists, such that $\gamma(g_0, g_1, \ldots, g_{k-1}, g_k) > \gamma(g_0, g_1, \ldots, g_{k-1})$. In case we want to extract more than one dense cluster, we remove all genes selected in the current iteration, recompute $K(g)$ and proceed as described above.

4.2.3 Condensing the joint expression of genes in a module to a consensus expression index

Recall that our goal is to estimate pathway activation for each tumor sample of T. By construction, the expression levels of genes in a extracted module are tightly correlated across the tumor samples. Further, in any tumor they are either unanimously up or down regulated. This allows us to condense their expression into a single number per tumor. This number can be used as a surrogate for pathway activity in the tumor. Remember that all genes were adjusted such that numerically all targets display 'high' expression upon pathway activation. Hence a high index points to an active pathway and a low index to an inactive pathway. Note that although genes in a module correlate strongly their expression values can deviate highly due to scale differences. A module can be composed of genes, where some have a greater expression level than others. Simple summing or averaging of all module genes per sample weights genes unequally according to their variance. Instead, we use a standard additive model that accounts for scale differences on the

log scale to compute the consensus index:

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \tag{4.8}$$

where y_{ij} is the observed expression of a gene, α_i a gene specific scale coefficient, β_j the sample specific index of pathway activation and ϵ_{ij} the error term. We fit the additive model using Tuckey's robust median polish procedure (Hoaglin *et al.*, 1977). In the following we will refer to the index as pathway activation index (PAI). Note that our approach for summarizing the expression of module genes to a PAI is identical to a method used to compute probeset summaries in the popular normalization package RMA for Affymetrix GeneChips.

4.2.4 Balancing both data sets & parameter selection

In the last section we described a method that identifies modules of genes that are coherently expressed in T and have joint differential expression in G. Further we summarized the gene modules to a single index that estimates response to perturbation in the tumor samples. We do not weight the influence of the two data sets T and G against each other. Such a weighting is necessary since the values of T and G might be on different scales.

The weighting is achieved by introducing a new parameter ω to the equations 4.3 and 4.4. We extend equation 4.3 and 4.4 by:

$$A_T(g,h) = \exp\left(\frac{-(1-\omega)d(g,h)^2}{2\sigma^2}\right) \tag{4.9}$$

and

$$A_G(g,g) = \exp\left(\frac{-\omega d(g,Y)^2}{2\sigma^2}\right) \tag{4.10}$$

where ω is chosen between 0 and 1 and therefore shifts the focus from the clinical data ($\omega = 0$) to the guiding data ($\omega = 1$). By setting $\omega = 0$ the nominator in equation 4.10 equals zero. Hence A_G becomes the identity matrix as the exponential of zero equals one. If we plug in the identity matrix for A_G in equation 4.5, $W = A_T$ since matrix multiplication with the identity matrix is self-preserving. Likewise setting $\omega = 1$ results in $A_T(g,h) = 1$ for all pairs of genes g and h. Thus all entries of the fused similarity matrix $W(g,h)$ equal $\sqrt{A_G(g,g) * A_G(h,h)}$, eliminating the influence of A_T on the identification of gene modules. Setting ω to any value between zero and one balances the influence between both data sets. This balancing is illustrated by

Figure 4.1(b). Moving the focus from T to G by increasing ω lifts the resulting pairwise gene similarities on T and decreases the resulting per gene similarities on G.

To select a optimal ω we need to evaluate the influence of both data sets by some measure. While the tumor data T is driving within cluster similarity, the guiding data G forces genes to have a high response to perturbation. For a gene module C_ω retrieved for a specific ω, we evaluate the within cluster similarity by $\phi(\omega)$, the average pairwise correlation between genes of C_ω:

$$\phi(\omega) = 1/|C_\omega|^2 \sum_{g,h \in C_\omega} \max(\rho(g,h), 0) \ , \tag{4.11}$$

where $|C_\omega|$ is the number of genes in C_ω and $\rho(g,h)$ is Spearmans correlation of any two genes g and h. The within cluster response to perturbation of a gene module C_ω is assessed by the average gene activation $\varphi(\omega)$ defined as:

$$\varphi(\omega) = 1/|C_\omega| \sum_{g \in C_\omega} \varrho(g, Y) \ , \tag{4.12}$$

where ϱ is Pearsons correlation of gene g and the label vector Y.

To select the optimal ω during gene module extraction, we generate a set of module candidates for different choices of ω. The parameter ω is varied from 0 to 1 in steps of 0.1. The best choice features a high within module correlation and a high average response to perturbation. Hence we choose ω by maximizing the sum of $\phi(\omega) + \varphi(\omega)$. Since $\phi(\omega)$ and $\varphi(\omega)$ are on different scales we rescale them before summation, such that they both range form 0 to 1. This means ,if ω is set to zero $\phi(\omega) = 1$ and $\varphi(\omega) = 0$, since all weight is on T. In contrast if ω is set to one $\phi(\omega) = 0$ and $\varphi(\omega) = 1$, since we focus only on G.

The parameter σ specifies the bandwidth of the smoothing kernel and influences the global sensitivity of the method. Larger bandwidths generate higher gene similarities as shown in Figure 4.1(b). This results in larger clusters that may include genes with low responses to perturbation. Smaller bandwidth enforce more rigorous restrictions on the genes from the guiding data. We recommend to tune σ manually starting from a large value and decreasing it in several steps while at the same time monitoring the cluster tightness and the distribution of perturbation responses in the guiding data. For our analysis we varied σ between $1/3 \approx 0.33$ and $0.1/3 \approx 0.03$ in steps of $0.1/3 \approx 0.03$. An example for parameter tuning will be given in section 4.3.2.

4.2.5 Extensions to other experimental settings

So far, we have discussed *guided clustering* in the context of a gene expression based perturbation experiment. Here the labeled guiding data set G consists of two classes of gene expression profiles: (i) The perturbed class, measuring the per gene response of a perturbation like gene silencing by RNAi or application of a compound and (ii) a corresponding control. *Guided clustering* can be easily adapted to other application scenarios such as protein binding or interaction assays, DNA methylation studies or genome aberration studies via array CGH. Basically it works with any method that gives a quantitative measurement per gene. The adoption is made by tailoring the similarity function in equation 4.10 to the application. The similarity values need to quantitatively rank genes that should be preferentially used to build gene clusters. The strongest preference possible is encoded by a value of 1. Smaller values gradually reduce the influence of a gene. The preference scores need to be calculated from guiding data. For example the preference score can reflect the connectivity of a gene in a protein-protein interaction network, thus guiding the formation of gene clusters seeded around hub genes. They can also reflect the binding abundance of a transcription factor assessed in a chromatin immunoprecipitation experiment thus guiding the gene clusters to be build from targets of a specific transcription factor. We will demonstrate the use of guided clustering in this context in section 4.4.1.

4.2.6 Runtime

To analyze the runtime of the *guided clustering* algorithm we dissect the algorithm into its three main parts: (i) The calculation of the pairwise gene distance matrix has a runtime of $O(n^2)$, where n is the number of genes. (ii) Transformation of the pairwise gene distances into affinities needs additional $n_\omega\, O(n^2)$ operations, were n_ω is the number of values used for ω when choosing the optimal weighting between both datasets. (iii) Extracting k gene modules has a complexity of $k\,O(n)$. In practice, the total complexity $O(n^2) + n_\omega\, O(n^2) + k\,O(n)$, is dominated by the square number of input genes n^2.

All calculations have been performed on a machine containing 16 Quad-Core AMD Opteron 8354 processors with 2.2 GHz each and 132 GB main memory. At the current state of development *guided clustering* is a single thread method using 1 of 16 processors available. For the analysis of simulated data sets used in section

4.3 each run needed 13 sec on average. Analysis of the lymphoma samples together with BCL6 and LPS data in section 4.4.1 and 4.4.2 took about 890 and 902 sec, respectively.

4.3 Simulations

Prior to testing *guided clustering* in real data integration contexts, we study its performance on data that is artificially generated and fulfills the underlying assumptions of our algorithm. This allows us to better understand its limitations alongside those of competing strategies. In simulations, the data generating process defines a ground truth, against which any analysis result can be evaluated. In real applications we often do not have a ground truth result. Moreover, focused simulations allow us to study individual difficulties in the analysis independently from each other, while real data usually comprises many of them in parallel. Finally, the difficulty of clustering problems can be scaled freely in simulations. In this section we compare *guided clustering* to the two competing sequential analysis concepts described in the literature which select genes sets only using the clinical (Läuter *et al.*, 2009) or the guiding data respectively (Bild *et al.*, 2006).

4.3.1 Simulation model

We simulate artificial data that mimics the application of *guided clustering* in the context of pathway activation prediction via guiding by perturbation experiments. The data consist of an artificial clinical data set T with 80 samples and a guiding data set G with 20 control and 20 perturbed samples. The data sets comprise 1500 features. Both data sets are generated by adding a signal component F_{ij} or I_{ij} and a noise component ϵ_{ij}:

$$T_{ij} = F_{ij} + \omega_T \, \epsilon_{ij} \tag{4.13}$$

and

$$G_{ij} = I_{ij} + \omega_G \, \epsilon_{ij} \tag{4.14}$$

The components F_{ij} and I_{ij} contain the ground truth, since they simulate the target signals, while the noise components ϵ_{ij} simulate technical measurement fluctuations and biological variability not related to the target signal. The tuning parameters

ω_T and ω_G are used to calibrate the signal-to-noise ratio. The noise component is simulated for both data sets using a multivariate normal distribution with a block structured covariance matrix Σ as proposed by Guo *et al.* (2007):

$$\Sigma = \begin{pmatrix} \Sigma_k & 0 & 0 & 0 & 0 \\ 0 & \Sigma_{-k} & 0 & 0 & 0 \\ 0 & 0 & \Sigma_k & 0 & 0 \\ 0 & 0 & 0 & \Sigma_{-k} & 0 \\ 0 & 0 & 0 & 0 & \Sigma_k \end{pmatrix}_{1000\times1000} \tag{4.15}$$

with

$$\Sigma_k = \begin{pmatrix} 1 & k & \dots & k^{98} & k^{99} \\ k & 1 & \ddots & \ddots & k^{98} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ k^{98} & \ddots & \ddots & \ddots & k \\ k^{99} & k^{98} & \dots & k & 1 \end{pmatrix}_{100\times100} \tag{4.16}$$

where k reflects the amount of co-regulation, similar to Guo *et al.* (2007) we set $k = 0.9$. Σ is a square matrix with the number of genes as dimensions and has a block structure. Each block has the dimensions 100×100 and represents one set of co-regulated genes.

For T_{ij} we generate signals in 3 clusters $E_1, \dots, E_3$ of 200 features each representing different biological activities across the samples. In analogy to the additive model we generate traces of pathway activity by drawing for each gene in a cluster a random number α_i uniformly from the interval $[0, 1]$. This number represents the strength with which a gene responds to pathway activation. Moreover, for every sample we draw a uniformly distributed random number β_j from the interval $[-1, 1]$, which represents the strength of the pathway activation in this sample. F_{ij} is then set to $\alpha_i + \beta_j$. Note that since clusters mimic different biological activities, β_j is constant throughout genes from the same cluster but not for genes from different clusters. For genes that do not fall in any of the three clusters F_{ij} is set to zero.

The simulation of the guiding data G_{ij} includes a set B_d of 600 *responding* genes. These genes are simulated differently for control and perturbation samples. For each of them we draw a random number γ_i uniformly from $[0, 1]$ and set $I_{ij} = -\gamma$ for control samples and $I_{ij} = \gamma$ for perturbation samples. For the remaining genes, we set I_{ij} to zero. The size of the intersection of the three clusters E_i with the set of responding genes varies across clusters (200 for E_1, 100 for E_2, and 50 for E_3).

Signal-to-noise ratios are not constant over genes but we can use the tuning parameters ω_T and ω_G to calibrate the max signal-to-noise ratios

$$R = \max_i \left(\frac{\text{rmsd}(F_{i,\cdot})}{\text{rmsd}(\epsilon_{i,\cdot})} \right) \tag{4.17}$$

where the maximum is taken over all genes and the root mean square distance (rmsd) of a gene expression vector is defined as

$$\text{rmsd}(x) = \sqrt{\frac{1}{n} \sum_{j=i}^{n} (x_j - \bar{x})^2} \tag{4.18}$$

where n is the number of samples in T. In our simulations we varied R for T between 2 and 0.5 in steps of 0.1. In the guiding data G the maximal signal to noise ratio was kept constant at 2. For each signal-to-noise setting 20 data sets were generated.

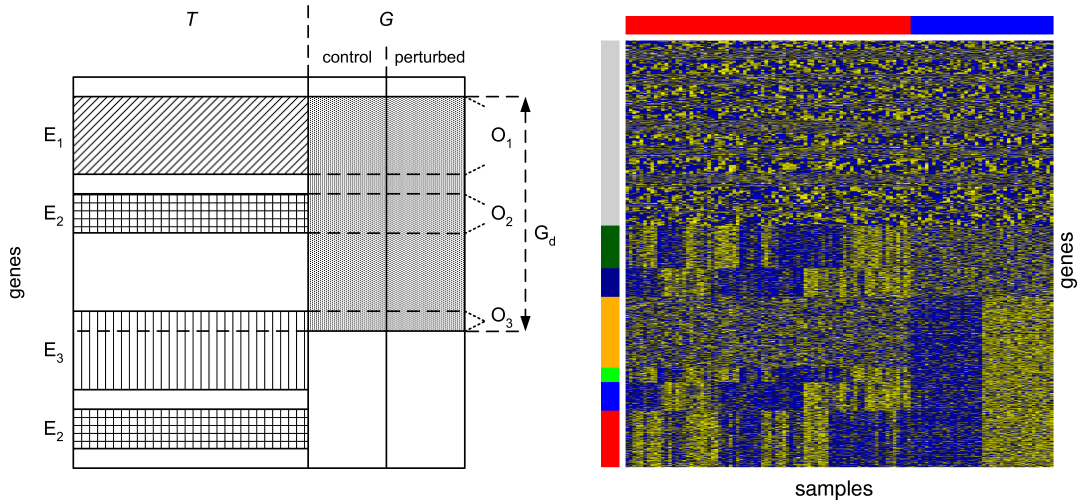
The simulation setup is summarized in Figure 4.3(a) and an exemplary data set with signal-to-noise ratio of 1 is shown in Figure 4.3(b). Note that we not only simulate signals and noise but also confounding structures including clusters of correlated genes in T that do not correspond to induced genes in G as well as induced genes in G that do not form tight clusters in T.

4.3.2 Simulation analysis

In this study we want to assess the performance of *guided clustering* reconstructing the hidden pathway activation signals β_j for all samples and all three clusters. We compare *guided clustering* to two possible sequential analysis concepts, namely gene selection based on experimental data followed by pathway activity prediction on the clinical data as described by Bild *et al.* (2006), and gene selection via identification of strongly correlated gene sets in the clinical data followed by multivariate tests in the experimental data as described by Läuter *et al.* (2009).

We ran *guided clustering* on a series of simulated data sets with increasing difficulty. To evaluate accuracy of the estimated signals, we calculated the maximal correlation of the top 3 estimated pathway indices $\hat{\beta}_j$ to any of the β_j underlying the simulation. For each signal-to-noise setting we averaged the accuracy across the 20 simulated data sets.

The whole simulation study includes 320 independent runs of *guided clustering*. As an example we give a detailed description of analysis results obtained for a



(a) (b)

Figure 4.3: **(a)** Structure of the simulated data. The three clusters with non-zero signals F_{ij} in T are named E_1, E_2 and E_3. Each effect overlaps with a certain number of differentially expressed genes in $B_d \subset G_d$. The overlaps are named O_1, O_2 and O_3. **(b)** Example of simulated data with signal-to-noise ratio set to 1. Different simulated effects are indicated by left side colorbar: red - O_1, blue - O_2, green - O_3, orange - $G_d \notin O_1 \cup O_2 \cup O_3$, dark blue - $E_2 \notin O_2$, dark green - $E_3 \notin O_3$. Tumor samples (T) and guiding data samples (G) are indicated by a colorbar on top: red - T and blue - G.

signal-to-noise ratio of 1. The analysis starts with calibrating the parameter σ, such that the resulting modules are highly enriched for genes with strong perturbation responses within the guiding data. We begin with a large value of $\sigma = 0.33$ and consecutively lower it while monitoring the distribution of signals from the guiding data. Figure 4.4 illustrates the selection process, with $\sigma = 0.23$ a good balance was found. All extracted modules show strong responses to the perturbation. We kept $\sigma = 0.23$ constant for all other analyzes.

Both sequential analysis concepts were evaluated similar to *guided clustering*. For the approach described by Bild *et al.* (2006) we employed the *limma* package to rank the genes according to their differential expression within the guiding data. The top 100 differentially expressed genes were used to estimate the perturbation response of the tumor samples by fitting the additive model described in section 4.2.3. Correlation with the simulated effects was calculated as described above. The concept described by Läuter *et al.* (2009) works in the opposite direction by

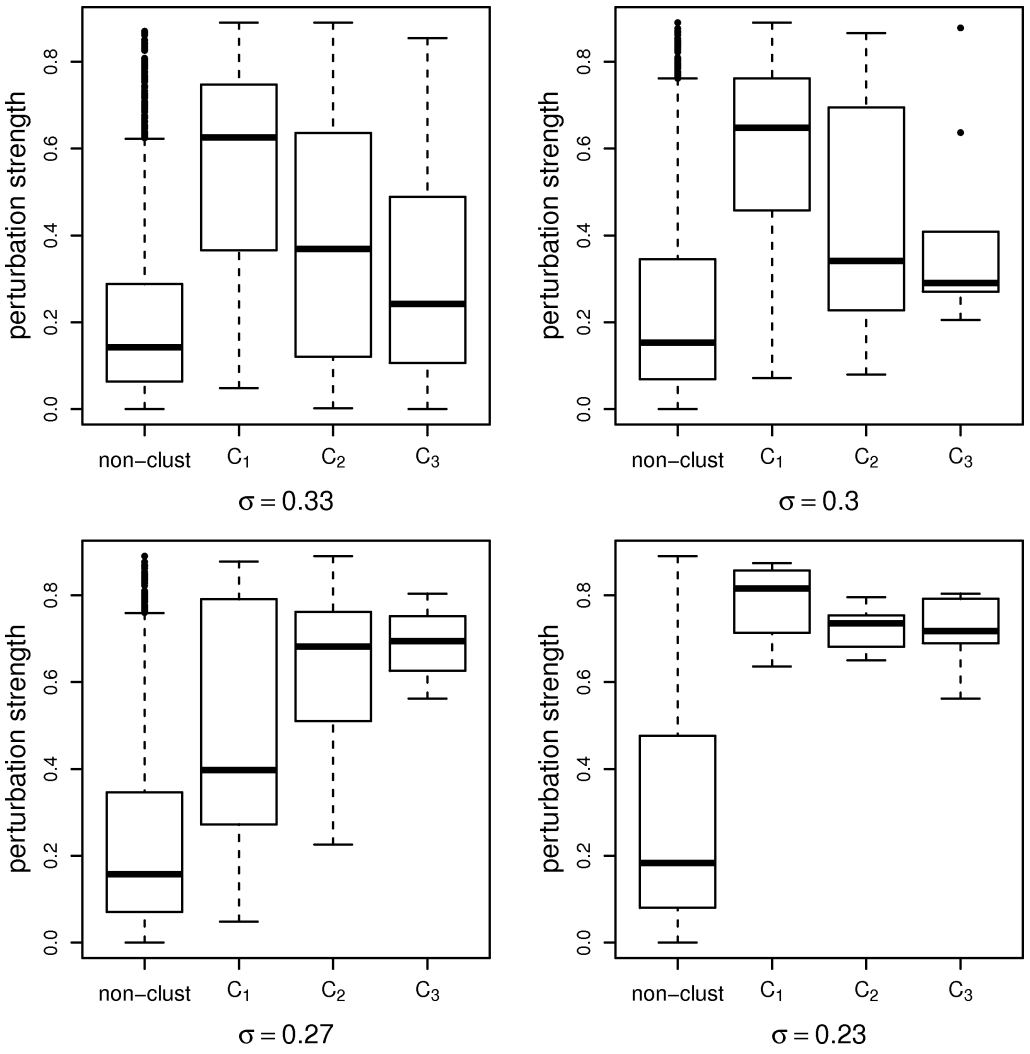
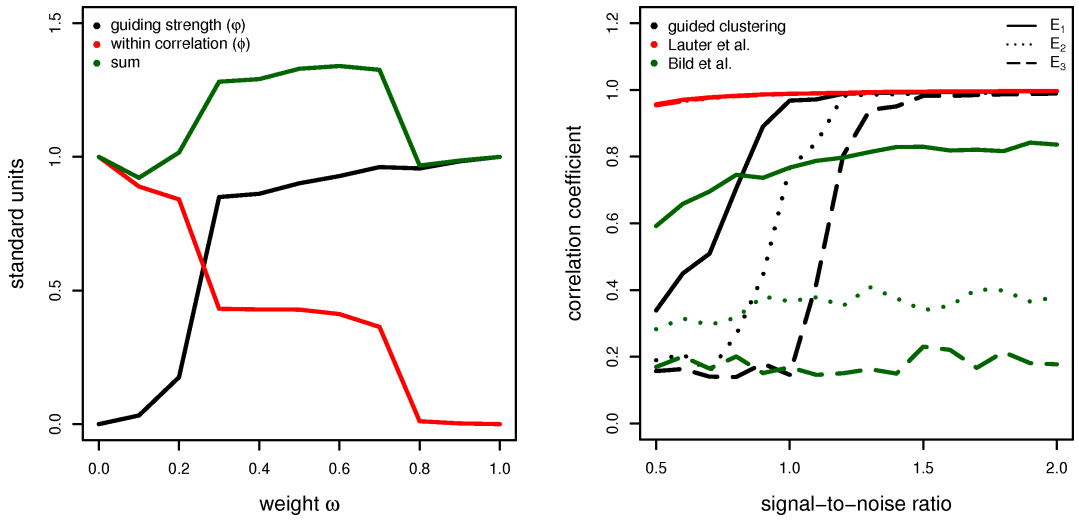


Figure 4.4: Guiding strength of top three clusters extracted with different choices of σ.

selecting clusters of genes highly correlated across T. Those gene clusters are then tested for joint differential expression in G. We assume that we have an optimal cluster algorithm at hand and use the correct simulated gene clusters to estimate the perturbation response in the tumor samples by fitting the additive model. The gene clusters are tested for joint differential expression in G using the *geneSetTest* function of the *limma* package. This function tests whether a set of genes is highly ranked relative to other genes in terms of a given statistic. We used the t-statistic supplied by the differential gene expression analysis of *limma*.

Figure 4.5(b) shows the performance of *guided clustering* and both sequential approaches on estimating the perturbation response within the tumor samples of



(a) (b)

Figure 4.5: **(a)** Trade-off between within cluster correlation and guiding strength depending on ω. **(b)** Accuracy of estimated effects depending on the signal-to-noise ratio with $\sigma \approx 0.23$: black - guided clustering; red - Läuter et al.; green - Bild et al.; E_1, E_2, E_3 simulated effect with $200, 100, 50$ genes overlapping between induced perturbation in T and differential expression in G.

the three simulated effects (E_1 solid, E_2 dotted, E_3 dashed line). *Guided clustering* reconstructs the pathway indices correctly for signal-to-noise ratios ≥ 1. However, the smaller the overlap between the correlated clusters in T and the differentially expressed genes in G the more difficult the reconstruction. The approach by Läuter *et al.* (2009) reconstructs the signals perfectly since we provided it with the correct gene sets. Using the approach of Bild *et al.* (2006) results in a poor reconstruction quality. The method is only able to coarsely reconstruct the cluster with the biggest overlap. We believe that the ignorance of the correlation structure in T that this method exercises when constructing gene sets compromises its performance in signal reconstruction.

To determine whether gene modules showed a significant grade of perturbation response we employed the *geneSetTest* function of the *limma* package as described above. We tested only modules identified by *guided clustering* or the approach of Läuter *et al.* (2009), since gene modules identified by the approach of Bild *et al.* (2006) are significant *per se* as they involve only the 100 top ranking genes. Table 4.1 shows how often a gene module was considered to be statistically significant given

Table 4.1: Percentage of simulation runs in which the identified gene sets were rated as significantly different within the guiding data G for varying signal-to-noise ratios. The significance threshold was set to 0.05 (0.01).

	guided clustering			tumor clustering		
snr	cluster 1	cluster 2	cluster 3	cluster 1	cluster 2	cluster 3
0.5	1.00 (1.00)	1.00 (0.95)	1.00 (1.00)	1 (1)	0.85 (0.40)	0 (0)
0.6	0.95 (0.95)	0.95 (0.85)	1.00 (1.00)	1 (1)	0.75 (0.25)	0 (0)
0.7	0.95 (0.80)	0.95 (0.80)	0.95 (0.90)	1 (1)	0.80 (0.30)	0 (0)
0.8	1.00 (1.00)	1.00 (0.95)	0.90 (0.85)	1 (1)	0.65 (0.45)	0 (0)
0.9	1.00 (0.90)	1.00 (1.00)	1.00 (0.95)	1 (1)	0.70 (0.30)	0 (0)
1.0	1.00 (0.95)	0.95 (0.80)	1.00 (0.80)	1 (1)	0.75 (0.40)	0 (0)
1.1	1.00 (0.95)	1.00 (1.00)	0.95 (0.95)	1 (1)	0.85 (0.35)	0 (0)
1.2	1.00 (1.00)	1.00 (0.95)	1.00 (0.95)	1 (1)	0.65 (0.40)	0 (0)
1.3	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1 (1)	0.85 (0.50)	0 (0)
1.4	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1 (1)	0.60 (0.45)	0 (0)
1.5	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1 (1)	0.85 (0.50)	0 (0)
1.6	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1 (1)	0.70 (0.35)	0 (0)
1.7	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1 (1)	0.80 (0.30)	0 (0)
1.8	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1 (1)	0.85 (0.45)	0 (0)
1.9	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1 (1)	0.55 (0.20)	0 (0)
2.0	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1 (1)	0.75 (0.35)	0 (0)

a threshold of 0.05 or 0.01, respectively. *Guided clustering* reliably detects gene sets with differential expression in G for both thresholds. For low signal-to-noise ratios the performance is slightly worse. In comparison, the approach of Läuter *et al.* (2009) identifies perturbation response only in gene modules were the majority of contributing genes are differentially expressed (cluster 1). Modules were only half of the genes (cluster 2) or less (cluster 3) are differentially expressed within the guiding data were detected only sometimes or never.

In summary, guided clustering cannot outperform the sequential methods both with respect to cluster tightness and enrichment of genes with strong signals from

the guiding data. However, conversely it is the only method that can achieve good performance in both aspects simultaneously.

4.4 Applications

The *guided clustering* procedure identifies gene modules that respond strongly to the perturbation present in the experimental data while at the same time display coherent expression in clinical expression profiles. This general setting allows the application of *guided clustering* in many different situations. In this thesis we present two of them: (i) The detection of functional targets of a transcription factor and the quantitative estimation of its activity in individual clinical samples, and (ii) the estimation of pathway activities in tumor samples as introduced by Bild *et al.* (2006) using a cell line stimulation experiment. All lab work concerning the stimulation experiment was done by our cooperation partners from the laboratory of Dieter Kube from the university clinic for hematology in Göttingen.

4.4.1 Identification of BCL6 target modules in diffuse large B-cell lymphomas

BCL6 is a key oncogene in B-cell lymphomas The transcription factor BCL6 acts predominantly as a repressor of transcription. In B-cells, BCL6 is required for the formation of germinal centers (GC). Hence its function is critical for the working of the acquired immune system. It was shown by Dent *et al.* (1997) and Ye *et al.* (1997) that BCL6-null mice lack the formation of GC and are unable to produce high-affinity antibodies. In diffuse large B-cell lymphomas (DLBCL) BCL6 is frequently translocated, hyperpermutated or hypermethylated. In fact, the BCL6 gene was identified in 1993 as the target of chromosomal translocations in DLBCL by Baron *et al.* (1993), Ye *et al.* (1993a) and Ye *et al.* (1993b). Its oncogenic potential was shown by Cattoretti *et al.* (2005) in mouse models where deregulated BLC6 expression lead to the development of lymphomas. It is assumed that this dysfunctional activity contributes to oncogenesis in a subset of DLBCL and potentially also in different malignancies (Iqbal *et al.*, 2007). DLBCL is a morphologically, genetically and clinically heterogeneous lymphoma entity (IARC, 2008) with several defined subentities (Rosenwald *et al.* (2002), Bentink *et al.* (2008)). Whether

differential BCL6 activity contributes to the heterogeneity is not fully known.

The influence of BCL6 on the transcriptional program of B-cells is accomplished via BCL6 targeted genes. The identification of such target genes is essential for understanding the underlying mechanisms of transcriptional regulation. Initially, BCL6 target discovery was based on educated guesses (Niu *et al.*, 2003) and gene expression profiling (Shaffer *et al.*, 2000). Recently, Polo *et al.* (2007), Ci *et al.* (2009) and Basso *et al.* (2010) identified large sets of genes whose promotor regions are bound by BCL6 in vivo. Ci *et al.* (2009) generated a data set by performing a ChIP-on-chip screen with primary human GC B-cells and the DLBCL cell lines OCI-Ly1 and OCI-Ly7 using high density oligonucleotide promotor microarrays. This data was combined with gene expression profiles from DLBCL. The authors followed a strictly sequential analysis strategy: (i) Only the ChIP data was used to identify distinct groups of genes bound by BCL6 exclusively in GC B-cells, DLBCL, or both. (ii) A look up of these genes in a tumor data set revealed that a large number of BCL6 target genes are silenced in primary human centroblasts compared to naive B-cells. But only half of those genes are also silenced in DLBCL. Among these non-silenced genes are critical mediators of survival, growth, proliferation and B-cell differentiation. Hence Ci *et al.* (2009) reason that the transcriptional programming of BCL6 is disturbed in DLBCL. Further, BCL6 target genes have a lower expression in GC B-cell-like (GCB) lymphomas compared with activated B-cell-like (ABC) lymphomas. In their analysis the authors considered DLBCL and its subentities GCB and ABC as lymphoma populations. They did not account for variability of BCL6 target expression across individual lymphomas. In fact, we observed that the expression of the top ranking genes in DLBCL is rather disperse (Figure 4.6(a)).

Genomewide chromatin immunoprecipitation assays identify binding sites of transcription factors in the neighborhood of genes and thus generate lists of potential targets of this transcription factor. Clearly, binding does not imply regulation. In general, several regulators and cofactors are needed for transcription and their presence depends on the cellular context. In the BCL6 context, various cofactors have been identified by Dhordain *et al.* (1997), Dhordain *et al.* (1998), Huynh and Bardwell (1998), Huynh *et al.* (2000), Lemercier *et al.* (2002) and others. However, if a transcription factor like BCL6 actively regulates a module of target genes in a defined cellular context like the DLBCL context, its functional targets should display correlated expression across clinical samples. The heterogeneity of DLBCL poses

an additional problem, since it is not clear whether there is only a single 'DLBCL context'. The expression and activation of transcriptional co-regulators might vary across DLBCL and we might find multiple modules of functional BCL6 targets, each expressed in different subsets of DLBCL.

Data In our analysis we used *guided clustering* to integrate the BCL6 ChIP-on-chip data of purified GC B-cells from Ci *et al.* (2009) with 220 expression profiles of DLBCL and Burkitt lymphoma samples from Hummel *et al.* (2006). Both data sets were taken from the Gene Expression Omnibus (Edgar *et al.*, 2002), GEO-accession: GSE15179 and GSE4475. Gene expression profiles were measured on the Affymetrix HG-U133A GeneChip^{TM}platform and the variance stabilization method proposed Huber *et al.* (2002) was used for normalization. All probe sets were summarized to gene expression values by fitting a standard additive model, employing Tuckey's median polish algorithm (Hoaglin *et al.*, 1977). The ChIP-on-chip data was available as normalized log2 ratios per gene locus representing the binding affinity fold change between sample and corresponding control. We averaged the log2 ratios across all three replicates after truncating ratios above the 95% quantile of all positive log2-ratios (≈ 2.67). Locus by locus we subtracted the log2 ratios from their maximum across samples and fed these values directly into the *guided clustering* algorithm. The BCL6 binding loci were matched to HG-U133A probe sets using the accession numbers and NCBI reference sequence (refseq) identifiers provided by Ci *et al.* (2009). Multiple probe sets for the same locus were summarized using the median polish algorithm. Each of the final data sets consisted of 9648 genes.

Result On this data sets, we ran *guided clustering* and identified the top three modules of BCL6 targets each expressed predominantly in a different subset of lymphomas. The smoothing parameter σ was selected as described in section 4.2.4 and set to $\sigma = 0.23$. The selection process is illustrated in Figure 4.7.

We extracted the top three gene modules. Their expression across the lymphoma samples is shown in Figure 4.6(e). Within each module the genes are coherently expressed and there is a continuous gradient when the samples are ordered by increasing BCL6 indices. All module genes exhibit strikingly large log2-ratios as shown in Figure 4.6(b), indicating that BCL6 indeed binds their promotor. Together both plots confirm that the two data sets are in good balance. In agreement to the observations of Ci *et al.* (2009) the second extracted index (BCL6-index2) is higher in

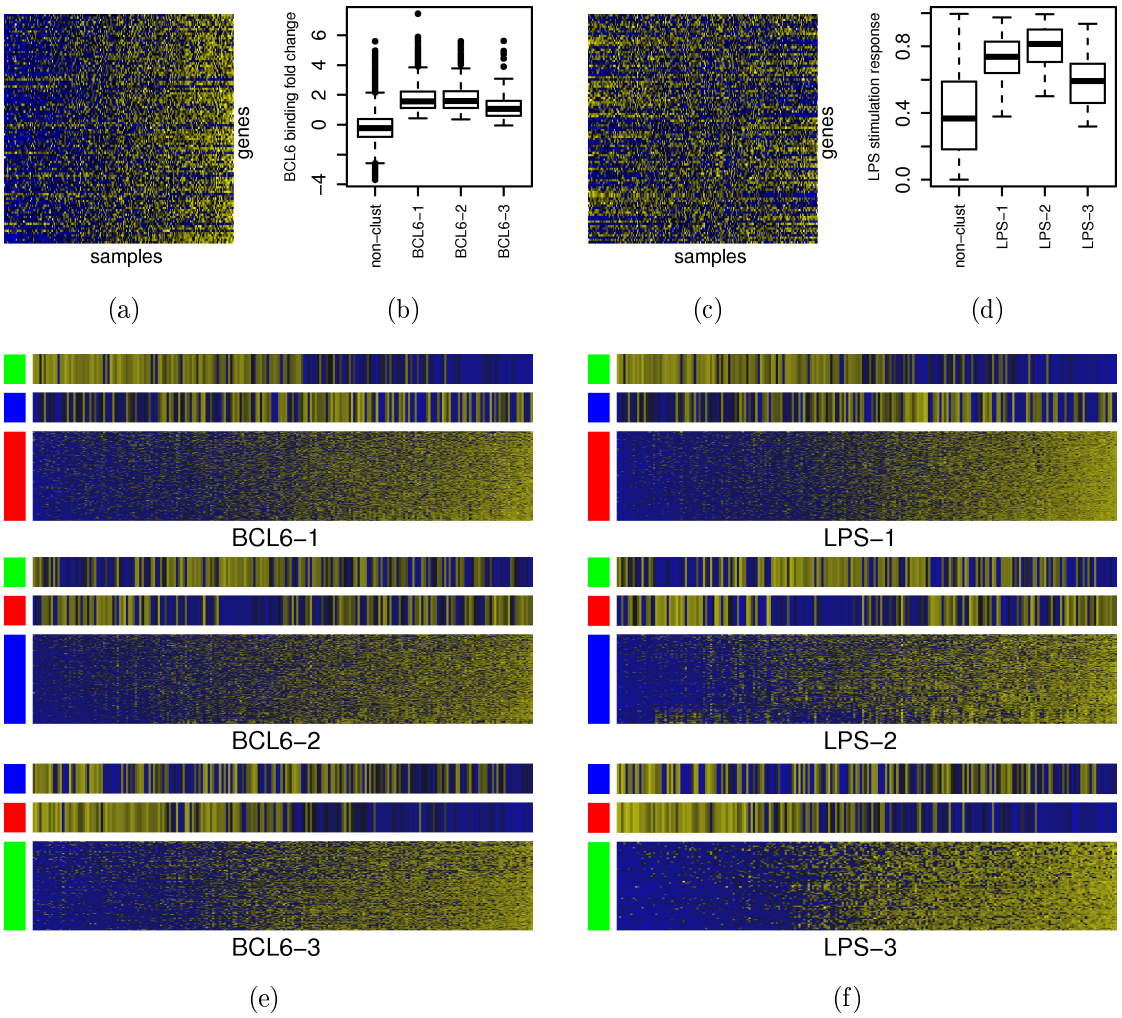
Table 4.2: Cox-regression analysis of DLBCL depending on the BCL6-index2, adjusted for activated B-cell (ABC) status, age and Ann Arbor staging.

factor	p-value		relative risk (95% confidence interval)
BCL6-index2	3.40e-6	***	0.0257 (0.0055 – 0.1205)
ABC-status	0.02298	*	2.2112 (1.1157 – 4.3823)
Ann Arbor stage	0.00173	**	3.1389 (1.5347 – 6.4203)
age	0.27981		1.4352 (0.7453 – 2.7634)

ABC than GCB type DLBCL ($P < 10^{-9}$, t-test).

To test for a clinical impact of BCL6 activity in DLBCL we fitted a cox proportional hazard model including the BCL6-index from module 2 as a continuous covariate together with the established categorical prognostic factors ABC/GCB status, age > 59 years, and Ann Arbor staging. Our survival analysis was restricted to a group of patients that received identical treatment, a combination of chemotherapy based on cyclophosphamide, doxorubicin, vincristine, and prednisone (CHOP) or similar. This was the case for 80 lymphoma patients in the study. We found the BCL6-index2 to be a significant independent predictor of survival ($P < 10^{-5}$). Patients with a high BCL6-index2 have a better outcome than patients with a low index. Notably, the hazard associated with the BCL6-index2 is higher than all other factors including the ABC/GCB status as shown in Table 4.2.

The BCL6-index2 accumulates the expression of 335 genes including several BCL6 targets that where also described in the primary analysis of the data by Ci *et al.* (2009) like the oncogenes BCL2A1, CCND1, CCND2, HSP90B1 and JUNB, as well as the transcription factors STAT1 and STAT3. A full list of all genes can be found in supplementary Table 6.1. We analyzed this gene set for enrichment of genes involved in certain aspects of B-cell functionality or malignant transformation using the Gene Set Analysis Toolkit V2 by Duncan *et al.* (2010). Genes involved in Toll-like receptor signaling were significantly enriched ($P < 0.004$, hypergeometric test). Similarly we found enrichment of Jak-STAT signaling genes ($P < 0.001$, hypergeometric test). These observations support the findings of Basso and Dalla-Favera (2010), who hypothesize that BCL6 modulates signaling through Toll-like receptors.



(a) (b) (c) (d)

(e) (f)

Figure 4.6: **(a, c)** Gene expression of top 100 genes with highest BCL6 binding fold change (a) or LPS activation (c), respectively. Yellow indicates high and blue low expression. The samples were ordered according to their mean expression. **(b, d)** BCL6 binding fold change (b) or LPS activation (d) of the extracted gene clusters compared to non cluster genes. **(e, f)** Gene expression of extracted PAI gene clusters across the lymphoma samples (1st - red, 2nd - blue, 3rd - green). Samples are ordered increasingly with respect to the PAI. On top of each gene cluster the two other PAIs are shown. Yellow indicates high and blue low expression.

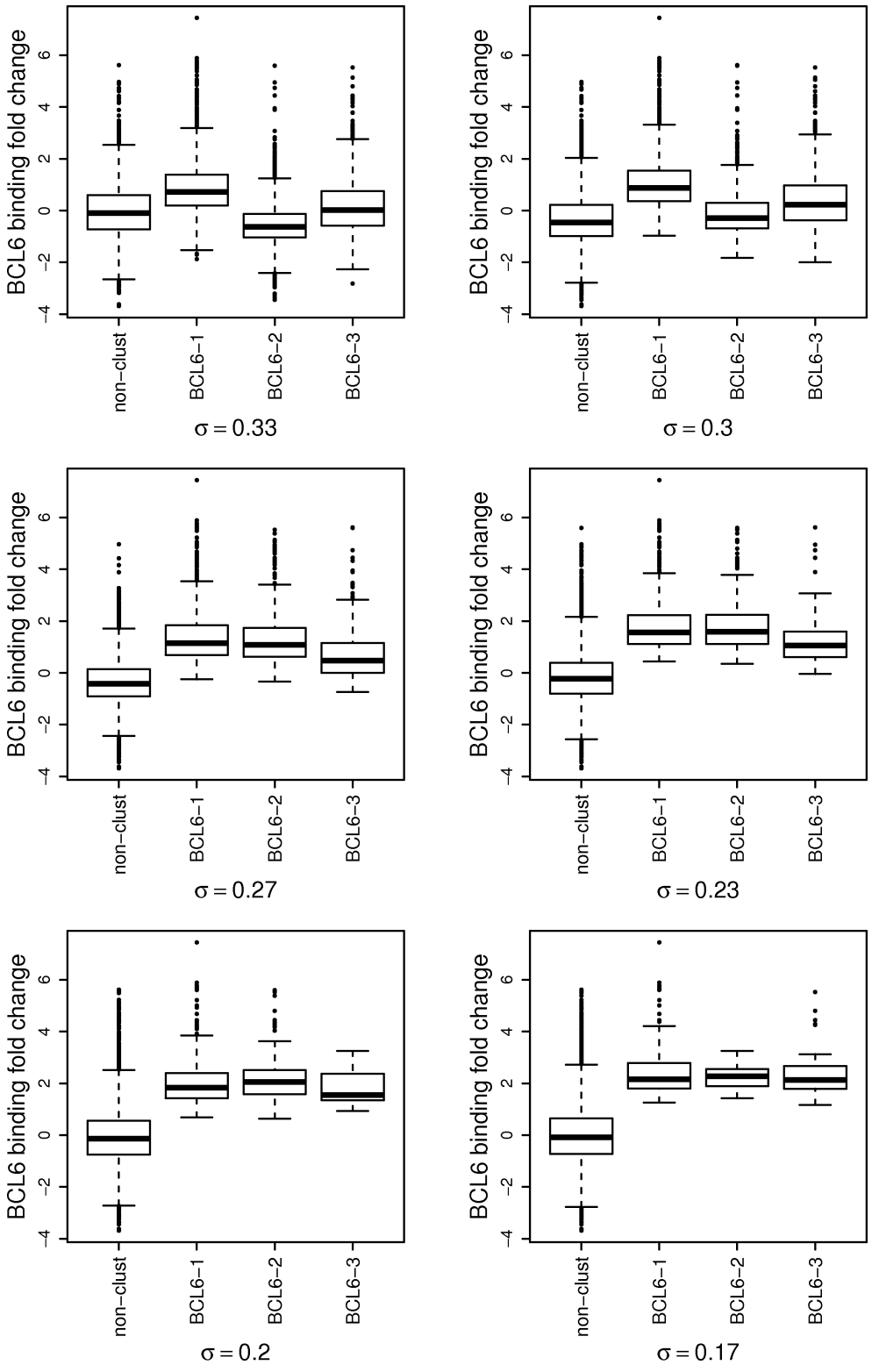


Figure 4.7: Guiding strength of top three BCL6 clusters extracted with different choices of σ. For $\sigma \geq 0.23$ the extracted clusters show outstanding high BCL6 fold changes.

4.4.2 LPS mediated Toll-like receptor signaling and BCL6 targets are coherently expressed in DLBCL

Data We follow-up our results from the previous section and support the hypothesis of Basso and Dalla-Favera (2010) by stimulating Toll-like receptors directly in the lymphoma cell line BL-2. The stimulation is achieved by treating the cells for 6 hours with lipopolysaccharide (LPS). Toll-like receptor mediated signal transduction of LPS stimulation has been shown by Chow *et al.* (1999) and Schwandner *et al.* (1999). In our experiment we compared expression profiles of stimulated cells with control profiles of unstimulated BL-2 cells. The expression profiles were generated on the Affymetrix HG-U133plus2.0^{TM}platform. Data normalization was performed as described in section 4.4.1. Only the subset of genes that was also represented within the lymphoma and BCL6 data sets was used for further analysis. Altogether 6 samples were hybridized, 3 independent biological replicates in each group. A detailed description of the sample preparation can be found in the supplementary material section 6.5.

Result We used *guided clustering* for a joint analysis of our stimulation data and the lymphoma data set by Hummel *et al.* (2006). This is a typical application of *guided clustering* in the context of integrating experimental cell perturbation data and clinical expression studies. In contrast to ChIP assays, cell perturbation experiments can identify functional targets of signaling pathways. However, they are not confined to direct targets. Transcriptional regulation is context specific and the molecular contexts of a cell culture significantly differs from that of a tumor. Nevertheless, if genes whose expression respond in the cell culture context also display a coherent expression across patient profiles, it is likely that their consensus expression reflects the activity of this pathway in individual patient probes as shown by Bentink *et al.* (2008). We applied *guided clustering* to identify transcriptional modules that are conserved between both cellular contexts. The smoothing parameter σ was selected as described in section 4.2.4 and set to $\sigma = 0.17$. Figure 4.8 illustrates the selection process. As in the BCL6 analysis, we extracted the top 3 modules and examined them for cluster tightness in the lymphoma data and joint differential expression in the guiding data. Figure 4.6(f) shows heatmaps of the extracted gene modules on the clinical data. The genes are coherently expressed across the lymphomas and form a continuous gradient when the samples are arranged by

Table 4.3: Cox-regression analysis of DLBCL depending on the LPS-index2, adjusted for activated B-cell (ABC) status, age and Ann Arbor staging.

factor	p-value		relative risk (95% confidence interval)
LPS-index2	9.48e-7	***	0.0257 (0.0059 – 0.1110)
ABC-status	0.01170	*	2.4722 (1.2231 – 4.9970)
Ann Arbor stage	0.00132	**	3.2994 (1.5927 – 6.8350)
age	0.27995		1.4355 (0.7451 – 2.7657)

increasing LPS indices. The distribution of correlations to the class label vector for module genes and non-module genes are shown in Figure 4.6(d). The module genes stand out and are clearly enriched for LPS stimulation.

Strikingly the LPS-index2 and BCL6-index2, although derived from completely different guiding data sets are almost perfectly correlated (r >0.98). Both gene modules, including 335 (BCL6) or 198 (LPS) genes, overlap only in 73 genes. A complete gene list of the LPS-index2 can be found in supplementary Table 6.2 it contains prominent oncogenes like BCL2A1 and the transcription factor STAT1 which were previously described as BCL6 targets by Ci *et al.* (2009). Furthermore fitting a Cox-regression model including established prognostic factors as described above showed that LPS-index2 is an independent highly significant predictor of survival. The associated hazard is higher than all other included prognostic factors as shown in Table 4.3.

This further supports the hypothesis of Basso and Dalla-Favera (2010) that BCL6 in fact modulates Toll-like receptor signaling in DLBCL.

4.5 Discussion

In this chapter we introduced *guided clustering*, a novel method for the combined analysis of clinical microarray gene expression data and experimental data. in contrast to completing analysis strategies, *guided clustering* does not analyze the two data sets sequentially, but in a single joint analysis.

In a simulation study *guided clustering* behaves favorable compared to sequential analysis approaches. These approaches use only one of the two data sets for the gene selection procedure, disregarding the structure or constrains of the second data

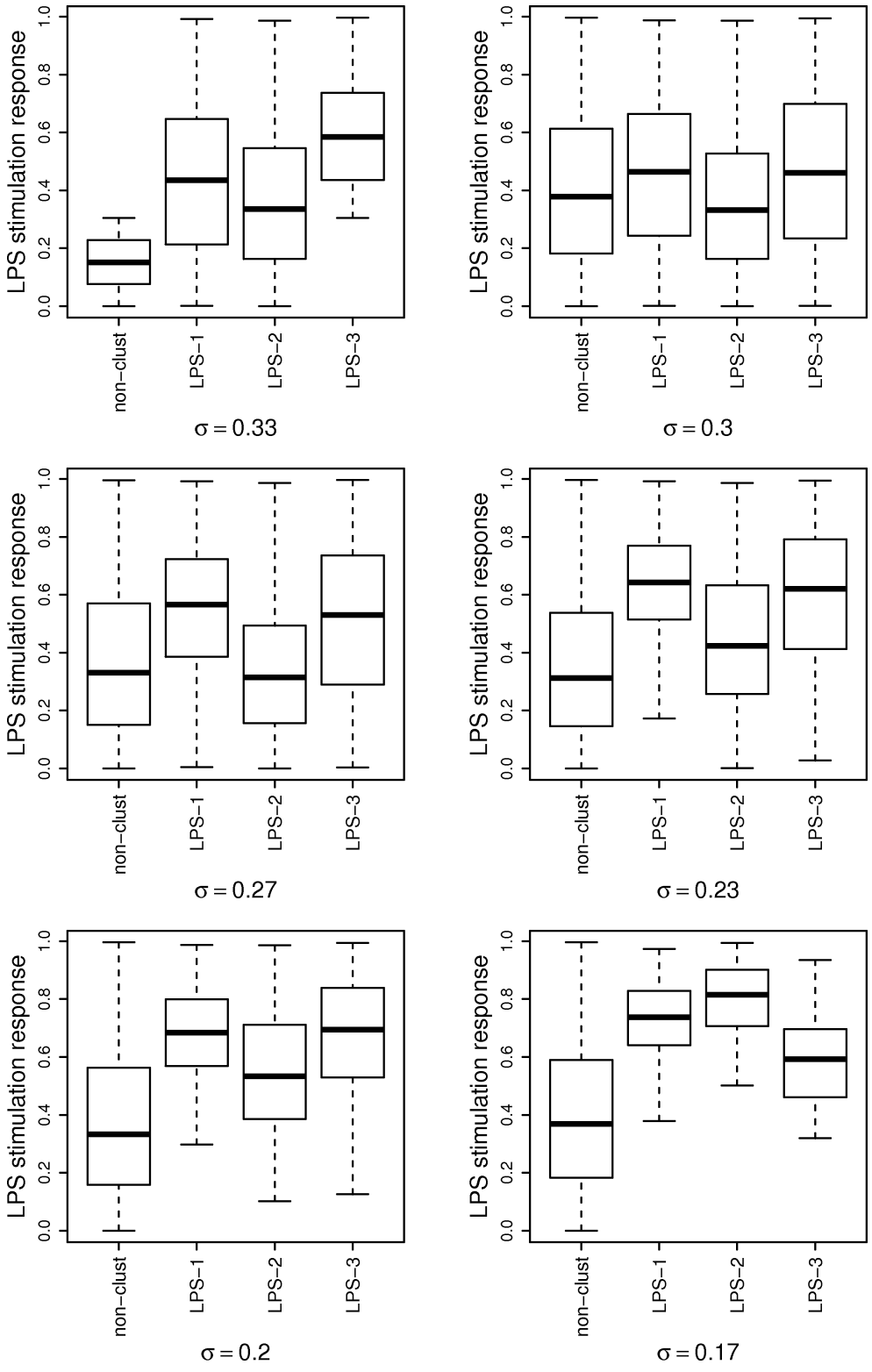


Figure 4.8: Guiding strength of top three LPS clusters extracted with different choices of σ. For $\sigma \geq 0.17$ the extracted clusters show outstanding high LPS stimulation response.

set. Our study showed that *guided clustering* is superior to sequential methods in situations when both data sets should have similar influence on the gene selection process. The method of Läuter *et al.* (2009) misses gene modules were only the minority of genes show response to perturbation, while the approach of Bild *et al.* (2006) only works correctly if the top responding genes form a cluster on the clinical data. Further *guided clustering* is more versatile and can integrate various kinds of guiding data, while the other approaches only focus on gene expression data. Our performance analysis showed that *guided clustering* favors bigger gene modules that show strong responses. Smaller gene modules with weak responses are also detectable, but will appear later in the cluster extraction process.

In this work we applied *guided clustering* in the context of a DLBCL study that was guided to focus on aspects of BCL6 and Toll-like receptor signaling. We established a novel prognostic index which holds more prognostic information than existing predictors of survival. The composition of the genes underlying this index point to a link between BCL6 activity and Toll-like receptor signaling. This link was also suggested by Basso and Dalla-Favera (2010). Among the underlying genes were known oncogenes and transcription factors that were described as BCL6 targets by Ci *et al.* (2009). We experimentally strengthened the link between BCL6 and Toll-like receptors by a LPS stimulation experiment combined with a second *guided clustering* analysis. We observed that targets of LPS mediated Toll-like receptor signaling and BCL6 targets are coherently expressed in a large collection of DLBCL suggesting that BCL6 in fact influences the transcriptional program by Toll-like receptor signaling in DLBCL.

Part III

Analysis of Single Cells

This part explores the prospects of gene expression profiling of single cells using microarray technology.

In chapter 5 we investigate the application of the Operon microarray platform to samples derived from single cells. In doing so we investigate dye specific effects within both available channels and determine the reproducibility of biological differences between samples. Further we compare the performance of different normalization methods in the context of sensitivity, classification and differential gene expression.

An application of single cell microarray analysis is given in chapter 6, where we investigate if embryonic pre-patterning is present in mice. We adopt the standard nearest shrunken centroids classifier to the scenario of paired samples. Further we develop a new strategy to detect conserved mRNA asymmetries within paired expression profiles.

Chapter 5

Evaluation of single cell microarray analysis using spike-in probes

High density oligonucleotide expression arrays proved to be a valuable tool in modern molecular biology. The gene expression microarray delivers a snapshot of the entire transcriptome of a tissue sample. Generally gene expression microarray experiments are performed on tissue samples consisting of several thousands of cells. Hence, the measured transcriptome abundances are an average across all cells of a sample.

Recently, the analysis of single cells has moved into the focus of cancer research. For instance, the model of early metastatic spread suggests that single tumor cells disseminate from the primary tumor in a very early stage of disease and may cause metastasis, even if the primary tumor is removed early as reported by Hüsemann *et al.* (2008). For a better understanding of the relation between disseminated tumor cells and metastasis microarray experiments seem promising. Another example for the use of single cell gene expression analysis is the area of reproductive medicine and embryogenic research were the properties of specific single cells are of importance. We will have a closer look on the application of gene expression microarrays in the context of embryogenesis in chapter 6. However, applying the gene expression microarray technology to single cells raises new challenges. The amount of mRNA available in one cell is small and needs additional amplification. Each amplification step introduces technical noise which exacerbates data analysis. Moreover, since each sample consists only of one single and not thousands of cells the measured expression levels are not averaged across cells, but represent the transcriptome of one cell. Hence, samples contain not only the biological variance between different

individuals but also the cellular variance within each individual. Latest research suggests that mRNA transcription within a cell is not a continuous but a rather stochastic process (Raj *et al.* (2006) and Raj and van Oudenaarden (2008)). This means that even mRNA profiles of cells from the same tissue of one patient may exhibit large differences.

In this chapter we will examine the prospects of microarray technology in the context of single cell analysis. We generated a spike-in data set using the Operon GeneChip platform. Different data normalization methods are compared and the ability of differential gene expression analysis and classification of different biological backgrounds examined.

5.1 Preliminary experiments

The Operon GeneChip is a 2-channel cDNA platform and available for different species like human and mouse. Before we created the spike-in data set to compare different normalization procedures, we performed two preliminary experiments analyzing the properties of the Operon platform. The first experiment was a color switch experiment including two arrays. The second experiment was based on the results of the color switch and investigated the reproducibility of fold changes within both channels. All data was background corrected using the *normexp* function of the *limma* package. Subsequently the data was log2 transformed.

5.1.1 Color switch

The color switch experiment consisted of two microarrays where the same amplified sample was hybridized against a common reference. However, while on one array the reference was hybridized to the Cy3 (green) and the cell to the Cy5 (red) channel, the other array was color switched. Thus the reference was hybridized to the Cy5 and cell to the Cy3 channel. This setup allows to investigate dye bias, which is the intensity difference between samples labeled with different dyes, that is attributable to the dyes instead of the gene expression in the samples. In Figure 5.1(a) and 5.1(d) the Cy3 channel of an array is plotted against the corresponding Cy5 channel. Both Figures show a banana like shape indicating differences between the intensities. Note that both Figures exhibit the same dependence between the Cy3 and Cy5 channel although the sample types are switched between the axes. Figures 5.1(b)

and 5.1(e) show the dye dependent difference clearer. Even if two samples of the same type are plotted against each other and only the dye differs, the banana shape is present. This indicates that the difference between the intensities that cause the banana shape are due to the dye and not the probe type. The channels with similar dyes were plotted against each other in Figures 5.1(c) and 5.1(f). Compared to the other plots the intensities show a high correlation. This means that intensities from the same channel of different arrays are more alike than intensities from similar samples hybridized to different channels. Additionally, the dynamic range of the two channels is different. While the Cy5 channel shows intensities on the full range (0-15) the Cy3 channel agglomerates the majority of spots in an area of (5-10).

This experiment exhibit a strong dye bias and differences in the dynamic range of the two channels of the Operon GeneChip. At this point we are unsure if the 2-color setup of the Operon platform is suitable for the analysis of single cells. Therefore we performed an experiment to investigate the reproducibility of fold changes within the two channels.

5.1.2 Conservation of fold changes

To analyze the conservation of fold changes by the two different channels of the Operon GeneChip platform we performed an experiment consisting of three microarrays. Each of the three samples that were hybridized contained 20 different mRNA fragments with defined copy numbers that were spiked into the hybridization solution. These mRNA fragments bind to specific control spots on the Operon GeneChip. The copy numbers differ between the arrays as shown in table 5.1. The intensity values of the corresponding probes of the microarrays should reflect copy numbers changes of the sample sequences. Figures 5.1(g) and 5.1(h) show the comparison of the expected and observed fold changes for the Cy5 and Cy3 channel. To evaluate the quality of fold change reproducibility more formally, we fit a linear model to the spike in probes

$$fc_{obs} = \alpha_0 + \alpha_1 fc_{exp} + \epsilon \tag{5.1}$$

where the observed fold change is a linear function of the expected fold change. The ideal result would be a line with slope 1. The resulting slopes for the Cy5 and Cy3 channel are 0.21 and 0.10 respectively. However, if the three highest and lowest expected fold changes were excluded from the analysis the slope of the Cy3 channel

drops to 0.03 while the slope of the Cy5 channel remains at 0.21. This shows that the Cy3 channel is not able to capture the spike-in fold changes while maintaining the linear differences. In contrast the Cy5 channel detects the spike-in fold changes. But also the Cy5 channel decreases the fold changes about a factor of $\frac{1}{0.21} \approx 4.5$. In all following experiments we will employ the Operon GeneChip platform in a single channel setup using the Cy5 channel only.

Table 5.1: Estimated copy number per spike-in oligo (log2) used to determine the conservation of fold changes within the Cy5 and Cy3 channel.

	Cy5			Cy3			fold change	
oligo id	A-1	A-2	A-3	A-1	A-2	A-3	Cy5	Cy3
AF159801	26.95	8.94	8.94	8.94	26.95	26.95	-18.01	18.01
opHsV04TC000049	26.25	9.41	9.41	9.41	26.25	26.25	-16.84	16.84
opHsV04TC000044	25.71	10.05	10.05	10.05	25.71	25.71	-15.66	15.66
opHsV04TC000045	23.89	10.83	10.83	10.83	23.89	23.89	-13.05	13.05
ATU91966	23.14	12.56	12.56	12.56	23.14	23.14	-10.58	10.58
opHsV04TC000002	22.55	13.15	13.15	13.15	22.55	22.55	-9.39	9.39
opHsV04TC000001	20.83	13.89	13.89	13.89	20.83	20.83	-6.94	6.94
AF168390	20.04	15.73	15.73	15.73	20.04	20.04	-4.31	4.31
AF159803	19.40	16.27	16.27	16.27	19.40	19.40	-3.13	3.13
X56062	18.92	16.96	16.96	16.96	18.92	18.92	-1.96	1.96
opHsV04TC000004	16.96	18.92	18.92	18.92	16.96	16.96	1.96	-1.96
opHsV04TC000005	16.27	19.40	19.40	19.40	16.27	16.27	3.13	-3.13
AF191028	15.73	20.04	20.04	20.04	15.73	15.73	4.31	-4.31
AF198054	13.89	20.83	20.83	20.83	13.89	13.89	6.94	-6.94
opHsV04TC000009	13.15	22.55	22.55	22.55	13.15	13.15	9.39	-9.39
opHsV04TC000008	12.56	23.14	23.14	23.14	12.56	12.56	10.58	-10.58
AF247559	10.83	23.89	23.89	23.89	10.83	10.83	13.05	-13.05
X58149	10.05	25.71	25.71	25.71	10.05	10.05	15.66	-15.66
opHsV04TC000052	9.41	26.25	26.25	26.25	9.41	9.41	16.84	-16.84
opHsV04TC000051	8.94	26.95	26.95	26.95	8.94	8.94	18.01	-18.01

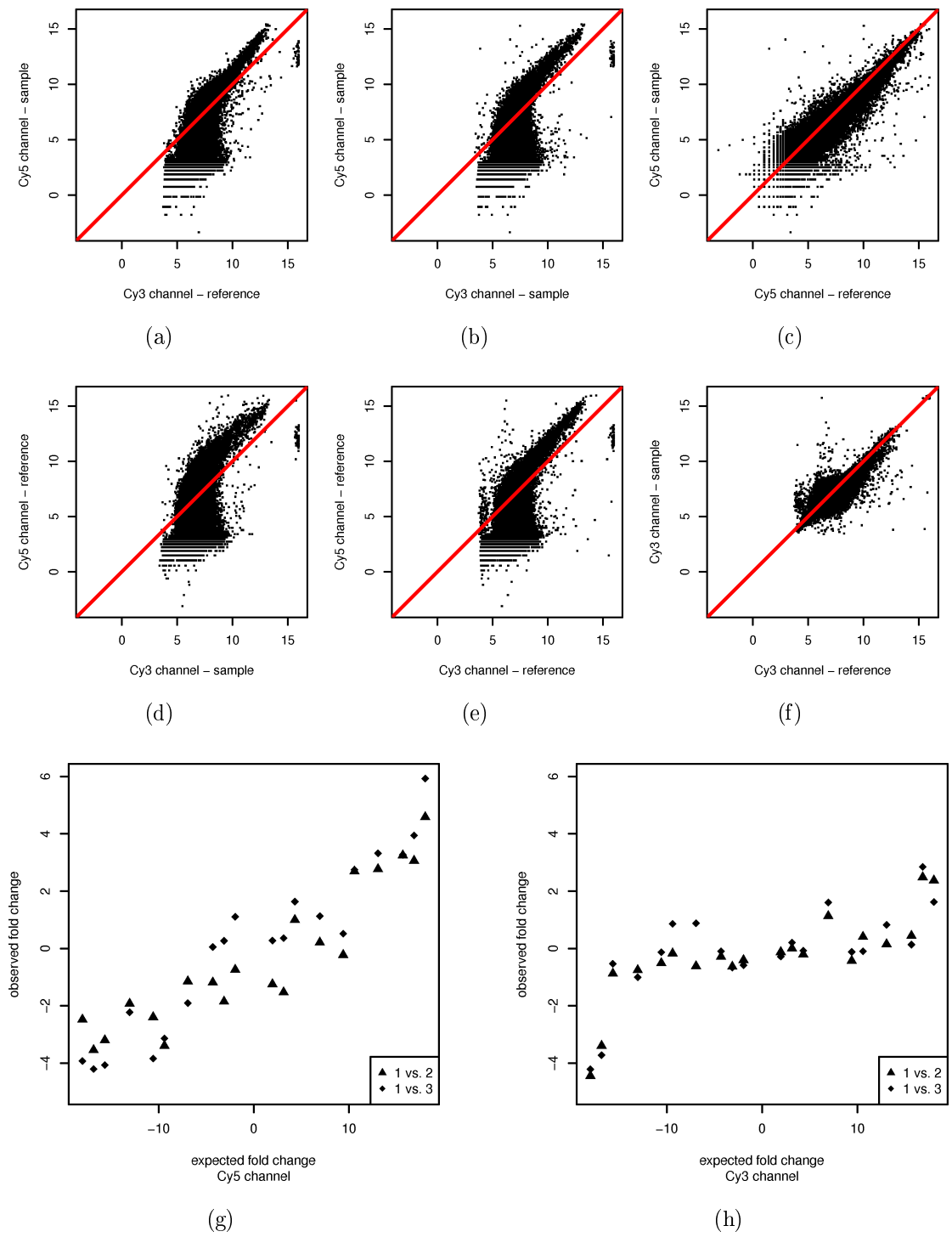


Figure 5.1: **(a-f)** Scatter plots of color switch experiment: Array1 has sample mRNA on the Cy5 and reference mRNA on the Cy3 channel, while array2 is vice versa. **(g,h)** Scatter plot expected vs. observed fold changes between spike-in oligos measured on the (g) Cy5 channel and (h) Cy3 channel.

5.2 Comparison of normalization methods

Normalization is needed when dealing with multiple microarrays. Between different arrays there exist two types of variation: biological variation and distracting variation. Researchers are interested in the biological variation and not in the distracting variation, which can have multitudinous sources. The purpose of normalization is to reduce the distracting variation, while conserving the biological. Many normalization procedures have been proposed and compared in the literature (Bolstad *et al.* (2003), Irizarry (2003)). However none of those have been applied to single cell data. Therefore we used different normalization procedures and compared their influence on our analysis results. We focus on three established algorithms namely loess normalization as proposed by Yang *et al.* (2002), quantile normalization, which was proposed by Bolstad *et al.* (2003) and variance stabilization normalization (vsn) proposed by Huber *et al.* (2002). While quantile normalization and vsn can be directly applied to the data, loess normalization needs a baseline towards each array is normalized. We defined this baseline as the median expression per gene across all arrays of our dataset. Additional we combined loess with the quantile normalization by applying quantile normalization to the expression ratios given by the loess method. Prior to data normalization background correction of the spot intensities is possible. Since it is unknown whether background correction is advantageous or not, we performed all normalization methods with and without background correction. If background correction was applied, we used the *normexp* background correction method as recommended by Ritchie *et al.* (2007). The exception is vsn which we applied directly to the raw data, since this method has a build in background correction model. Altogether we compared 9 normalization procedures: none, quantile, loess, loess-quantile, all with and without background correction and vsn. All methods were used as implemented in R employing the packages *limma* (Smyth and Speed, 2003) and *vsn* (Huber *et al.*, 2003).

5.2.1 Experimental setup

To properly analyze the gene expression measurements in terms of sensitivity, stability, classification and differential expression, data for which we know the truth is required. Assessments can only be performed where specific results are expected. Therefore we generated a spike-in data set consisting of 40 samples derived from two

different cell lines, namely Cal51 and T47D, contributing 20 samples each. Whereas the Cal51 samples were single cells the T47D samples were single cell equivalents derived from a 20 cell pool. Each cell line consist of two sample types, activated and arrested cells. Altogether the data set consists of four groups with 10 samples each: Cal51 activated, Cal51 arrested, T47D activated and T47D arrested. Likewise to the experiment described in section 5.1.2 20 unique mRNA fragments were spiked into the hybridization mixture. These fragments originate from as different species (*Arabidopsis Thaliana*) and hence do not interfere with the sample mRNA. Within each of the 4 sample groups the spike-in concentrations were arranged in a 20×10 cyclic Latin square, with each concentration appearing one per row and twice per column as showed in table 5.2.

5.2.2 Reproduction of spike-in concentrations and fold changes

To determine the quality of the spike-in probes we examined if the spike-in spot intensities increase proportional to the spike-in concentrations. Figure 5.2(a) shows average log2 intensities of the spike-in probes given a certain concentration. The log2 intensities increase with higher spike in concentrations. Background corrected intensities are lower than uncorrected ones. This effect especially affects lower and diminishes with higher intensities since background intensities are not increasing with the foreground intensities of the spots. Further the average intensity of the lowest three concentrations are similar between different background and not background corrected methods. Therefore we will not use spots with a spike-in concentration lower than 21.57 for further analysis. To specify the bias we fit the following linear model to the spike-in probes

$$\log_2 I = \alpha_0 + \alpha_1 \log_2 C + \epsilon \tag{5.2}$$

where I are the intensity values measured, C the spike-in concentrations and ϵ the error term. In the ideal case the slope would be equal to 1. Table 5.3 shows the estimated slopes for the different normalization methods. There are only small differences between the methods with similar background treatment. Slopes are ≈ 0.6 if no background correction was performed and ≈ 0.7 if background correction was done. Hence background correction is beneficial in terms of intensity bias, especially in the low intensity range.

Table 5.2: Estimated copy number (log2) per spike-in oligo for the spike-in data set.

oligo id	A-1	A-2	A-3	A-4	A-5	...	A-9	A-10
AF159801	27.58	26.58	25.58	24.58	23.58	...	19.58	18.58
opHsV04TC000049	26.58	25.58	24.58	23.58	22.58	...	18.58	27.58
opHsV04TC000044	25.58	24.58	23.58	22.58	21.58	...	27.58	26.58
opHsV04TC000045	24.58	23.58	22.58	21.58	20.58	...	26.58	25.58
ATU91966	23.58	22.58	21.58	20.58	19.58	...	25.58	24.58
opHsV04TC000002	22.58	21.58	20.58	19.58	18.58	...	24.58	23.58
opHsV04TC000001	21.58	20.58	19.58	18.58	27.58	...	23.58	22.58
AF168390	20.58	19.58	18.58	27.58	26.58	...	22.58	21.58
AF159803	19.58	18.58	27.58	26.58	25.58	...	21.58	20.58
X56062	18.58	27.58	26.58	25.58	24.58	...	20.58	19.58
opHsV04TC000004	27.58	26.58	25.58	24.58	23.58	...	19.58	18.58
opHsV04TC000005	26.58	25.58	24.58	23.58	22.58	...	18.58	27.58
AF191028	25.58	24.58	23.58	22.58	21.58	...	27.58	26.58
AF198054	24.58	23.58	22.58	21.58	20.58	...	26.58	25.58
opHsV04TC000009	23.58	22.58	21.58	20.58	19.58	...	25.58	24.58
opHsV04TC000008	22.58	21.58	20.58	19.58	18.58	...	24.58	23.58
AF247559	21.58	20.58	19.58	18.58	27.58	...	23.58	22.58
X58149	20.58	19.58	18.58	27.58	26.58	...	22.58	21.58
opHsV04TC000052	19.58	18.58	27.58	26.58	25.58	...	21.58	20.58
opHsV04TC000051	18.58	27.58	26.58	25.58	24.58	...	20.58	19.58

Besides a low intensity bias the conservation of fold changes between different samples is essential for the comparison of different biological backgrounds. To examine the reproducibility of fold changes between different samples of our data set, we calculated the absolute fold changes of the spike-in probes between all pairs of samples and compared these observed values with the expected fold changes given by the concentration differences. Figure 5.2(b) shows the results for the different normalization procedures. For all normalization methods, the observed fold changes increase linearly with the expected fold changes. Background corrected methods

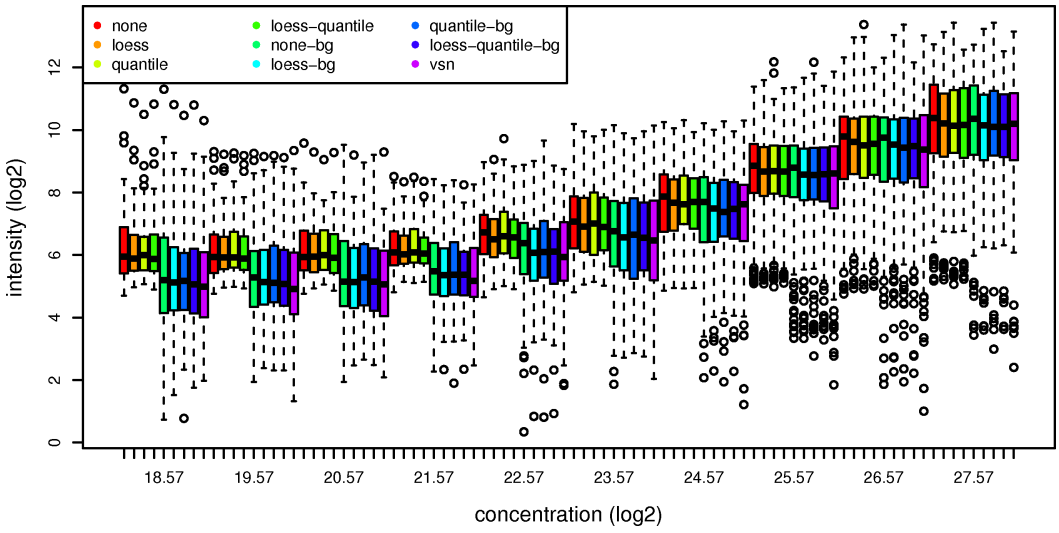
seem to achieve slightly higher fold changes than methods without background correction. Similar to the analysis above we fit the linear model

$$fc_{obs} = \alpha_0 + \alpha_1 fc_{exp} + \epsilon \tag{5.3}$$

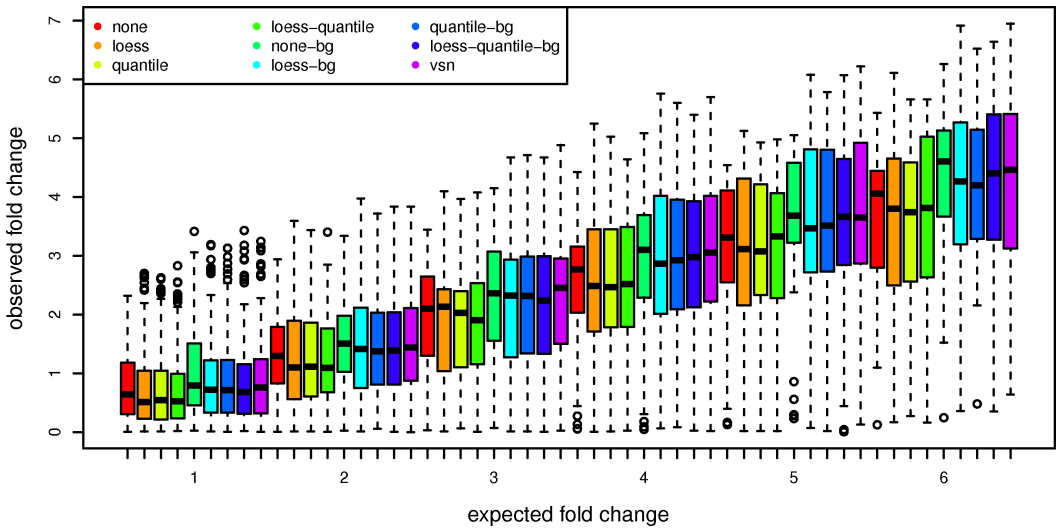
where the observed fold change fc_{obs} is a linear function of the expected fold change fc_{exp}. This model is similar to the one used in section 5.1.2. The ideal result would be a slope of 1. The results are summarized in table 5.3. Slopes are around 0.6 when no background correction was done and 0.68 if the data was background corrected. There is a bias that lowers the observed fold changes. Removing this bias from the probes by performing background correction improves conservation of fold changes.

5.2.3 Receiver operating characteristic (ROC) analysis

For any analysis that involves the comparison of microarrays it is essential to distinguish between true and random differences between the microarrays. This is only possible if the true differences are stronger than such that occur by chance. We use ROC curves to illustrate how well fold changes between spike-in probes of different arrays can be separated from the biological background (non spike-in probes). For a series of fold change thresholds ROC analysis compares the resulting true positives ratio (TPR) with the false positive ratio (FPR). A TPR is the fraction of fold changes originating from spike-in probes that are bigger than a certain threshold. In contrast the FPR is the fraction of fold changes originating from non spike-in probes that are bigger than this threshold. Optimally there exists a threshold that is smaller than all spike-in fold changes, but larger than all non spike-in fold changes. In this case, the ROC curve climbs rapidly away from the origin (lower left hand corner). A common application of microarray analysis is the comparison of different groups of samples. Probe fold changes between these groups are averaged across all samples within a group. Since our data set contains four groups of 10 samples, each featuring the same spike-in Latin square table, we can mimic four spike-in replicates by averaging across the groups. The resulting data set consists of 10 profiles were each probe intensity is an average across four samples. We calculated the absolute fold changes between all 45 pairs of the 10 averaged profiles and applied ROC analysis. Figures 5.3(a) and 5.3(b) show the resulting ROC curves. The better the result the closer a ROC curve gets to the upper right corner of the plot, which represents a perfect separation between spike-in and non spike-in fold changes. The diagonal



(a)



(b)

Figure 5.2: Boxplots illustrating sensitivity and fold change conservation. The different normalization methods are color coded as shown in the legend ('bg' indicates background correction). **(a)** Observed spot intensities as a function of spike-in concentrations. Each box consists of the intensity values measured from the spike in probes given a certain concentration across all arrays. **(b)** Observed fold changes in comparison with expected fold changes between the spike-in probes of different arrays. The absolute fold changes were calculated for each pair of arrays.

line represents the case were the fractions of spike-in and non spike-in fold changes that are above a given fold change threshold are equal. This situation is equivalent to randomly picking probes that are above the threshold. How close the ROC curve gets to the upper right corner can be quantified by measuring the area under the curve (AUC).

All normalization methods are in the upper right triangle of the plots, meaning that the separation between true and false positives is better than chance. Comparing Figure 5.3(a) with 5.3(b) reveals that normalization procedures perform better when no background correction was done. Likewise the AUC is bigger for methods without background correction, as shown in Table 5.3. Within a specific background correction mode all normalization procedures perform better than unnormalized data except the vsn method. The differences between the other normalization methods are only marginal. The loess and loess-quantile method performs slightly better than the quantile method.

5.2.4 Classification analysis

In this section we investigate, if the application of different normalization methods affects the classification problems inherent in our data. There are three such problems, two involving 2-classes and one involving 4-classes. For the 2-class problems we train classifiers separating the activated from arrested cell samples for the Cal51 and T47D cell lines separately. In the 4-class problem we train a classifier that separates all classes of the 2-class problems at once. We used a nearest shrunken centroid (NSC) classifier implemented in the R-package *PAMr* (Tibshirani *et al.*, 2002) with default settings for all classification tasks. Each classifier was evaluated by 10-fold cross validation (CV). The CV-errors were averaged across 50 independent CV runs for the 2-class problems and across 100 independent CV runs for the 4-class problem. The results are shown in Figures 5.3(c), 5.3(d) and 5.3(e).

For both 2-class problems we were able to train classifiers with low CV errors ≤ 0.05 (1 sample) with all normalization methods using between 100 and 1000 genes. Only unnormalized data, with and without background correction, performed worse. The 4-class problem is more complex. It could also be solved independent of the normalization methods used. But more genes (≥ 1000) were needed to achieve a CV error of ≤ 0.03. Again unnormalized data, with and without background correction, performed worse (CV error: $0.09 \approx 4$ samples). Table 5.3 summarizes the CV errors

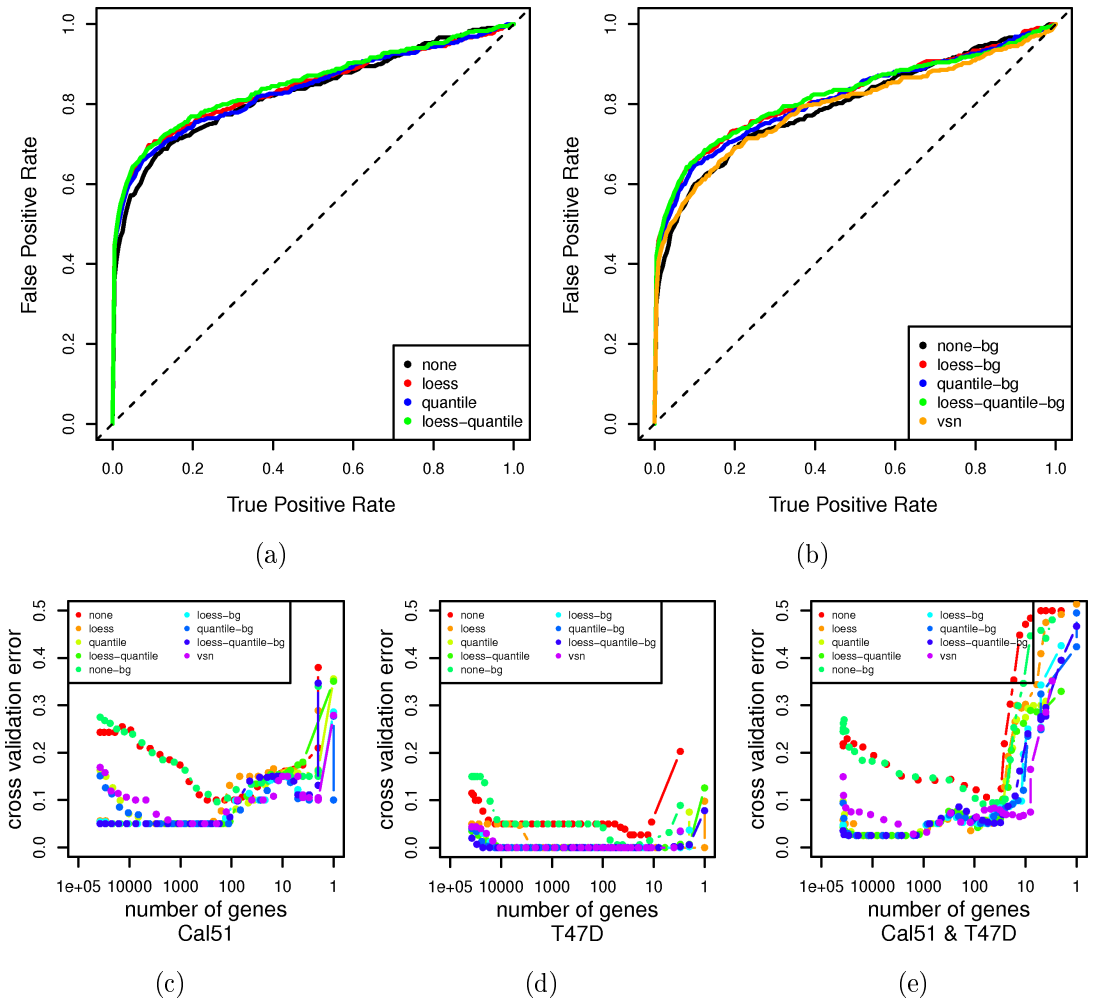


Figure 5.3: **(a,b)** ROC analysis using fold changes between spike-in probes as positives and fold changes between non spike-in probes as negatives. (a) normalization without background correction (b) normalization with background correction. **(c-e)** PAMr 10-fold cross validation error averaged across 10 independent runs as a function of the number of genes used by the classifier. (c) Cal51 activated vs. arrested (d) T47D activated vs. arrested (e) 4-class problem: Cal51 & T47D activated vs. arrested.

achieved by the different normalization methods. For all investigated classification problems application of background correction had no influence on the classifier performance.

Table 5.3: Regression slope estimates for spike-in concentration vs. probe intensities (A) and expected vs. observed fold changes (B). ROC analysis: Area Under the Curve was calculated using the trapezoidal rule. Classification: 10-fold cross validation error for activated vs. arrested, Cal51 and T47D separate and combined.

	slope estimates			CV error		
method	A	B	ROC AUC	Cal51	T47D	combined
none	0.62	0.59	0.83	0.10	0.03	0.09
loess	0.62	0.59	0.84	0.05	0.00	0.02
quantile	0.61	0.59	0.84	0.05	0.00	0.02
loess-quantile	0.61	0.59	0.85	0.05	0.00	0.02
none-bg	0.71	0.66	0.80	0.10	0.00	0.09
loess-bg	0.71	0.68	0.82	0.05	0.00	0.02
quantile-bg	0.71	0.68	0.81	0.05	0.00	0.02
loess-quantile-bg	0.71	0.68	0.82	0.05	0.00	0.02
vsn	0.73	0.70	0.79	0.05	0.00	0.03

5.2.5 Differential gene expression analysis

Differential gene expression analysis will be one of the major applications of single cell analysis. Using our data set we searched for genes that are differentially expressed between the activated and arrested samples within each cell line. We employed the R package *limma* (Smyth, 2004) to perform the calculations.

The number of significant differentially expressed genes ($P_{adj} \leq 0.05$) ranged from 47 to 91 genes for Cal51 depending on the normalization method and for T47D from 330 to 641, respectively. The composition of genes sets was similar for the different methods. Considering the overlap between all pairs of normalization methods the fraction of genes included in both sets relative to the smaller set was $\approx 90\%$ for Cal51 and $\approx 74\%$ for T47D.

To check if the biological difference of our samples (activated vs. arrested) agrees with the differentially expressed genes, we performed gene ontology (GO) enrichment analysis. This analysis checks whether sets of genes that belong to a specific biological process, molecular function or cellular component are significantly enriched in the set of differentially expressed genes. Here we use the sets of genes identified af-

ter loess-quantile normalization with background correction. There were 76 probes differentially expressed in the Cal51 and 580 in the T47D cell line. Examples of GO-terms most strongly enriched are shown in table 5.4.

Within the activated samples of the Cal51 cell line genes are up-regulated that are connected with the absorption of metal ions. Metal ions play a key role in the nutrition of cells and are essential for cell growth and cell division (Nelson, 1999). GO-terms enriched in the T47D comparison suggest that the activated samples are in the process of ribosome biogenesis. This process might be in control of cell proliferation as it antagonizes cell cycle progression until the cell has grown to an adequate size as described by Thomas (2000) and Bernstein *et al.* (2007). This suggests that serum activation has a detectable impact on the transcriptome of Cal51 and T47D cell lines.

5.3 Discussion

In this chapter we analyzed the capability of the Operon microarray platform in terms of single cell gene expression profiling. This platform is based on the 2-channel cDNA technology. Hence in a standard setup one would measure a sample on one channel and a corresponding control or common reference on the other channel. Since it is well known that dye specific effects may occur we performed a color switch experiment investigating dye specific measurement differences between similar samples. We observed huge differences between the measured probe intensities of the two channels that were independent from the hybridized sample. Such behavior can cause major problems in subsequent data analysis, as it disguises intensity differences reflecting biological properties of the samples. Aware of this problem, we investigated the conservation of fold changes between different samples within the two channels. This is the most important property of a microarray technology measuring gene expression, as expression differences between arrays should reflect biological differences between samples. Our results showed that only the Cy5 channel is able to correctly reproduce defined fold changes between biological samples. As a consequence we used only the Cy5 channel of this platform in all following applications.

We continued with comparing different normalization procedures in terms of bias and their influence on typical data analysis methods like classification and differ-

Table 5.4: GO-terms overrepresented in genes differentially expressed between activated and arrested samples within the Cal51 or T47D cell line. Regulation is with respect to the activated samples. Ontology abbr.: MF - molecular function, BP - biological process, CP - cellular component.

GO-ID	GO-term	p-value	regulation	ontology
Cal51				
GO:0046870	cadmium ion binding	5e-13	up	MF
GO:0005507	copper ion binding	2e-07	up	MF
GO:0008270	zinc ion binding	0.006	up	MF
GO:0022614	membrane to membrane docking	3e-05	up	BP
GO:0007159	leukocyte cell-cell adhesion	0.001	up	BP
GO:0022406	membrane docking	0.002	up	BP
T47D				
GO:0003735	structural constituent of ribosome	<2e-16	up	MF
GO:0003723	RNA binding	2e-05	up	MF
GO:0022890	inorganic cation transmembrane transporter activity	3e-04	up	MF
GO:0006412	translation	2e-12	up	BP
GO:0022613	ribonucleoprotein complex biogenesis	2e-06	up	BP
GO:0016072	rRNA metabolic process	3e-05	up	BP
GO:0006355	regulation of transcription, DNA-dependent	6e-04	down	BP
GO:0022626	cytosolic ribosome	4e-16	up	CP
GO:0030529	ribonucleoprotein complex	6e-13	up	CP
GO:0033279	ribosomal subunit	3e-12	up	CP
GO:0022627	cytosolic small ribosomal subunit	2e-11	up	CP

ential gene expression based on a well defined spike-in dataset. We did not detect large differences between the different normalization methods. Bias, conservation of fold changes between spike-in probes of different samples and classification accuracy were comparable. Generally results on normalized data were better than when un-

normalized data was used. The biggest influence on performance has the application of background correction. It notably reduces the intensity bias, especially in the low intensity range, but increases the variance. This increase is disadvantageous and can disguise real fold changes as it was shown in the ROC analysis.

Generally the noise present in single cell data is high, which might be due to additional amplification of sample mRNA. Unfortunately we can not make any statement about this amplification step, since the spike-ins were made after the amplification procedure. Only spike-ins before the amplification would have allowed us to asses the influence of the amplification procedure on the ratios of mRNA abundances within a sample. It is technically impossible to reliable spike-in such a low amount of sequence copy numbers that would match biological concentrations. Still spike-ins after the amplification step are useful to assess bias and sensitivity of the platform.

A final choice of the best normalization procedure is difficult since no method had outstanding performance. We favor loess-quantile normalization which had a consistently high performance. Background correction has pros and cons, but we accept the increase of variance in change for bias reduction.

At the present state of technology we can reliably classify and predict single cell samples of different biological backgrounds. Even complex multi class problems can be solved correctly. Additionally differential gene expression analysis identified probes with significant differences in expression between the activated and arrested cells. Gene set analysis of those genes confirmed the biological differences intended by serum stimulation. The Cal51 cell line responded with the absorption of metal ions, which are essential for cell growth and division (Nelson, 1999). The T47D cell line seems to be further progressed as regulated genes belong to the process of ribosome biogenesis. Thomas (2000) and Bernstein *et al.* (2007) suggest that this process is in control of the cell cycle as it antagonizes the cell cycle progression until the cell has grown to an adequate size.

Chapter 6

Analysis of transcriptome asymmetry within mouse zygotes and embryonic sister blastomeres

In the last chapter we examine on the prospects of gene expression analysis of single cells using microarray technology. We compared arrested with serum activated cell line samples and showed that typical applications of gene expression analysis, namely classification and differential gene expression analysis, are feasible. In this chapter we focus on an application of single cell analysis. Embryogenesis is one of the most fascinating processes in biology. A whole organism consisting of millions of cells, dedicated to various tasks, originates from one single totipotent cell. This cell is the zygote, which is formed by two highly differentiated gamete cells, sperm and egg. The gamete cells are not totipotent themselves, in the contrary they face assured cell death if they do not combine. This change of cell fate, that occurs during fertilization may be archetypical and promises insights into other changes of cellular potency.

Until today the processes underlying cell fate changes are not fully understood. However, it is known that cellular potency is sometimes regulated by controlling the asymmetric distribution of mRNAs during cell division, such that each daughter cell retrieves a distinct mixture of mRNAs that contribute to lineage commitment (Macara and Mili, 2008). This mechanism enables stem cells to divide such that one daughter keeps the stem cell properties, while the other is committed to further differentiation (Lin, 2008). Examples come from *Drosophila* neuroblasts and

mammalian skin stratification and differentiation (Chia *et al.* (2008), Lechler and Fuchs (2005)). These observations imply that underlying principles of asymmetric cell division are conserved between different species and tissues. Asymmetric localization of cytoplasmic mRNAs has been shown to be a conserved strategy by which cell polarity is regulated. However, its direct contribution to the formation of cell pluripotency is poorly understood.

Non-uniform distribution of cytoplasmic mRNA during gametogenesis and embryogenesis can be observed in different species. Dubowy and Macdonald (1998) showed that at least 10% of the transcriptome in Drosophila oocytes is distributed non-uniformly. This fraction increases to 71% in early embryos (Lécuyer *et al.*, 2007). In vertebrates like *Xenopus laevis* it has been shown that the synthesis of maternal proteins is located by segmenting mRNAs (King (1995), Mowry and Cote (1999), King *et al.* (2005), Holt and Bullock (2009)). For the formation of the animal and vegetal poles in the oocyte this segmentation is of critical importance (Nieuwkoop, 1985). It involves the Balbiani body formation during maturation, which can be also observed in mouse embryos (Pepling *et al.* (2007), Kloc *et al.* (2008)). This suggests that mechanisms involving asymmetry have been conserved. Although asymmetric mRNA distribution has been observed in different mammalian cell types it remains unclear whether it is present in mammalian oocytes and early embryos (Mili *et al.*, 2008).

The answer to this question is essential for clarifying whether or not oocyte and/or embryonic mRNA patterning influence the establishment of cellular totipotence in early mammalian development. So far related experiments deliver arguments for both, positive and negative answers. On the one hand Torres-Padilla *et al.* (2007) showed with 4-cell mouse embryos that methylation levels of the histone H3 correlate with the fate of each blastomere lineage. Cells with low methylation levels contribute more to the trophectoderm, while cells with high methylation levels contribute more to the inner cell mass. On the other hand data of Kurotaki *et al.* (2007) suggests that early embryonic lineages are not pre-determined maternally or in the earliest phase of embryogenesis. However, it is certain that cellular totipotency is present in mouse cell short after fertilization until at least the 8-cell embryo (Tarkowski (1959), Kelly (1977)).

We address the question of asymmetric mRNA patterning by measuring transcriptome distributions within mouse oocytes and pre-nuclear zygotes, and between

sister blastomeres of single embryos after first and second mitotic division. This analysis was performed in cooperation with the laboratory of mammalian molecular embryology of Dr. Perry from the RIKEN center for developmental biology in Kobe, Japan, and the Klein lab from the department of pathology in Regensburg, Germany. The results of our joined work are published in the *EMBO* journal (Vermilyea *et al.*, 2011).

6.1 Data set

Collection and culture of mouse oocytes, zygotes and blastomeres from 2- and 3-cell blastomeres and all necessary microsurgery was done by the lab of T. Perry. Additional qPCR measurements for transcript quantification in single cells and cell fragments to validate the microarray data were done there too. Microarray preparation, including PCR, sample labeling and array hybridization was done in the lab of C. Klein. The gene expression data set consisted of 120 microarray samples composed of 4 subsets of different sample types, namely 10 spindle-oocyte pairs, 16 polar body II (Pb2)-zygote couplets, 22 pairs of sister blastomeres derived from 2-cell embryos and 8 triplets of sister blastomeres derived from 3-cell embryos.

The 10 oocyte-spindle pairs were generated by microsurgical removal of the spindle apparatus form unfertilized metaphase II oocytes. The spindle and remnant samples were then subject to gene expression microarray analysis. This will answer the question whether certain mRNAs are more associated with the spindle than others.

Pb2-zygote couplets were harvested as soon as the Pb2 was generated by the zygote after fertilization. Comparison of the transcriptomes may reveal specific mRNAs that are removed from the zygote via the Pb2.

Sister blastomeres of 2- and 3-cell embryos were collected as soon as the 1st or 2nd mitotic cleavage occurred. This data can be used to decide whether an asymmetric mRNA distribution exist between sister blastomeres of a single embryo.

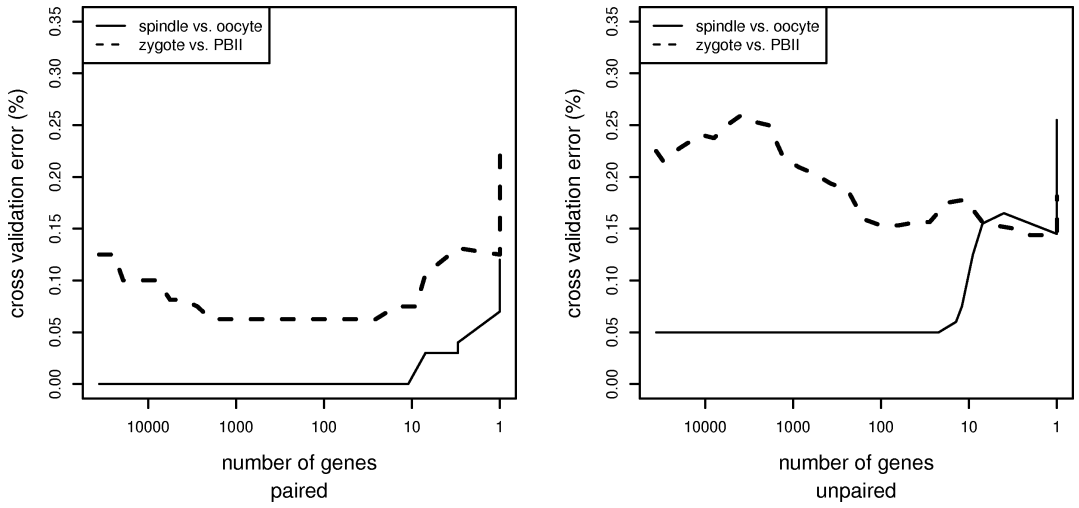
The whole data set was preprocessed by applying loess-quantile normalization with background correction as described in chapter 5.

6.2 Supervised analysis of spindle-oocyte and Pb2-zygote couplets

We begin by performing supervised analysis on the spindle-oocyte and Pb2-zygote samples, namely classification and differential gene expression. Classification will determine whether it is possible to distinguish spindle from oocyte and zygote from Pb2 samples using a particular set of genes. The differential gene expression analysis will order the measured genes according to their difference between spindle and oocyte, and Pb2 and zygote samples and determine if the differences are statistically significant.

6.2.1 Classification

Classification analysis was done using the nearest shrunken centroid method proposed by Tibshirani *et al.* (2002). We employed the R implementation from the *PAMr* package with default settings. For further details on this classification method see section 2.1.1. The default implementation of the NSC classifier does not take the pairing of our data into account. In a standard unpaired 2-class setting, where the classes are denoted A and B, a classifier decides whether an unknown sample belongs to class A or B. However, in the paired setting two samples are given at a time. The classifier now needs to decide which of the two samples belongs to class A and which to class B. If one sample is assigned to class A or B the other one has to be of the different class. We can use this additional knowledge to our advantage by adopting the classification procedure to the paired setting. To do so, we will not train the classifier on the samples themselves, but on the differences within the pairs. As an example we will use the spindle-oocyte couplets. We randomly sort the couplets into two groups of equal size, A and B. For each couplet of type A we subtract the spindle profile from the oocyte profile transcript by transcript. In contrast, couplets of type B are reversely subtracted, oocyte profile from spindle profile transcript by transcript. This procedure transforms the two profiles of a couplet into one difference. This difference is either of type A or B. The NSC classifier is now trained on these differences. Given an unknown couplet we can calculate its within difference and predict whether the difference was of type A or B, automatically implying the sample types. Classification accuracy was assessed by 10-fold cross



(a) (b)

Figure 6.1: Cross validation errors achieved by NSC classifiers trained on the spindle-oocyte and zygote-Pb2 couplets in a **(a)** paired and **(b)** unpaired setting.

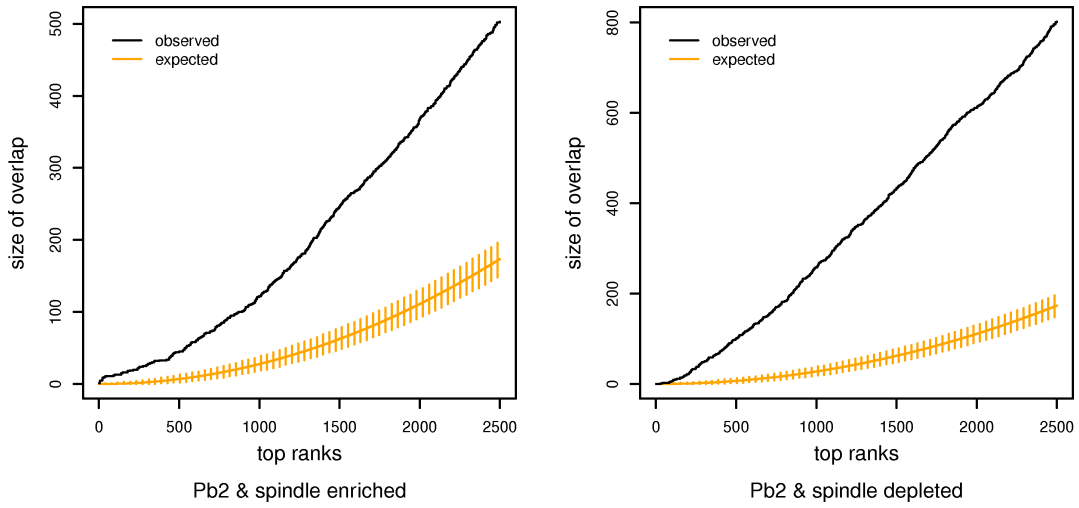
validation. The classification results are shown in Figure 6.1(a). The spindle-oocyte classifier had a CV error of zero when using at least 10 genes. The CV error of the Pb2-zygote classifier was 0.0625. Both classifiers had an increased performance due to the adoption of the paired setting. In Figure 6.1(b) the CV errors for the unpaired scenario are shown, which are higher than in the paired analysis.

The low CV errors suggest that the transcriptomes of spindles on oocytes and zygotes and Pb2 are distinguishable. An explanation could be a programmed sorting of mRNAs within oocytes, assembling certain mRNAs close to the spindle. Further the zygote might eject certain mRNAs to the Pb2. We will now address the question which genes are differently expressed between spindle and oocytes, and Pb2 and zygote.

6.2.2 Differential gene expression analysis

Differential gene expression analysis was done by employing the *limma* package implemented in R. *Limma* allows direct specification of the paired setting in the analysis design. Raw $P-$values were adjusted for multiple testing as proposed by Benjamini and Hochberg (1995). Two comparisons are of interest: (i) spindle vs. oocyte and (ii) Pb2 vs. zygote.

We begin with spindle vs. oocyte. We identified 384 transcripts to be significantly



(a) (b)

Figure 6.2: Ordered list comparisons: The observed overlap sizes between the top ranking mRNAs from the lists of mRNAs **(a)** enriched in spindle and Pb2 or **(b)** depleted in spindle and Pb2 are significantly greater than the magnitudes of overlaps expected for randomly selected mRNAs. Vertical orange bars indicate the spread of overlap sizes that occur if transcripts are chosen at random from both lists.

differentially expressed ($P_{adj} \leq 0.05$). 95 of those transcripts showed a higher and 289 a lower mRNA abundance in the spindle compared to the oocyte samples. Microarray data was verified by the laboratory of T. Perry using quantitative PCR (qPCR) analysis of 30 transcripts in independently isolated single spindle-oocyte couplets. Abundance changes were in agreement between microarray and qPCR data for 25 of 30 transcripts (83.3%).

Our comparison of Pb2 vs. zygotes determined 1069 transcripts to be differently expressed. Here 572 transcripts showed higher and 497 lower mRNA abundance in Pb2 compared to zygote samples. Again the microarray data was confirmed by qPCR experiments on independent Pb2-zygote couplets.

Given that the spindle and Pb2 samples are derived from the same intracellular region, their transcriptomes should overlap. We tested this hypothesis by comparing the list of differentially partitioned transcripts. Both lists were ordered, so that the transcript most strongly enriched in spindle or Pb2 is on the first position and the transcript most strongly enriched in the oocyte or zygote is on the last position. Then we employed the *OrderedList* package of Lottaz *et al.* (2006) to test whether the lists overlap more strongly than would have been expected by chance. Figure

6.2 shows the results. Both, spindle and Pb2 enriched as well as depleted transcripts show highly significant overlaps. Compared with Pb2-zygote pairs, fewer transcripts achieved significant concentration changes in spindle-oocyte couplets. Nevertheless, transcript rankings correlate strongly, even when including transcripts falling below the significance threshold.

6.3 Unsupervised analysis of sister blastomeres

In the previous section we described the identification of transcriptome differences within spindle-oocyte and Pb2-zygote pairs following a supervised strategy. Now we focus on the sister blastomeres derived from 2- and 3-cell embryos. Since a spatial transcript distribution was detected in oocytes and zygotes, it might be transmitted to the early embryo such that sister blastomere products of the first or second mitosis inherit distinct transcriptomes. We begin the analysis with the sister blastomeres derived from 2-cell embryos in sections 6.3.1 and 6.3.2 and advance to analysis of 3-cell embryos in section 6.3.3.

Considering the paired structure of the expression data one can rephrase the problem of asymmetric mRNA distribution as follows: Is there a set of genes that separates all pairs of sister blastomeres into two classes A and B, so that one sample of each pair belongs to class A and the other one to class B. Compared to the situation in the last section this problem is unsupervised, since it is not *a priori* possible to assign a given cell to a cell type. However, in this scenario, if one of the cells could be assigned to class A the other would have to be of class B, distinguishing the situation from standard unsupervised learning problems. Note that we are only interested in expression differences within pairs of samples, not between samples. We address this by centering the expression values of all transcripts around zero within each pair of sister blastomeres. Let X^* the matrix of pair-centered expression values, then

$$x^*_{i,j} = x_{i,j} - \bar{x}_{i,j} \ , \tag{6.1}$$

where i runs across all transcripts and j across all samples, and $\bar{x}_{i,j}$ is the mean expression of transcript i within the pair including sample j. Obviously within each pair of sister blastomeres we have

$$x^*_{i,1} = -x^*_{i,2} \ , \tag{6.2}$$

where 1 and 2 indicate the two samples within a pair and i runs through all transcripts.

6.3.1 A clustering approach for analyzing asymmetry of mRNA distributions

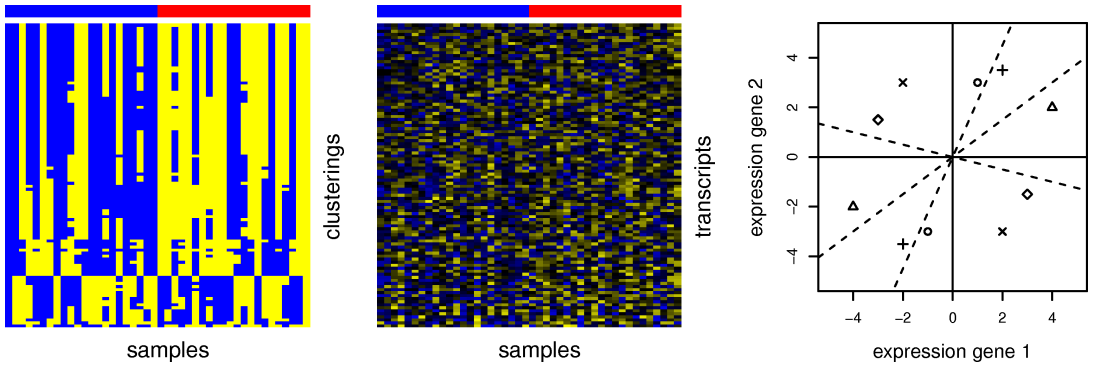
An approach to determine whether it is possible to group the sister blastomeres into the desired classes is clustering. If a conserved asymmetric distribution of mRNAs within each pair exists, a cluster algorithm should group the samples accordingly such that the two samples of each pair are in different clusters. Here we use the partitioning around medoids (PAM) cluster algorithm described in section 2.1.1. In a first attempt we perform global clustering involving all genes. The distance between any two samples s and t is defined as

$$d(s,t) = \sqrt{\sum_{i=1}^{p}(x^*_{i,s} - x^*_{i,t})^2}\ , \tag{6.3}$$

where p is the number of transcripts. This is the Euclidean distance in p dimensional space. We choose this distance measure, because it reflects abundance differences of transcripts between different profiles. We achieved a valid clustering, in that each pair of sister blastomeres was separated. However, we do not expect that an asymmetric mRNA distribution will affect the whole transcriptome. Therefore we rank all transcripts according to their contribution to the global pairwise distance. We define the contribution of a transcript g as the sum of univariate distances between any two samples s and t

$$c(g) = \sum_{i=1}^{n}\sum_{j=1}^{n} x^*_{g,i} - x^*_{g,j}\ . \tag{6.4}$$

The top 1000 transcripts have large pairwise distance sums. We try to cluster the pairs of sister blastomeres using only those top scoring transcripts and see whether they are correctly separated or not. We increase the number of transcripts used from 10 to 1000 in steps of 10. The resulting clusters are shown in Figure 6.3(a). Non of the retrieved clusterings matches the global clustering result. But, all clusterings separate the sister blastomeres so that the samples of one pair are in different clusters. This result is puzzling, since we assumed that if there is an asymmetric mRNA distribution it will separate the sister blastomeres in one specific way. We



(a) (b) (c)

Figure 6.3: **(a)** Separation of 2-cell sister blastomeres by clustering using the top $10 - 1000$ transcripts with maximal variance. The color bar on top indicates the global clustering result involving all transcripts. **(b)** Heatmap of the top 100 most variable transcripts. Yellow indicates high and blue low transcript abundance. The color bar on top indicates the global cluster result involving all transcripts. **(c)** 2-dimensional example of paired data were expression values are centered around zero for all genes. The data points represent paired samples, where the pairing is indicated by similar symbols. Due to the centering each pair is mirrored by the origin. Hence any line through the origin gives a trivial solution to separate the pairs into two groups, such that each group holds one sample of each pair.

investigate the clusterings result further by looking directly at the expression values. Figure 6.3(b) shows a heatmap of the top 100 distance contributing transcripts. The expression values are mixed, even though the samples are arranged according to the clustering. It is reasonable to assume that this observation is an artifact of the data or analysis method.

Indeed it can be easily shown that the pairwise centering of the sister blastomeres makes a correct clustering possible with any set of genes. Given a set of transcript wise centered pairs in a p-dimensional Euclidean space, where each transcript defines a dimension, all pairs can be separated by any plane that runs through the origin, given that no sample is located directly on this plane. This can be seen if we assume the plane to be defined by a vector $\vec{v}$ of unit length. The projections of the sample vectors $\vec{s}$ and $\vec{t}$ of any pair of sister blastomeres onto the axis defined by $\vec{v}$ is given by the scalar products $\vec{s} \cdot \vec{v}$ and $\vec{t} \cdot \vec{v}$. This projections are always on different sides

of the origin since

$$\vec{s} \cdot \vec{v} = -(\vec{t} \cdot \vec{v}) \tag{6.5}$$

as

$$\vec{s} \cdot \vec{v} = \sum_{i=1}^{p} x_{1,s}^{*} v_i \tag{6.6}$$

and

$$\vec{t} \cdot \vec{v} = \sum_{i=1}^{p} -x_{1,s}^{*} v_i \tag{6.7}$$

given equation 6.2, where p is the number of transcripts used. Figure 6.3(c) illustrates this artifact in a 2-dimensional example.

To answer the question if asymmetric mRNA distribution it truly present or not we need to overcome this artifact. This can be done by refining and extending the clustering approach as described in the next section.

6.3.2 Measuring group separation of sister blastomeres derived from 2-cell embryos by cluster quality

We improve the clustering approach in two ways: (i) We refine the transcript filtering, so that only those genes are used for clustering that exhibit a stable and strong separation of pairs. (ii) We introduce a resampling based significance test deciding whether a separation of sister blastomeres is artificial or real.

In the last section we filtered transcripts due to their univariate pairwise distances between samples. We improve this filtering by directly focussing on the separation of pairs. Transcripts should have large abundance differences between the two samples within a pair. Further the abundance differences of all pairs should have low variance across the pairs to ensure equally strong separation. Abundance differences were calculated transcript wise for each pair of sister blastomeres by subtracting the two expression values from each other. Since this difference can be either positive or negative we used the absolute values. To rank the genes according to separation strength and variance we employed a moderated one sided t-test. The t-statistic of each transcript is given as

$$t_{score} = \sqrt{n} \frac{\bar{x} - \mu_0}{\sigma + \sigma_0} \,, \tag{6.8}$$

where

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\tilde{x} - \bar{x})^2} \tag{6.9}$$

is the empirical standard deviation and

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i \tag{6.10}$$

is the mean across all within pair differences $\tilde{x}$ of a transcript. The number of pairs is denoted by n, and μ_0 is the value from which means $\bar{x}$ should differ, here $\mu_0 = 0$. The fudge factor σ_0 is a positive constant that is added to the standard deviation of each transcript. It guards against large t_{score} values arising from transcripts with low abundances but even lower variability. We set σ_0 equal to the median value of standard deviations over all transcripts. High values of t_{score} are achieved by transcripts with large within pair differences that show low variations. Using the ranking by t_{scores} we discarded all but the top N separating transcripts. We continue the analysis with the resulting truncated expression profiles.

To decide whether a clustering resulting from the top N separating transcripts is an artifact we compare its quality with a null-distribution of cluster qualities. This null-distribution is generated by determining the quality of 10000 clusterings obtained from independent sets of N randomly chosen transcripts. A clustering is of high quality if the two resulting classes are well separated. We measure this separation strength by calculating cluster silhouettes $\bar{s}$ as proposed by Rousseeuw (1987). A cluster silhouette is a value between -1 and 1, that describes the relation of distances between samples of different clusters and distances between samples of the same cluster. Clusterings of high quality achieve values close to 1 and have high between cluster and low within cluster distances. Values around zero indicate that no cluster structure is present. Negative values arise from wrong clusterings. A detailed description of cluster silhouettes can be found in section 2.1.3. The probability that the silhouette $\bar{s}_{top}$ of a clustering using the top N separating transcripts appeared by chance is given by the fraction of silhouettes $\bar{s}_r$ achieved by clustering using N randomly selected transcripts that are bigger than $\bar{s}_{top}$.

Silhouettes achieved by clusterings using the top N separating transcripts in comparison to their null-distribution are shown in Figure 6.4(a). The number of used transcripts N was varied from 2 to 100 in steps of 1. The observed silhouettes did not exceed those from random clusters. Inspection of the top 100 separating transcripts reveal no structure as shown in Figure 6.4(d). We thus conclude that there is no detectable difference within the transcriptomes of sister blastomeres derived from 2-cell embryos. However, even though our approach was well-thought-out it might

be that it is unsuited for this type of data. Hence we validate our approach using the spindle-oocyte and Pb2-zygote pairs. For the analysis we leave the known classes aside and check whether the resulting clustering generate the right assignments. Within pair differences were calculated and transcripts ranked as described above. Likewise cluster silhouettes were calculated for the top N separating transcripts as well as for 10000 random clusters. In both cases separation strengths exceed by-chance expectations and confirmed spindle-oocyte and Pb2-zygote assignments as shown in Figures 6.4(c) and 6.4(b). Only one Pb2-zygote pair was assigned wrong. Further inspection of the top 100 separating transcripts show a clear cluster structure (Figures 6.4(f) and 6.4(e)). This validates our approach and shows that the analysis strategy has the power to expose hidden classes of transcriptomes. Consistent with this analysis we were unable to classify 2-cell blastomeres by qPCR (Vermilyea *et al.*, 2011).

6.3.3 Detection of asymmetric mRNA distribution in sister blastomeres derived from 3-cell embryos

Undetectable transcriptome segregation between sister blastomeres of 2-cell embryos does not formally exclude the possibility that asymmetries occur in the second mitotic products. A spatial mRNA distribution may be inherited and conserved during the first mitotic division as illustrated by Figure 6.5. To evaluate this alternative we analyzed transcript profiles of sister blastomeres derived from 3-cell embryos. Altogether 8 triplets blastomeres were available, measured by 24 individual profiles.

We employ the same analysis approach as described above. However, the triplet scenario yields constraints that need to be reflected in the statistical analysis. Given that during the first mitotic division the transcriptome is distributed uniformly between parent and daughter cell, we further assume that the second mitotic division generates asymmetric transcriptomes, then for some transcripts the mRNA abundance of the remaining first mitotic product is lower than one of the second mitotic products and higher than that of the other one. That means, given a blastomere triplet, where A is the remaining undivided first mitotic product and B, β are the two second mitotic products, each asymmetric distributed transcript can order the triplet according to its abundance as β, A, B or B, A, β. Therefore we aim to cluster the blastomere triplets into three classes (type A, B and β). Since transcript

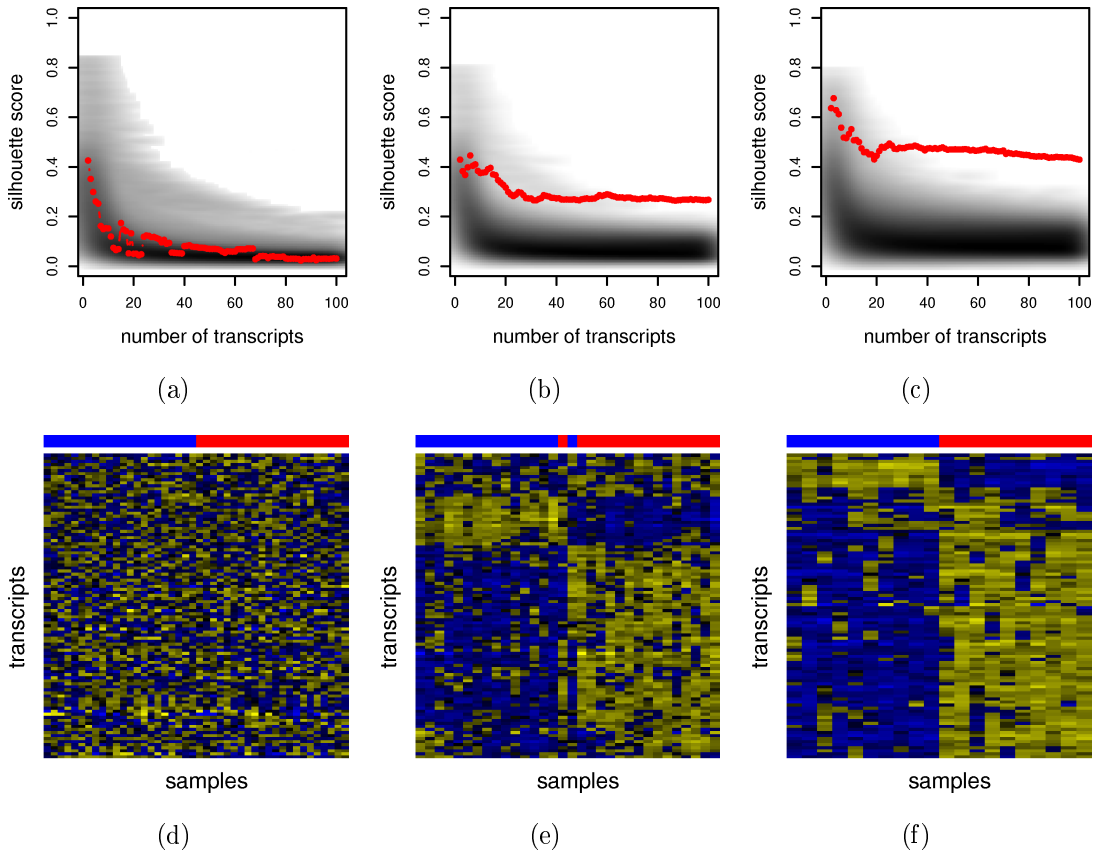


Figure 6.4: Cluster analysis of sister blastomeres derived from 2-cell embryos on the left, spindle-oocyte pairs in the middle and Pb2-zygote pairs on the right. **(a-c)** Silhouettes achieved by clusterings using the top N separating transcripts are indicated by the red line. Grey background shading reflects the distribution of matching random silhouettes. Darker colors represent higher densities. **(d-f)** Heatmaps of top 100 separating transcripts, yellow indicates high and blue low mRNA abundance. The two clusters are indicated by the color bar on top.

abundances can be on different levels we center each triplet around the remaining first mitotic product A, by subtracting A from B and β for each transcript. Let X^{**} be the matrix of centered triplets, then

$$x^{**}_{i,j} = x_{i,j} - x_{i,A} \tag{6.11}$$

where i runs across all transcripts and j across all samples. $x_{i,A}$ is the matching A-type sample of triplet including j. Hence all A-type profiles have a constant centered transcript abundance of zero. Likewise to the analysis above we rank the transcripts according to their potential separation strength within the triplets.

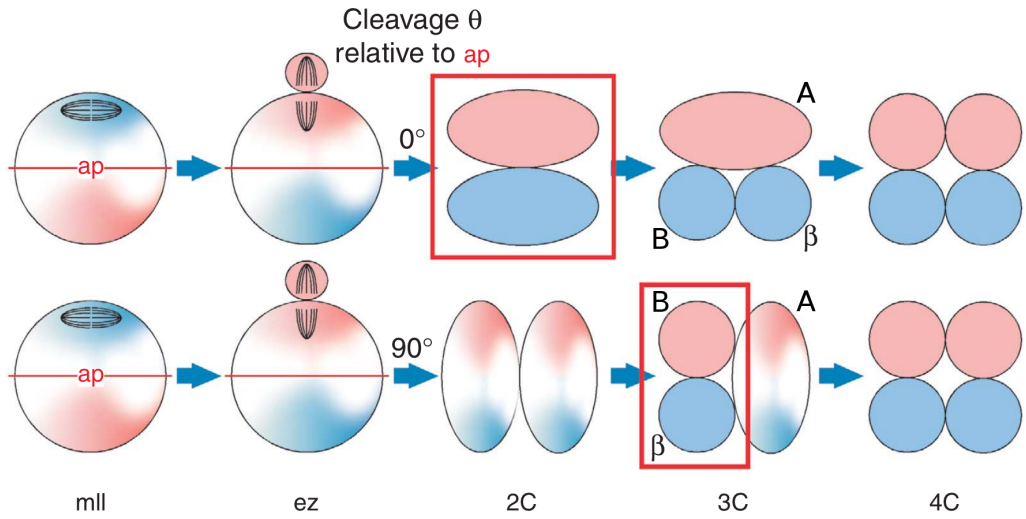


Figure 6.5: Asymmetric maternal transcriptome inheritance may persist through the early zygote (ez) and not be detected until the first (upper) or second mitotic division. Polarity determinants marking the lineages are shown as red or blue. Asymmetry detection (boxed) is possible after first (upper) or second mitosis. Image was taken from Vermilyea *et al.* (2011).

Instead of the t-test we use the Jonckheere-Terpstra test (Jonckheere, 1954), which test for a monotone trend in terms of given classes. In the three class case the test statistic is defined as

$$jt_{score} = \sum_{i=1}^{2} \sum_{j=i+1}^{3} \vartheta(c_i, c_j) \tag{6.12}$$

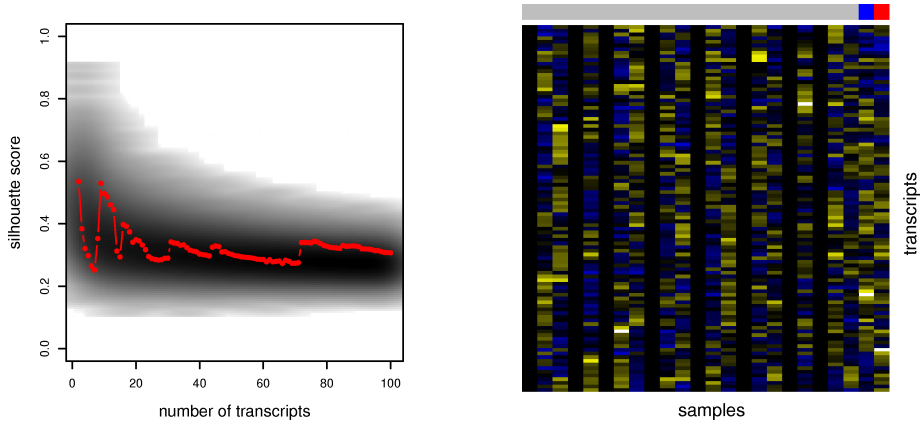
with

$$\vartheta(c_i, c_j) = \sum_{e \in c_i} \sum_{f \in c_j} \psi(x^{**}_{t,e} - x^{**}_{t,f}) \tag{6.13}$$

and

$$\psi(v) = \begin{cases} 1 & \text{if } v < 0 \\ 0 & \text{else} \end{cases} \tag{6.14}$$

where t is any transcript and c are the sample classes with c_1 including samples of type β, c_2 samples of type A and c_3 samples of type B. Here we assume that for any triplet the abundance value of type β is smaller than the abundance value of type B. We ensure this by sorting the triplets. We continue with truncated profiles using only the top N separating transcripts. Clustering and calculation of silhouettes was done similar to the paired scenario.



(a) (b)

Figure 6.6: Cluster analysis of sister blastomeres derived from 3-cell embryos. **(a)** Silhouettes achieved by clusterings using the top N separating transcripts are indicated by the red line. Grey background shading reflects the distribution of matching random silhouettes. Darker colors represent higher densities. **(b)** Heatmap of top 100 separating transcripts, yellow indicates high and blue low mRNA abundance. The three clusters are indicated by the color bar (grey - A-, red - B- and blue - β-type).

The resulting silhouettes are shown in Figure 6.6(a). Again the separation strength achieved by using the top N transcripts did not exceed random expectations. Moreover inspection of the top 100 separating transcripts shows no cluster structure. This result was confirmed by qPCR analysis (Vermilyea *et al.*, 2011). Thus, we found no evidence for programmed transcriptome asymmetry between sister blastomeres of 3-cell embryos.

6.4 Discussion

In this chapter we reported on the problem of intracellular mRNA regionalization by single subcellular transcriptome profiling. Two question were addressed: (i) Does non-uniform mRNA distribution exist in mammalian oocytes? (ii) Are transcriptome asymmetries present in early embryonic development, that might result from embryonic pre-patterning? We addressed the first question by analyzing spindle-oocyte and Pb2-zygote pairs. Investigation of sister blastomeres of 2- and 3-cell blastomeres resolved the second question.

From the bioinformatics perspective the comparison of spindle with oocyte samples as well as Pb2 with zygote samples can be accomplished with standard supervised approaches. Here we employed classification methods to showed that transcriptome differences within the spindle-oocyte and Pb2-zygote pairs exist. We adopted the standard NSC classifier to the paired setting by learning on within pair abundance differences instead of individual abundance values. This increased the classifier performance and allowed almost errorless separation of the sample groups. Differential gene expression analysis identified sets of transcripts that differ significantly within the pairs and ranked all transcripts according to their within abundance difference. Further analysis of those ranked lists showed that both spindle/Pb2 up regulated and spindle/Pb2 down regulated transcripts overlap stronger than expected by chance. This asymmetry might be a remnant from earlier oocyte development, during which mRNA localization had a critical role. It is known that spindle complexes are sites of targeted polysomal mRNA localization in diverse systems (Blower *et al.* (2007) Mili and Macara (2009)) and characteristically orchestrate key elements of mouse oocyte maturation (Brunet and Verlhac, 2011). Spindle enrichment of mRNAs might thus localize translation to target proteins during the establishment of a fertilizable mII oocyte. Alternatively the strong overlap between spindle and Pb2 enriched transcripts may suggest a unique tier of gene regulation during the gamete-to-embryo transition.

The analysis of sister blastomeres derived from 2- and 3-cell embryos aimed to detect sets of transcripts that consistently separate the pairs or triplets. This is an unsupervised analysis scenario that we approached by clustering. However the paired structure of samples derived from 2-cell embryos introduced an artifact that generated consistent pair separation for any set of transcripts. To overcome this artifact we developed a strategy that combined the gene filtering with analysis of cluster separation to detect non artificial separations. We did not detect conserved transcriptome differences within sister blastomeres of 2- or 3-cell embryos. However, as a proof of principle we applied our approach to the spindle-oocyte and Pb2-zygote pairs. We were able to identify the correct separation of pairs in both cases. Our results are in line with the work of Hiiragi and Solter (2004), Motosugi *et al.* (2005), Kurotaki *et al.* (2007) and Guo *et al.* (2010) who were unable to classify 2-cell blastomeres according to reproducible differences between them. Neither of the cells makes a biased contribution to the inner cell mass or trophectoderm of the early

embryo. Vermilyea *et al.* (2011) reason that if there is maternal or early embryonic pattering in mammals, it is more likely to be mediated post-translationally. This is in contrast to the targeted localization of oocyte derived mRNAs in *Drosophila* and *Xenopus*, which result in developmentally critical transcriptome partitioning between embryonic cells.

In summary, mice do not seem to utilize the developmental mechanism of non-uniform mRNA distribution between sister blastomeres of the early embryo that is highly conserved in other species, even though transcriptome asymmetries exist in mouse oocytes.

Part IV

Appendix

Summary and Conclusions

Today microarray gene expression profiling has become a routine high throughput method in modern molecular biology laboratories. Gene expression data is used to characterize and distinguish tissues, cells or tumors. However, the regulation and interaction of genes is still not fully understood. In this context combination of gene expression data with other per gene measurements like transcription factor binding or response to a certain stimulus can improve the understanding of the underlying regulatory mechanisms. Further, this approach enables us to categorize gene expression profiles according to the additional information like the presence of a transcription factor. This thesis aims to develop an algorithm that allows the integration of clinical gene expression data with additional quantitative per gene measurements.

Another trend in the field of gene expression profiling is the analysis of single cells. The employment of microarray technology for single cell analysis is rather new, but applications range from the analysis of rare cell populations, like disseminated tumor cells, to fertility medicine or embryogenesis, where single cells are of major importance. This thesis evaluates the prospects of single cell gene expression analysis using the Operon GeneChip platform.

Genomic data integration This part of the thesis is dealing with integrating gene expression data from clinical biopsies (clinical data) with additional quantitative per gene measurements (guiding data). We begin with an introduction to the field of genomic data integration in the context of gene expression data in chapter 3. We criticize the sequential nature of existing analysis concepts. Genes are selected either based on a coherent expression across the clinical or are high scoring in terms of some statistic on the guiding data. Subsequently the gene set is applied to the other data set. Hence the resulting signatures are either not influenced by the clinical, or they are not influenced by the guiding data. In chapter 4 we propose

a novel algorithm that selects genes based on both, the clinical and guiding data in one single joint analysis. The algorithm is based on a kernel density estimation approach and automatically weights both data sets against each other. It is named *guided clustering*. A detailed description is given in section 4.2. We investigated the performance of *guided clustering* in-depth by running a simulation study. The results are compared with the sequential approaches in section 4.3. We observed that *guided clustering* cannot outperform the sequential approaches with respect to both, cluster tightness and enrichment of high scoring genes. But, it is the only method able to perform good in both aspects simultaneously. In section 4.4 we show two applications of *guided clustering* to real data. We integrate 220 lymphoma gene expression profiles with DNA binding affinity of BCL6. This analysis identified a set of genes whose summarized expression index is a highly significant independent predictor of survival. The associated hazard of this index is higher than other established factors including ABC/GCB status, age and Ann Arbor staging. Gene set analysis showed a significant enrichment of genes involved in the Toll-like receptor pathway. This connection between the Toll-like receptor pathway and BCL6 was already described by Basso and Dalla-Favera (2010). To verify this finding we conducted a LPS stimulation experiment on BL-2 cell lines, since LPS directly activated the Toll-like receptors. Application of the *guided clustering* algorithm identified a gene module whose summarized expression index is highly correlated with the BCL6 index, even though only the minority of genes overlap between both modules. This further supports the hypothesis that BCL6 modulates the transcriptome of DLBCL via Toll-like receptor signaling.

Within this thesis a novel data integration algorithm has been developed and proven to work properly in the desired data setting. *Guided clustering* complements existing standard approaches and even outperforms them. The application of *guided clustering* to a lymphoma cancer data set discovered a novel independent factor for survival and supported the hypothesis of Toll-like receptor signaling mediating BCL6 induced gene regulation.

Analysis of single cells In this part of the thesis we explore the prospects of single cell gene microarray expression analysis. Chapter 5 evaluates the microarray platform used for the analysis of single cells throughout this thesis. This evaluation becomes necessary as measurement and analysis of single cells are more difficult than normal tissue samples, that consist of several thousand cells. Investigation of

the individual channels available on the Operon platform in section 5.1.1 and 5.1.2 showed that only the red channel is suitable for single cell analysis. Subsequently we compared the performance of several normalization methods in section 5.2. We designed a spike-in data set to asses the influence of different normalization procedures on the gene expression data in terms of sensitivity, stability, classifiability and detection of differential gene expression. We did not detect significant differences between the different normalization methods, but normalization is beneficial compared to unnormalized data. Our analysis showed that standard applications of gene expression data like classification and differential gene expression analysis are possible with single cell data.
To summarize, gene expression measurement of single cells is more difficult and the resulting data is more noisy than those of usual tissue samples. Nevertheless analysis is possible, but has to be interpreted more carefully.

Encouraged by these results we applied our single cell gene expression platform to a data set from developmental biology. This analysis is described in chapter 6 and investigates the two questions: (i) How is cellular pluripotency established within the zygote? (ii) Is there an asymmetric mRNA distribution present in the early mammalian development, like it is observed in other species like *Drosophila* or *Xenopus*? We investigate the first question in section 6.2 by analyzing spindle-oocyte and Pb2-zygote couplets. We show that there are conserved differences within the transcriptomes of spindle-oocyte and Pb2-zygote couplets. This was achieved by adopting the standard NSC classifier to the paired setting of the data. Further we detected transcripts enriched or depleted in the spindle and Pb2 and observed that overlaps between these sets of transcripts were significantly larger than expected by chance. It is possible that the spindle is utilized to eject specific mRNAs from the zygote via the Pb2. To answer the second question we analyze pairs and triplets of expression profiles derived from 2- or 3-cell sister blastomeres. In section 6.3 we develop a novel analysis strategy to decide whether a conserved asymmetric mRNA distribution exist within the pairs and triplets of sister blastomeres. We do not observe any conserved mRNA asymmetries. As a proof of principle we show that our approach is able to detect mRNA asymmetries within the Pb2-zygote or spindle-oocyte pairs. This supports our result that no asymmetric mRNA distributions are present within 2- or 3- cell sister blastomeres.

In summary, transcriptomic prepatterning during the first two mitotic divisions is either to subtle or not-existent. Hence, if there exists maternal prepatterning it is more likely to be mediated post translationally. However, transcriptomic differences were identified between spindle and oocyte as well as between Pb2 and zygote profiles. In this context the spindle might be utilized by the cell to eject specific mRNAs from the Pb2. Further analysis of the genes differently expressed between Pb2 and zygote might help do develop RNA-induced pluripotent stem cells. Additionally those genes might provide markers that can be used judge the viability of and embryo. This would be a valuable tool in human assisted reproduction as it allows the selection of single embryos.

Supplement

6.5 LPS sample preparation

BL2 cells were cultivated as described recently by Vockerodt *et al.* (2001). Cells were incubated with 1μg/ml LPS (Lipopolysaccharide from E. coli 055:B5, Sigma) for 6hrs. Cells were harvested, washed with PBS containing 1mM Sodium orthovanadate (Sigma) as described by Holtick *et al.* (2005). RNA was isolated using the RNeasy Plus Mini Kit (Qiagen). For whole genome microarrays RNA was labelled for microarray hybridization using Affymetrix GeneChip^{TM} IVT Labelling Kit (Affymetrix). Fragmentation and hybridization of labelled anti sense RNA on Human Genome U133Plus2.0^{TM}arrays (Affymetrix) was performed according to manufacturer's recommendations.

All lab work was performed by the lab of D. Kube from the university clinic for hematology of Göttingen.

6.6 Additional tables

Table 6.1: List of genes that are members of the BCL6-index2 module.

Accession Number	Entrez Id	Symbol	Description
NM_139276	6774	STAT3	signal transducer and activator of transcription 3 (acute-phase response factor)
NM_030938	81671	TMEM49	TMEM49 transmembrane protein 49
NM_002956	6249	CLIP1	CAP-GLY domain containing linker protein 1
NM_004049	597	BCL2A1	BCL2-related protein A1
NM_002110	3055	HCK	hemopoietic cell kinase
NM_000880	3574	IL7	interleukin 7
NM_006399	10538	BATF	basic leucine zipper transcription factor, ATF-like
NM_003929	8934	RAB7L1	RAB7, member RAS oncogene family-like 1
NM_016545	51278	IER5	immediate early response 5
NM_000594	7124	TNF	tumor necrosis factor
NM_006465	10620	ARID3B	AT rich interactive domain 3B (BRIGHT-like)
NM_007237	11262	SP140	SP140 nuclear body protein
NM_000560	963	CD53	CD53 molecule
NM_001558	3587	IL10RA	interleukin 10 receptor, alpha
D29642	9938	ARHGAP25	Rho GTPase activating protein 25
NM_181078	50615	IL21R	interleukin 21 receptor
NM_005204	1326	MAP3K8	mitogen-activated protein kinase kinase kinase 8
NM_001759	894	CCND2	cyclin D2
NM_018639	55884	WSB2	WD repeat and SOCS box-containing 2
NM_021105	5359	PLSCR1	phospholipid scramblase 1
NM_015149	23179	RGL1	ral guanine nucleotide dissociation stimulator-like 1
NM_002356	4082	MARCKS	myristoylated alanine-rich protein kinase C substrate
NM_002114	3096	HIVEP1	human immunodeficiency virus type I enhancer binding protein 1
NM_014372	26994	RNF11	ring finger protein 11
NM_000391	1200	TPP1	tripeptidyl peptidase I
NM_138444	115207	KCTD12	potassium channel tetramerisation domain containing 12
NM_000043	355	FAS	Fas (TNF receptor superfamily, member 6)
NM_007318	5663	PSEN1	presenilin 1
NM_016546	51279	C1RL	complement component 1, r subcomponent-like
NM_015364	23643	LY96	lymphocyte antigen 96
NM_002727	5552	SRGN	serglycin
NM_001610	53	ACP2	acid phosphatase 2, lysosomal
NM_000201	3383	ICAM1	intercellular adhesion molecule 1
NM_000211	3689	ITGB2	integrin, beta 2 (complement component 3 receptor 3 and 4 subunit)
NM_005561	3916	LAMP1	lysosomal-associated membrane protein 1
NM_003461	7791	ZYX	zyxin
NM_005627	6446	SGK1	serum/glucocorticoid regulated kinase 1
NM_000593	6890	TAP1	transporter 1, ATP-binding cassette, sub-family B (MDR/TAP)
NM_019034	54509	RHOF	ras homolog gene family, member F (in filopodia)
NM_005658	7185	TRAF1	TNF receptor-associated factor 1
NM_002229	3726	JUNB	jun B proto-oncogene
NM_000595	4049	LTA	lymphotoxin alpha (TNF superfamily, member 1)
NM_005574	4005	LMO2	LIM domain only 2 (rhombotin-like 1)
NM_001165	330	BIRC3	baculoviral IAP repeat-containing 3
NM_004125	NA	NA	NA
NM_007315	6772	STAT1	signal transducer and activator of transcription 1, 91kDa
NM_022168	64135	IFIH1	interferon induced with helicase C domain 1
NM_000434	4758	NEU1	sialidase 1 (lysosomal sialidase)

Accession Number	Entrez Id	Symbol	Description
NM_016582	51296	SLC15A3	solute carrier family 15, member 3
NM_014314	23586	DDX58	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58
NM_019027	54502	RBM47	RNA binding motif protein 47
NM_000878	3560	IL2RB	interleukin 2 receptor, beta
NM_018295	55281	TMEM140	transmembrane protein 140
NM_004454	2119	ETV5	ets variant 5
NM_003768	8682	PEA15	phosphoprotein enriched in astrocytes 15
NM_003263	7096	TLR1	toll-like receptor 1
NM_003332	7305	TYROBP	TYRO protein tyrosine kinase binding protein
NM_007161	7940	LST1	leukocyte specific transcript 1
NM_006407	10550	ARL6IP5	ADP-ribosylation-like factor 6 interacting protein 5
NM_006317	10409	BASP1	brain abundant, membrane attached signal protein 1
NM_000399	1959	EGR2	early growth response 2
NM_001665	391	RHOG	ras homolog gene family, member G (rho G)
NM_016061	51646	YPEL5	yippee-like 5 (Drosophila)
NM_001310	1389	CREBL2	cAMP responsive element binding protein-like 2
NM_007229	11252	PACSIN2	protein kinase C and casein kinase substrate in neurons 2
NM_032895	84981	C17orf91	chromosome 17 open reading frame 91
NM_002118	3109	HLA-DMB	major histocompatibility complex, class II, DM beta
NM_199418	5547	PRCP	prolylcarboxypeptidase (angiotensinase C)
NM_002468	4615	MYD88	myeloid differentiation primary response gene (88)
NM_000958	5734	PTGER4	prostaglandin E receptor 4 (subtype EP4)
NM_017791	55640	FLVCR2	feline leukemia virus subgroup C cellular receptor family, member 2
NM_030762	79365	BHLHE41	basic helix-loop-helix family, member e41
NM_021259	58986	TMEM8A	transmembrane protein 8A
NM_053056	595	CCND1	cyclin D1
NM_006148	3927	LASP1	LIM and SH3 protein 1
NM_152862	10109	ARPC2	actin related protein 2/3 complex, subunit 2, 34kDa
NM_000952	5724	PTAFR	platelet-activating factor receptor
NM_002406	4245	MGAT1	mannosyl (alpha-1,3-)-glycoprotein beta-1,2-N-acetylglucosaminyltransferase
NM_001084	8985	PLOD3	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3
NM_000332	6310	ATXN1	ataxin 1
NM_031858	4077	NBR1	neighbor of BRCA1 gene 1
NM_007278	11337	GABARAP	GABA(A) receptor-associated protein
NM_002117	3107	HLA-C	major histocompatibility complex, class I, C
NM_004383	1445	CSK	c-src tyrosine kinase
NM_018668	26276	VPS33B	vacuolar protein sorting 33 homolog B (yeast)
NM_004184	7453	WARS	tryptophanyl-tRNA synthetase
NM_030796	81552	VOPP1	vesicular, overexpressed in cancer, prosurvival protein 1
NM_021009	7316	UBC	ubiquitin C
NM_002984	6351	CCL4	chemokine (C-C motif) ligand 4
NM_018247	55754	TMEM30A	transmembrane protein 30A
NM_002349	4065	LY75	lymphocyte antigen 75
NM_020755	57515	SERINC1	serine incorporator 1
NM_003900	8878	SQSTM1	sequestosome 1
NM_005669	7905	REEP5	receptor accessory protein 5
L78132	3964	LGALS8	lectin, galactoside-binding, soluble, 8
NM_006813	10957	PNRC1	proline-rich nuclear receptor coactivator 1
NM_030797	81553	FAM49A	family with sequence similarity 49, member A
NM_003820	8764	TNFRSF14	tumor necrosis factor receptor superfamily, member 14 (herpesvirus entry mediator)
NM_001859	1317	SLC31A1	solute carrier family 31 (copper transporters), member 1

Accession Number	Entrez Id	Symbol	Description
NM_001995	2180	ACSL1	acyl-CoA synthetase long-chain family member 1
NM_001500	2762	GMDS	GDP-mannose 4,6-dehydratase
NM_006403	4739	NEDD9	neural precursor cell expressed, developmentally down-regulated 9
NM_000206	3561	IL2RG	interleukin 2 receptor, gamma
NM_080593	85236	HIST1H2BK	histone cluster 1, H2bk
NM_002355	4074	M6PR	mannose-6-phosphate receptor (cation dependent)
NM_022371	64222	TOR3A	torsin family 3, member A
NM_005717	10092	ARPC5	actin related protein 2/3 complex, subunit 5, 16kDa
NM_014313	23585	TMEM50A	transmembrane protein 50A
NM_198892	55589	BMP2K	BMP2 inducible kinase
NM_001659	377	ARF3	ADP-ribosylation factor 3
NM_002199	3660	IRF2	interferon regulatory factor 2
NM_000271	4864	NPC1	Niemann-Pick disease, type C1
NM_006827	10972	TMED10	transmembrane emp24-like trafficking protein 10 (yeast)
NM_002983	NA	NA	NA
NM_003272	7107	GPR137B	G protein-coupled receptor 137B
NM_024602	79654	HECTD3	HECT domain containing 3
NM_003407	7538	ZFP36	zinc finger protein 36, C3H type, homolog (mouse)
NM_003299	7184	HSP90B1	heat shock protein 90kDa beta (Grp94), member 1
NM_004419	1847	DUSP5	dual specificity phosphatase 5
NM_000572	3586	IL10	interleukin 10
NM_018320	55298	RNF121	ring finger protein 121
NM_006875	11040	PIM2	pim-2 oncogene
NM_145804	25841	ABTB2	ankyrin repeat and BTB (POZ) domain containing 2
NM_004418	1844	DUSP2	dual specificity phosphatase 2
NM_002037	2534	FYN	FYN oncogene related to SRC, FGR, YES
NM_016323	51191	HERC5	hect domain and RLD 5
NM_020739	9236	CCPG1	cell cycle progression 1
NM_022130	64083	GOLPH3	golgi phosphoprotein 3 (coat-protein)
NM_002872	5880	RAC2	ras-related C3 botulinum toxin substrate 2 (rho family, small GTP binding protein Rac2)
NM_005668	7903	ST8SIA4	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 4
NM_004811	9404	LPXN	leupaxin
NM_006889	942	CD86	CD86 molecule
NM_002209	3683	ITGAL	integrin, alpha L (antigen CD11A (p180), lymphocyte function-associated antigen 1; alpha polypeptide)
NM_015999	51094	ADIPOR1	adiponectin receptor 1
NM_007207	11221	DUSP10	dual specificity phosphatase 10
NM_000520	3073	HEXA	hexosaminidase A (alpha polypeptide)
NM_032582	84669	USP32	ubiquitin specific peptidase 32
NM_014153	29066	ZC3H7A	zinc finger CCCH-type containing 7A
NM_004045	475	ATOX1	ATX1 antioxidant protein 1 homolog (yeast)
NM_016133	51141	INSIG2	insulin induced gene 2
NM_020374	57102	C12orf4	chromosome 12 open reading frame 4
NM_006785	10892	MALT1	mucosa associated lymphoid tissue lymphoma translocation gene 1
NM_001154	308	ANXA5	annexin A5
NM_001946	1848	DUSP6	dual specificity phosphatase 6
NM_005125	9973	CCS	copper chaperone for superoxide dismutase
NM_014705	9732	DOCK4	dedicator of cytokinesis 4
NM_002648	5292	PIM1	pim-1 oncogene
NM_152244	29916	SNX11	sorting nexin 11
NM_016951	51192	CKLF	chemokine-like factor
NM_004691	9114	ATP6V0D1	ATPase, H+ transporting, lysosomal 38kDa, V0 subunit d1

Accession Number	Entrez Id	Symbol	Description
NM_000127	2131	EXT1	exostosin 1
NM_002228	3725	JUN	jun proto-oncogene
NM_024508	79413	ZBED2	zinc finger, BED-type containing 2
NM_025079	80149	ZC3H12A	zinc finger CCCH-type containing 12A
NM_006748	6503	SLA	Src-like-adaptor
NM_006437	143	PARP4	poly (ADP-ribose) polymerase family, member 4
NM_012249	23433	RHOQ	ras homolog gene family, member Q
NM_006537	9960	USP3	ubiquitin specific peptidase 3
NM_007126	7415	VCP	valosin-containing protein
NM_003896	8869	ST3GAL5	ST3 beta-galactoside alpha-2,3-sialyltransferase 5
NM_205847	29926	GMPPA	GDP-mannose pyrophosphorylase A
NM_005923	4217	MAP3K5	mitogen-activated protein kinase kinase kinase 5
NM_018009	55080	TAPBPL	TAP binding protein-like
NM_004833	9447	AIM2	absent in melanoma 2
NM_001693	526	ATP6V1B2	ATPase, H+ transporting, lysosomal 56/58kDa, V1 subunit B2
NM_001382	1798	DPAGT1	dolichyl-phosphate (UDP-N-acetylglucosamine) N-acetylglucosaminephosphotransferase 1 (GlcNAc-1-P transferase)
NM_017699	54847	SIDT1	SID1 transmembrane family, member 1
NM_020657	57343	ZNF304	zinc finger protein 304
NM_015346	23503	ZFYVE26	zinc finger, FYVE domain containing 26
NM_003113	6672	SP100	SP100 nuclear antigen
NM_003945	8992	ATP6V0E1	ATPase, H+ transporting, lysosomal 9kDa, V0 subunit e1
NM_002087	2896	GRN	granulin
NM_006016	8763	CD164	CD164 molecule, sialomucin
NM_014923	22862	FNDC3A	fibronectin type III domain containing 3A
NM_006384	10519	CIB1	calcium and integrin binding 1 (calmyrin)
NM_015388	25844	YIPF3	Yip1 domain family, member 3
NM_006923	6388	SDF2	stromal cell-derived factor 2
NM_014182	29095	ORMDL2	ORM1-like 2 (S. cerevisiae)
NM_006313	9958	USP15	ubiquitin specific peptidase 15
NM_004161	5861	RAB1A	RAB1A, member RAS oncogene family
NM_006405	10548	TM9SF1	transmembrane 9 superfamily member 1
NM_007348	22926	ATF6	activating transcription factor 6
NM_004079	1520	CTSS	cathepsin S
NM_000259	4644	MYO5A	myosin VA (heavy chain 12, myoxin)
NM_017445	NA	NA	NA
NM_000633	596	BCL2	B-cell CLL/lymphoma 2
NM_006432	10577	NPC2	Niemann-Pick disease, type C2
NM_006811	10955	SERINC3	serine incorporator 3
NM_003199	6925	TCF4	transcription factor 4
NM_014350	25816	TNFAIP8	tumor necrosis factor, alpha-induced protein 8
NM_003190	6892	TAPBP	TAP binding protein (tapasin)
NM_002298	3936	LCP1	lymphocyte cytosolic protein 1 (L-plastin)
AB014574	23307	FKBP15	FK506 binding protein 15, 133kDa
NM_015994	51382	ATP6V1D	ATPase, H+ transporting, lysosomal 34kDa, V1 subunit D
NM_006284	6881	TAF10	TAF10 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 30kDa
NM_005345	3303	HSPA1A	heat shock 70kDa protein 1A
NM_004288	9595	CYTIP	cytohesin 1 interacting protein
NM_005067	6478	SIAH2	seven in absentia homolog 2 (Drosophila)
NM_001777	961	CD47	CD47 molecule
NM_005160	157	ADRBK2	adrenergic, beta, receptor kinase 2
NM_003879	8837	CFLAR	CASP8 and FADD-like apoptosis regulator

Accession Number	Entrez Id	Symbol	Description
NM_032263	84223	IQCG	IQ motif containing G
NM_003764	8676	STX11	syntaxin 11
NM_018447	55831	TMEM111	transmembrane protein 111
NM_001375	1777	DNASE2	deoxyribonuclease II, lysosomal
NM_001295	1230	CCR1	chemokine (C-C motif) receptor 1
NM_005506	950	SCARB2	scavenger receptor class B, member 2
NM_005761	10154	PLXNC1	plexin C1
NM_004430	1960	EGR3	early growth response 3
NM_006457	10611	PDLIM5	PDZ and LIM domain 5
NM_004800	9375	TM9SF2	transmembrane 9 superfamily member 2
NM_020532	57142	RTN4	reticulon 4
NM_012428	27020	NPTN	neuroplastin
NM_006066	10327	AKR1A1	aldo-keto reductase family 1, member A1 (aldehyde reductase)
NM_203463	253782	LASS6	LAG1 homolog, ceramide synthase 6
NM_004604	6810	STX4	syntaxin 4
NM_015497	25963	TMEM87A	transmembrane protein 87A
BC002646	3725	JUN	jun proto-oncogene
NM_005433	7525	YES1	v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1
NM_004075	1407	CRY1	cryptochrome 1 (photolyase-like)
NM_134421	3241	HPCAL1	hippocalcin-like 1
NM_005514	3106	HLA-B	major histocompatibility complex, class I, B
NM_004339	754	PTTG1IP	pituitary tumor-transforming 1 interacting protein
NM_006736	3300	DNAJB2	DnaJ (Hsp40) homolog, subfamily B, member 2
NM_057158	1846	DUSP4	dual specificity phosphatase 4
NM_018179	55729	ATF7IP	activating transcription factor 7 interacting protein
NM_005335	3059	HCLS1	hematopoietic cell-specific Lyn substrate 1
NM_002502	4791	NFKB2	nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (p49/p100)
NM_003033	6482	ST3GAL1	ST3 beta-galactoside alpha-2,3-sialyltransferase 1
NM_005620	6282	S100A11	S100 calcium binding protein A11
NM_000876	3482	IGF2R	insulin-like growth factor 2 receptor
NM_019045	54521	WDR44	WD repeat domain 44
NM_033535	26234	FBXL5	F-box and leucine-rich repeat protein 5
NM_002064	2745	GLRX	glutaredoxin (thioltransferase)
NM_000153	2581	GALC	galactosylceramidase
NM_030790	81533	ITFG1	integrin alpha FG-GAP repeat containing 1
NM_006368	10488	CREB3	cAMP responsive element binding protein 3
NM_030666	1992	SERPINB1	serpin peptidase inhibitor, clade B (ovalbumin), member 1
NM_006184	4924	NUCB1	nucleobindin 1
NM_003290	7171	TPM4	tropomyosin 4
NM_015303	23355	VPS8	vacuolar protein sorting 8 homolog (S. cerevisiae)
NM_001334	1519	CTSO	cathepsin O
NM_002868	5869	RAB5B	RAB5B, member RAS oncogene family
NM_001690	523	ATP6V1A	ATPase, H+ transporting, lysosomal 70kDa, V1 subunit A
NM_007001	11046	SLC35D2	solute carrier family 35, member D2
NM_003009	6415	SEPW1	selenoprotein W, 1
NM_023079	65264	UBE2Z	ubiquitin-conjugating enzyme E2Z
NM_198183	9246	UBE2L6	ubiquitin-conjugating enzyme E2L 6
NM_017631	55601	DDX60	DEAD (Asp-Glu-Ala-Asp) box polypeptide 60
NM_002194	3628	INPP1	inositol polyphosphate-1-phosphatase
NM_021960	4170	MCL1	myeloid cell leukemia sequence 1 (BCL2-related)
NM_004926	677	ZFP36L1	zinc finger protein 36, C3H type-like 1
NM_006624	10771	ZMYND11	zinc finger, MYND domain containing 11

Accession Number	Entrez Id	Symbol	Description
NM_015602	26092	TOR1AIP1	torsin A interacting protein 1
NM_015368	24145	PANX1	pannexin 1
NM_017582	55585	UBE2Q1	ubiquitin-conjugating enzyme E2Q family member 1
NM_014301	23479	ISCU	iron-sulfur cluster scaffold homolog (E. coli)
NM_004972	3717	JAK2	Janus kinase 2
NM_022750	64761	PARP12	poly (ADP-ribose) polymerase family, member 12
NM_006714	10924	SMPDL3A	sphingomyelin phosphodiesterase, acid-like 3A
X56841	3133	HLA-E	major histocompatibility complex, class I, E
NM_022821	64834	ELOVL1	elongation of very long chain fatty acids (FEN1/Elo2, SUR4/Elo3, yeast)-like 1
D84454	7355	SLC35A2	solute carrier family 35 (UDP-galactose transporter), member A2
NM_012248	22928	SEPHS2	selenophosphate synthetase 2
NM_020199	56951	C5orf15	chromosome 5 open reading frame 15
NM_004637	7879	RAB7A	RAB7A, member RAS oncogene family
NM_004688	9111	NMI	N-myc (and STAT) interactor
NM_005533	3430	IFI35	interferon-induced protein 35
NM_000161	2643	GCH1	GTP cyclohydrolase 1
NM_002076	2799	GNS	glucosamine (N-acetyl)-6-sulfatase
NM_001822	1123	CHN1	chimerin (chimaerin) 1
NM_000081	1130	LYST	lysosomal trafficking regulator
NM_015040	200576	PIKFYVE	phosphoinositide kinase, FYVE finger containing
NM_199193	9577	BRE	brain and reproductive organ-expressed (TNFRSF1A modulator)
NM_138720	3017	HIST1H2BD	histone cluster 1, H2bd
NM_014701	9728	SECISBP2L	SECIS binding protein 2-like
NM_004034	310	ANXA7	annexin A7
NM_004566	5209	PFKFB3	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3
NM_144977	163486	DENND1B	DENN/MADD domain containing 1B
NM_005981	6302	TSPAN31	tetraspanin 31
NM_152866	931	MS4A1	membrane-spanning 4-domains, subfamily A, member 1
NM_014713	9741	LAPTM4A	lysosomal protein transmembrane 4 alpha
NM_014741	9776	ATG13	ATG13 autophagy related 13 homolog (S. cerevisiae)
NM_006472	10628	TXNIP	thioredoxin interacting protein
NM_004951	1880	GPR183	G protein-coupled receptor 183
NM_005730	10106	CTDSP2	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase 2
NM_017740	55625	ZDHHC7	zinc finger, DHHC-type containing 7
NM_020119	56829	ZC3HAV1	zinc finger CCCH-type, antiviral 1
NM_006079	10370	CITED2	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2
NM_025076	80146	UXS1	UDP-glucuronate decarboxylase 1
NM_017583	54765	TRIM44	tripartite motif-containing 44
NM_021729	55823	VPS11	vacuolar protein sorting 11 homolog (S. cerevisiae)
NM_005449	9214	FAIM3	Fas apoptotic inhibitory molecule 3
NM_015292	23344	ESYT1	extended synaptotagmin-like protein 1
NM_003331	7297	TYK2	tyrosine kinase 2
NM_016248	11215	AKAP11	A kinase (PRKA) anchor protein 11
NM_007131	7626	ZNF75D	zinc finger protein 75D
AF055376	4094	MAF	v-maf musculoaponeurotic fibrosarcoma oncogene homolog (avian)
NM_005713	10087	COL4A3BP	collagen, type IV, alpha 3 (Goodpasture antigen) binding protein
NM_078476	11120	BTN2A1	butyrophilin, subfamily 2, member A1
NM_145725	7187	TRAF3	TNF receptor-associated factor 3
NM_002618	5194	PEX13	peroxisomal biogenesis factor 13
NM_199072	29969	MDFIC	MyoD family inhibitor domain containing

Accession Number	Entrez Id	Symbol	Description
NM_017567	55577	NAGK	N-acetylglucosamine kinase
NM_015568	26051	PPP1R16B	protein phosphatase 1, regulatory (inhibitor) subunit 16B
NM_021175	57817	HAMP	hepcidin antimicrobial peptide
NM_002827	5770	PTPN1	protein tyrosine phosphatase, non-receptor type 1
NM_001001481	55284	UBE2W	ubiquitin-conjugating enzyme E2W (putative)
NM_016053	51019	CCDC53	coiled-coil domain containing 53
NM_032283	84243	ZDHHC18	zinc finger, DHHC-type containing 18
NM_013314	29760	BLNK	B-cell linker
NM_001779	965	CD58	CD58 molecule
NM_004310	399	RHOH	ras homolog gene family, member H
NM_019111	3122	HLA-DRA	major histocompatibility complex, class II, DR alpha
NM_014548	29767	TMOD2	tropomodulin 2 (neuronal)
NM_032227	84187	TMEM164	transmembrane protein 164
NM_005475	10019	SH2B3	SH2B adaptor protein 3
NM_004290	9604	RNF14	ring finger protein 14
NM_177538	57404	CYP20A1	cytochrome P450, family 20, subfamily A, polypeptide 1
NM_000484	351	APP	amyloid beta (A4) precursor protein
NM_000183	3032	HADHB	hydroxyacyl-CoA dehydrogenase/3-ketoacyl-CoA thiolase/enoyl-CoA hydratase (trifunctional protein), beta subunit
NM_004855	9488	PIGB	phosphatidylinositol glycan anchor biosynthesis, class B
NM_014506	27348	TOR1B	torsin family 1, member B (torsin B)
NM_033339	840	CASP7	caspase 7, apoptosis-related cysteine peptidase
NM_003877	8835	SOCS2	suppressor of cytokine signaling 2
NM_006277	50618	ITSN2	intersectin 2
NM_003144	6745	SSR1	signal sequence receptor, alpha
NM_006633	10788	IQGAP2	IQ motif containing GTPase activating protein 2
NM_004635	7867	MAPKAPK3	mitogen-activated protein kinase-activated protein kinase 3
NM_002250	3783	KCNN4	potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4
NM_003516	NA	NA	NA
NM_003512	8334	HIST1H2AC	histone cluster 1, H2ac
NM_005180	648	BMI1	BMI1 polycomb ring finger oncogene
NM_003930	8935	SKAP2	src kinase associated phosphoprotein 2
NM_014937	22876	INPP5F	inositol polyphosphate-5-phosphatase F

Table 6.2: List of genes that are members of the LPS-index2 module.

Acession Number	Entrez Id	Symbol	Description
NM_016623	51571	FAM49B	family with sequence similarity 49, member B
NM_004504	3267	AGFG1	ArfGAP with FG repeats 1
NM_000945	5534	PPP3R1	protein phosphatase 3, regulatory subunit B, alpha
NM_006136	830	CAPZA2	capping protein (actin filament) muscle Z-line, alpha 2
NM_014044	25972	UNC50	unc-50 homolog (C. elegans)
NM_004034	310	ANXA7	annexin A7
NM_012428	27020	NPTN	neuroplastin
NM_003816	8754	ADAM9	ADAM metallopeptidase domain 9
NM_003101	6646	SOAT1	sterol O-acyltransferase 1
NM_003851	8804	CREG1	cellular repressor of E1A-stimulated genes 1
NM_014028	28962	OSTM1	osteopetrosis associated transmembrane protein 1
NM_000310	5538	PPT1	palmitoyl-protein thioesterase 1
NM_003338	7321	UBE2D1	ubiquitin-conjugating enzyme E2D 1 (UBC4/5 homolog, yeast)
NM_007246	11275	KLHL2	kelch-like 2, Mayven (Drosophila)
NM_016078	51030	FAM18B1	family with sequence similarity 18, member B1
NM_006016	8763	CD164	CD164 molecule, sialomucin
NM_002408	4247	MGAT2	mannosyl (alpha-1,6-)-glycoprotein beta-1,2-N-acetylglucosaminyltransferase
NM_002485	4683	NBN	nibrin
NM_003968	9039	UBA3	ubiquitin-like modifier activating enzyme 3
NM_033339	840	CASP7	caspase 7, apoptosis-related cysteine peptidase
AJ131244	10802	SEC24A	SEC24 family, member A (S. cerevisiae)
NM_024920	79982	DNAJB14	DnaJ (Hsp40) homolog, subfamily B, member 14
NM_001920	NA	NA	NA
NM_003144	6745	SSR1	signal sequence receptor, alpha
NM_030939	81688	C6orf62	chromosome 6 open reading frame 62
NM_181836	51014	TMED7	transmembrane emp24 protein transport domain containing 7
NM_014570	26286	ARFGAP3	ADP-ribosylation factor GTPase activating protein 3
NM_002884	5906	RAP1A	RAP1A, member of RAS oncogene family
NM_006936	6612	SUMO3	SMT3 suppressor of mif two 3 homolog 3 (S. cerevisiae)
NM_018247	55754	TMEM30A	transmembrane protein 30A
NM_016041	51009	DERL2	Der1-like domain family, member 2
NM_005347	3309	HSPA5	heat shock 70kDa protein 5 (glucose-regulated protein, 78kDa)
NM_005719	10094	ARPC3	actin related protein 2/3 complex, subunit 3, 21kDa
NM_000395	1439	CSF2RB	colony stimulating factor 2 receptor, beta, low-affinity (granulocyte-macrophage)
NM_002349	4065	LY75	lymphocyte antigen 75
NM_004125	NA	NA	NA
NM_005669	7905	REEP5	receptor accessory protein 5
NM_004580	5873	RAB27A	RAB27A, member RAS oncogene family
NM_001779	965	CD58	CD58 molecule
NM_198892	55589	BMP2K	BMP2 inducible kinase
NM_020375	57103	C12orf5	chromosome 12 open reading frame 5
NM_181777	7319	UBE2A	ubiquitin-conjugating enzyme E2A (RAD6 homolog)
NM_004800	9375	TM9SF2	transmembrane 9 superfamily member 2
NM_006827	10972	TMED10	transmembrane emp24-like trafficking protein 10 (yeast)
NM_016607	51566	ARMCX3	armadillo repeat containing, X-linked 3
NM_022168	64135	IFIH1	interferon induced with helicase C domain 1
NM_000201	3383	ICAM1	intercellular adhesion molecule 1
NM_004049	597	BCL2A1	BCL2-related protein A1
NM_005204	1326	MAP3K8	mitogen-activated protein kinase kinase kinase 8
NM_002053	2633	GBP1	guanylate binding protein 1, interferon-inducible, 67kDa

Acession Number	Entrez Id	Symbol	Description
NM_004099	2040	STOM	stomatin
NM_000434	4758	NEU1	sialidase 1 (lysosomal sialidase)
NM_002127	3135	HLA-G	major histocompatibility complex, class I, G
NM_000593	6890	TAP1	transporter 1, ATP-binding cassette, sub-family B (MDR/TAP)
NM_002117	3107	HLA-C	major histocompatibility complex, class I, C
X56841	3133	HLA-E	major histocompatibility complex, class I, E
NM_005514	3106	HLA-B	major histocompatibility complex, class I, B
NM_002116	3105	HLA-A	major histocompatibility complex, class I, A
NM_004390	1512	CTSH	cathepsin H
NM_004604	6810	STX4	syntaxin 4
NM_017631	55601	DDX60	DEAD (Asp-Glu-Ala-Asp) box polypeptide 60
NM_004184	7453	WARS	tryptophanyl-tRNA synthetase
NM_004172	6507	SLC1A3	solute carrier family 1 (glial high affinity glutamate transporter), member 3
NM_001995	2180	ACSL1	acyl-CoA synthetase long-chain family member 1
NM_014880	9936	CD302	CD302 molecule
NM_004938	1612	DAPK1	death-associated protein kinase 1
NM_203330	966	CD59	CD59 molecule, complement regulatory protein
NM_000712	644	BLVRA	biliverdin reductase A
NM_006283	6867	TACC1	transforming, acidic coiled-coil containing protein 1
D29642	9938	ARHGAP25	Rho GTPase activating protein 25
NM_015033	23048	FNBP1	formin binding protein 1
NM_003768	8682	PEA15	phosphoprotein enriched in astrocytes 15
NM_003808	8741	TNFSF13	tumor necrosis factor (ligand) superfamily, member 13
NM_005561	3916	LAMP1	lysosomal-associated membrane protein 1
NM_002087	2896	GRN	granulin
NM_003896	8869	ST3GAL5	ST3 beta-galactoside alpha-2,3-sialyltransferase 5
NM_019027	54502	RBM47	RNA binding motif protein 47
NM_014737	9770	RASSF2	Ras association (RalGDS/AF-6) domain family member 2
NM_002110	3055	HCK	hemopoietic cell kinase
NM_030796	81552	VOPP1	vesicular, overexpressed in cancer, prosurvival protein 1
NM_006399	10538	BATF	basic leucine zipper transcription factor, ATF-like
NM_016545	51278	IER5	immediate early response 5
NM_199072	29969	MDFIC	MyoD family inhibitor domain containing
NM_197966	637	BID	BH3 interacting domain death agonist
NM_004710	9144	SYNGR2	synaptogyrin 2
NM_001621	196	AHR	aryl hydrocarbon receptor
NM_001109	101	ADAM8	ADAM metallopeptidase domain 8
NM_021105	5359	PLSCR1	phospholipid scramblase 1
NM_030666	1992	SERPINB1	serpin peptidase inhibitor, clade B (ovalbumin), member 1
NM_052847	2788	GNG7	guanine nucleotide binding protein (G protein), gamma 7
NM_002827	5770	PTPN1	protein tyrosine phosphatase, non-receptor type 1
NM_001777	961	CD47	CD47 molecule
NM_005949	4494	MT1F	metallothionein 1F
NM_004281	9531	BAG3	BCL2-associated athanogene 3
NM_000043	355	FAS	Fas (TNF receptor superfamily, member 6)
NM_001421	2000	ELF4	E74-like factor 4 (ets domain transcription factor)
NM_017491	9948	WDR1	WD repeat domain 1
NM_006209	5168	ENPP2	ectonucleotide pyrophosphatase/phosphodiesterase 2
NM_006403	4739	NEDD9	neural precursor cell expressed, developmentally down-regulated 9
NM_007315	6772	STAT1	signal transducer and activator of transcription 1, 91kDa
NM_018009	55080	TAPBPL	TAP binding protein-like
NM_001084	8985	PLOD3	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3

Acession Number	Entrez Id	Symbol	Description
AF333388	645745	MT1P2	metallothionein 1 pseudogene 2
NM_144653	138151	NACC2	NACC family member 2, BEN and BTB (POZ) domain containing
NM_001719	655	BMP7	bone morphogenetic protein 7
NM_007289	4311	MME	membrane metallo-endopeptidase
NM_000594	7124	TNF	tumor necrosis factor
NM_005761	10154	PLXNC1	plexin C1
NM_006504	5791	PTPRE	protein tyrosine phosphatase, receptor type, E
NM_172217	3603	IL16	interleukin 16 (lymphocyte chemoattractant factor)
NM_199040	11163	NUDT4	nudix (nucleoside diphosphate linked moiety X)-type motif 4
NM_005461	9935	MAFB	v-maf musculoaponeurotic fibrosarcoma oncogene homolog B (avian)
NM_012296	9846	GAB2	GRB2-associated binding protein 2
NM_030925	81617	CAB39L	calcium binding protein 39-like
AF249277	400410	ST20	suppressor of tumorigenicity 20
NM_002118	3109	HLA-DMB	major histocompatibility complex, class II, DM beta
NM_001276	1116	CHI3L1	chitinase 3-like 1 (cartilage glycoprotein-39)
NM_001334	1519	CTSO	cathepsin O
NM_014153	29066	ZC3H7A	zinc finger CCCH-type containing 7A
NM_006995	10385	BTN2A2	butyrophilin, subfamily 2, member A2
NM_018639	55884	WSB2	WD repeat and SOCS box-containing 2
NM_175617	4493	MT1E	metallothionein 1E
NM_004688	9111	NMI	N-myc (and STAT) interactor
NM_001066	7133	TNFRSF1B	tumor necrosis factor receptor superfamily, member 1B
NM_003978	9051	PSTPIP1	proline-serine-threonine phosphatase interacting protein 1
NM_003900	8878	SQSTM1	sequestosome 1
NM_002355	4074	M6PR	mannose-6-phosphate receptor (cation dependent)
NM_003929	8934	RAB7L1	RAB7, member RAS oncogene family-like 1
NM_003299	7184	HSP90B1	heat shock protein 90kDa beta (Grp94), member 1
NM_015510	25979	DHRS7B	dehydrogenase/reductase (SDR family) member 7B
NM_021194	7779	SLC30A1	solute carrier family 30 (zinc transporter), member 1
NM_000958	5734	PTGER4	prostaglandin E receptor 4 (subtype EP4)
NM_021136	6252	RTN1	reticulon 1
NM_007229	11252	PACSIN2	protein kinase C and casein kinase substrate in neurons 2
NM_001375	1777	DNASE2	deoxyribonuclease II, lysosomal
NM_000876	3482	IGF2R	insulin-like growth factor 2 receptor
NM_022750	64761	PARP12	poly (ADP-ribose) polymerase family, member 12
NM_014435	27163	NAAA	N-acylethanolamine acid amidase
NM_015194	4642	MYO1D	myosin ID
NM_016602	2826	CCR10	chemokine (C-C motif) receptor 10
NM_002356	4082	MARCKS	myristoylated alanine-rich protein kinase C substrate
NM_013313	29799	YPEL1	yippee-like 1 (Drosophila)
NM_001006109	56941	C3orf37	chromosome 3 open reading frame 37
NM_025079	80149	ZC3H12A	zinc finger CCCH-type containing 12A
NM_006441	10588	MTHFS	5,10-methenyltetrahydrofolate synthetase (5-formyltetrahydrofolate cyclo-ligase)
NM_002231	3732	CD82	CD82 molecule
NM_020177	56929	FEM1C	fem-1 homolog c (C. elegans)
NM_003454	7752	ZNF200	zinc finger protein 200
NM_002298	3936	LCP1	lymphocyte cytosolic protein 1 (L-plastin)
NM_002527	4908	NTF3	neurotrophin 3
NM_003219	7015	TERT	telomerase reverse transcriptase
NM_019107	56005	C19orf10	chromosome 19 open reading frame 10
NM_212481	10865	ARID5A	AT rich interactive domain 5A (MRF1-like)

Acession Number	Entrez Id	Symbol	Description
NM_017996	55070	DET1	de-etiolated homolog 1 (Arabidopsis)
NM_006277	50618	ITSN2	intersectin 2
NM_012198	25801	GCA	grancalcin, EF-hand calcium binding protein
NM_032121	84061	MAGT1	magnesium transporter 1
NM_004090	1845	DUSP3	dual specificity phosphatase 3
NM_002068	2769	GNA15	guanine nucleotide binding protein (G protein), alpha 15 (Gq class)
NM_024324	79174	CRELD2	cysteine-rich with EGF-like domains 2
NM_002416	4283	CXCL9	chemokine (C-X-C motif) ligand 9
NM_006389	10525	HYOU1	hypoxia up-regulated 1
NM_017899	54997	TESC	tescalcin
NM_001693	526	ATP6V1B2	ATPase, H+ transporting, lysosomal 56/58kDa, V1 subunit B2
NM_007161	7940	LST1	leukocyte specific transcript 1
NM_003190	6892	TAPBP	TAP binding protein (tapasin)
NM_004059	883	CCBL1	cysteine conjugate-beta lyase, cytoplasmic
NM_016548	51280	GOLM1	golgi membrane protein 1
NM_012214	11320	MGAT4A	mannosyl (alpha-1,3-)-glycoprotein beta-1,4-N-acetylglucosaminyltransferase, isozyme A
NM_024310	79156	PLEKHF1	pleckstrin homology domain containing, family F (with FYVE domain) member 1
NM_147780	1508	CTSB	cathepsin B
NM_002985	6352	CCL5	chemokine (C-C motif) ligand 5
NM_015441	25903	OLFML2B	olfactomedin-like 2B
NM_014372	26994	RNF11	ring finger protein 11
NM_144977	163486	DENND1B	DENN/MADD domain containing 1B
AB029551	23429	RYBP	RING1 and YY1 binding protein
NM_002372	4124	MAN2A1	mannosidase, alpha, class 2A, member 1
NM_014923	22862	FNDC3A	fibronectin type III domain containing 3A
NM_005966	4664	NAB1	NGFI-A binding protein 1 (EGR1 binding protein 1)
NM_014624	6277	S100A6	S100 calcium binding protein A6
NM_005026	5293	PIK3CD	phosphoinositide-3-kinase, catalytic, delta polypeptide
NM_001659	377	ARF3	ADP-ribosylation factor 3
NM_139265	30844	EHD4	EH-domain containing 4
NM_000521	3074	HEXB	hexosaminidase B (beta polypeptide)
NM_006417	10561	IFI44	interferon-induced protein 44
NM_002528	4913	NTHL1	nth endonuclease III-like 1 (E. coli)
NM_001408	1952	CELSR2	cadherin, EGF LAG seven-pass G-type receptor 2 (flamingo homolog, Drosophila)
NM_022365	64215	DNAJC1	DnaJ (Hsp40) homolog, subfamily C, member 1
NM_015567	26050	SLITRK5	SLIT and NTRK-like family, member 5
NM_006623	26227	PHGDH	phosphoglycerate dehydrogenase
NM_016289	51719	CAB39	calcium binding protein 39
NM_006226	5334	PLCL1	phospholipase C-like 1
NM_014795	9839	ZEB2	zinc finger E-box binding homeobox 2
NM_003442	7702	ZNF143	zinc finger protein 143
NM_000487	410	ARSA	arylsulfatase A
NM_013296	29899	GPSM2	G-protein signaling modulator 2
NM_001565	3627	CXCL10	chemokine (C-X-C motif) ligand 10
NM_006892	1789	DNMT3B	DNA (cytosine-5-)-methyltransferase 3 beta

Bibliography

Alberts, B., Bray, Dennis, Hopkin, K., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2005). *Lehrbuch der Molekularen Zellbiologie*. Wiley-VCH Verlag GmbH & Co. KGaA, 3rd edition.

Alizadeh, a., Eisen, M., Davis, R. E., Ma, C., Sabet, H., Tran, T., Powell, J. I., Yang, L., Marti, G. E., Moore, D. T., Hudson, J. R., Chan, W. C., Greiner, T., Weisenburger, D., Armitage, J. O., Lossos, I., Levy, R., Botstein, D., Brown, P. O., and Staudt, L. M. (1999). The lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. In *Cold Spring Harbor symposia on quantitative biology*, volume 64, pages 71–78.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**(6769), 503–11.

Baron, B., Nucifora, G., McCabe, N., Espinosa, R., Le Beau, M., and McKeithan, T. (1993). Identification of the gene associated with the recurring chromosomal translocations t (3; 14)(q27; q32) and t (3; 22)(q27; q11) in B-cell lymphomas. *PNAS*, **90**(11), 5262–66.

Basso, K. and Dalla-Favera, R. (2010). BCL6: Master Regulator of the Germinal Center Reaction and Key Oncogene in B Cell Lymphomagenisis. *Advances in Immunology*, **105**, 193–210.

Basso, K., Saito, M., Sumazin, P., Margolin, A. a., Wang, K., Lim, W.-K., Kitagawa, Y., Schneider, C., Alvarez, M. J., Califano, A., and Dalla-Favera, R. (2010). Integrated biochemical and computational approach identifies BCL6 direct target genes controlling multiple pathways in normal germinal center B cells. *Blood*, **115**(5), 975–84.

Bea, S., Zettl, A., Wright, G., Salaverria, I., Jehn, P., Moreno, V., Burek, C., Ott, G., Puig, X., Yang, L., Lopez-Guillermo, A., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Gascoyne, R. D., Connors, J. M., Grogan, T. M., Braziel, R., Fisher, R. I., Smeland, E. B., Kvaloy, S., Holte, H., Delabie, J., Simon, R., Powell, J., Wilson, W. H., Jaffe, E. S., Montserrat, E., Muller-Hermelink, H.-K., Staudt, L. M., Campo, E., and Rosenwald, A. (2005). Diffuse large B-cell lymphoma subgroups have distinct genetic profiles that influence tumor biology and improve gene-expression-based survival prediction. *Blood*, **106**(9), 3183–90.

Beissbarth, T. and Speed, T. P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**(9), 1464–5.

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)*, **57**(1), 289–300.

Bentink, S., Wessendorf, S., Schwaenen, C., Rosolowski, M., Klapper, W., Rosenwald, A., Ott, G., Banham, A. H., Berger, H., Feller, A. C., Hansmann, M.-L., Hasenclever, D., Hummel, M., Lenze, D., Möller, P., Stuerzenhofecker, B., Loeffler, M., Truemper, L., Stein, H., Siebert, R., and Spang, R. (2008). Pathway activation patterns in diffuse large B-cell lymphomas. *Leukemia*, **22**(9), 1746–54.

Bernstein, K., Bleichert, F., Bean, J., Cross, F., and Baserga, S. (2007). Ribosome biogenesis is sensed at the start cell cycle checkpoint. *Molecular biology of the cell*, **18**(3), 953–964.

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., Joshi, M.-B., Harpole, D., Lancaster, J. M., Berchuck, A., Olson, J. A., Marks, J. R., Dressman, H. K., West, M., and Nevins, J. R. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**(7074), 353–7.

Blower, M. D., Feric, E., Weis, K., and Heald, R. (2007). Genome-wide analysis demonstrates conserved localization of messenger RNAs to mitotic microtubules. *The Journal of cell biology*, **179**(7), 1365–73.

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*, **19**(2), 185–93.

Boulesteix, A.-L., Porzelius, C., and Daumer, M. (2008). Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics*, **24**(15), 1698–706.

Bowman, A. W. (1984). An Alternative Method of Cross-Validation for the Smoothing of Density Estimates. *Biometrika*, **71**(2), 353.

Brunet, S. and Verlhac, M. H. (2011). Positioning to get out of meiosis: the asymmetry of division. *Human reproduction update*, **17**, 68–75.

Cattoretti, G., Pasqualucci, L., Ballon, G., Tam, W., Nandula, S. V., Shen, Q., Mo, T., Murty, V. V., and Dalla-Favera, R. (2005). Deregulated BCL6 expression recapitulates the pathogenesis of human diffuse large B cell lymphomas in mice. *Cancer cell*, **7**(5), 445–55.

Chapelle, O., Schölkopf, B., and Zien, A. (2006). *Semi-Supervised Learning*. The MIT Press, Cambridge, Massachusetts.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S., and Fodor, S. P. a. (1996). Accessing Genetic Information with High-Density DNA Arrays. *Science*, **274**(5287), 610–614.

Chia, W., Somers, W. G., and Wang, H. (2008). Drosophila neuroblast asymmetric divisions: cell cycle regulators, asymmetric protein localization, and tumorigenesis. *The Journal of cell biology*, **180**(2), 267–272.

Chow, J. C., Young, D. W., Golenbock, D. T., Christ, W. J., and Gusovsky, F. (1999). Toll-like receptor-4 mediates lipopolysaccharide-induced signal transduction. *The Journal of Biological Chemistry*, **274**(16), 10689–92.

Ci, W., Polo, J. M., Cerchietti, L., Shaknovich, R., Wang, L., Yang, S. N., Ye, K., Farinha, P., Horsman, D. E., Gascoyne, R. D., Elemento, O., and Melnick, A. (2009). The BCL6 transcriptional program features repression of multiple oncogenes in primary B cells and is deregulated in DLBCL. *Blood*, **113**(22), 5536–48.

Dave, S., Fu, K., Wright, G., Lam, L., Kluin, P., Boerma, E., Greiner, T., Weisenburger, D., Rosenwald, A., Ott, G., and Others (2006). Molecular diagnosis of Burkitt's lymphoma. *New England Journal of Medicine*, **354**(23), 2431–2442.

Dent, A. L., Schaffer, A. L., Yu, X., Allman, D., and Staudt, L. M. (1997). Control of Inflammation, Cytokine Expression, and Germinal Center Formation by BCL-6. *Science*, **276**(5312), 589–592.

Dhordain, P., Albagli, O., Lin, R. J., Ansieau, S., Quief, S., Leutz, A., Kerckaert, J. P., Evans, R. M., and Leprince, D. (1997). Corepressor SMRT binds the BTB/POZ repressing domain of the LAZ3/BCL6 oncoprotein. *PNAS*, **94**(20), 10762–7.

Dhordain, P., Lin, R. J., Quief, S., Lantoine, D., Kerckaert, J. P., Evans, R. M., and Albagli, O. (1998). The LAZ3(BCL-6) oncoprotein recruits a SMRT/mSIN3A/histone deacetylase containing complex to mediate transcriptional repression. *Nucleic acids research*, **26**(20), 4645–51.

Dubowy, J. and Macdonald, P. M. (1998). Localization of mRNAs to the oocyte is common in Drosophila ovaries. *Mechanisms of development*, **70**, 193–195.

Duncan, D., Prodduturi, N., and Zhang, B. (2010). WebGestalt2: an updated and expanded version of the Web-based Gene Set Analysis Toolkit. *BMC Bioinformatics*, **11**(Suppl 4), P10.

Edén, P., Ritz, C., Rose, C., Fernö, M. r., and Peterson, C. (2004). "Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *European journal of cancer*, **40**(12), 1837–1841.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, **30**(1), 207–10.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, **95**(25), 14863–14868.

Gebhard, C., Schwarzfischer, L., Pham, T.-H., Schilling, E., Klug, M., Andreesen, R., and Rehli, M. (2006). Genome-wide profiling of CpG methylation identifies novel targets of aberrant hypermethylation in myeloid leukemia. *Cancer research*, **66**(12), 6118–28.

Gentle, J. E. (2002). *Elements of Computational Statistics*. Springer Science+Business Media, New Yorck, USA.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. a., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439), 531–7.

Guo, G., Huss, M., Tong, G. Q., Wang, C., Li Sun, L., Clarke, N. D., and Robson, P. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental cell*, **18**(4), 675–85.

Guo, Y., Hastie, T., and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, **8**(1), 86–100.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Science+Business Media, New York, USA.

Hiiragi, T. and Solter, D. (2004). First cleavage plane of the mouse egg is not predetermined but defined by the topology of the two apposing pronuclei. *Nature*, **430**(6997), 360–4.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors (1977). *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, New York, USA.

Holt, C. E. and Bullock, S. L. (2009). Subcellular mRNA localization in animal cells and why it matters. *Science*, **326**(5957), 1212–1216.

Holtick, U., Vockerodt, M., Pinkert, D., Schoof, N., Stürzenhofecker, B., Kussebi, N., Lauber, K., Wesselborg, S., Löffler, D., Horn, F., Trümper, L., and Kube, D. (2005). STAT3 is essential for Hodgkin lymphoma cell proliferation and is a target of tyrphostin AG17 which confers sensitization for apoptosis. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K*, **19**(6), 936–44.

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*, **18 Suppl 1**(1997), S96–104.

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*, **2**(1), 3.

Hummel, M., Bentink, S., Berger, H., Klapper, W., Wessendorf, S., Barth, T., Bernd, H., Cogliatti, S., Dierlamm, J., Feller, A., and Others (2006). A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *N Engl J Med*, **354**(23), 2419.

Hüsemann, Y., Geigl, J. B., Schubert, F., Musiani, P., Meyer, M., Burghart, E., Forni, G., Eils, R., Fehm, T., Riethmüller, G., and Klein, C. A. (2008). Systemic spread is an early step in breast cancer. *Cancer Cell*, **13**(1), 58–68.

Huynh, K. D. and Bardwell, V. J. (1998). The BCL-6 POZ domain and other POZ domains interact with the co-repressors N-CoR and SMRT. *Oncogene*, **17**(19), 2473–84.

Huynh, K. D., Fischle, W., Verdin, E., and Bardwell, V. J. (2000). BCoR, a novel corepressor involved in BCL-6 repression. *Genes & development*, **14**(14), 1810–23.

IARC (2008). *WHO Classification of Tumors of Haematopoietic and Lymphoid Tissues*. WHO Press, Lyon, France, 4th edition.

Iqbal, J., Greiner, T. C., Patel, K., Dave, B. J., Smith, L., Ji, J., Wright, G., Sanger, W. G., Pickering, D. L., Jain, S., Horsman, D. E., Shen, Y., Fu, K., Weisenburger, D. D., Hans, C. P., Campo, E., Gascoyne, R. D., Rosenwald, a., Jaffe, E. S., Delabie, J., Rimsza, L., Ott, G., Müller-Hermelink, H. K., Connors, J. M., Vose, J. M., McKeithan, T., Staudt, L. M., and Chan, W. C. (2007). Distinctive patterns of BCL6 molecular alterations and their functional consequences in different subgroups of diffuse large B-cell lymphoma. *Leukemia*, **21**(11), 2332–43.

Irizarry, R. a. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*, **31**(4), 15e–15.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**(5644), 449–453.

Jares, P. (2006). DNA microarray applications in functional genomics. *Ultrastructural pathology*, **30**(3), 209–19.

Jonckheere, A. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika*, **41**(1), 133–145.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, New Yorck.

Kelly, S. J. (1977). Studies of the developmental potential of 4- and 8-cell stage mouse blastomeres. *J. Exp. Zool.*, **200**(3), 365–376.

King, M. L. (1995). mRNA localization during frog oogenesis. In H. D. Lipshitz, editor, *Localized RNAs*, pages 137–148. R.G. Landes, Austin, Texas, USA.

King, M. L., Messitt, T. J., and Mowry, K. L. (2005). Putting RNAs in the right place at the right time: RNA localization in the frog oocyte. *Biol. Cell*, **97**(1), 19–33.

Kloc, M., Jaglarz, M., Dougherty, M., Stewart, M. D., Nel-Themaat, L., and Bilinski, S. (2008). Mouse early oocytes are transiently polar: three-dimensional and ultrastructural analysis. *Experimental cell research*, **314**(17), 3245–3254.

Kurotaki, Y., Hatta, K., Nakao, K., Nabeshima, Y.-I., and Fujimori, T. (2007). Blastocyst axis is specified independently of early cell lineage but aligns with the ZP shape. *Science*, **316**(5825), 719–723.

Läuter, J., Horn, F., Rosołowski, M., and Glimm, E. (2009). High-dimensional data analysis: selection of variables, data compression and graphics–application to gene expression. *Biometrical journal*, **51**(2), 235–51.

Lechler, T. and Fuchs, E. (2005). Asymmetric cell divisions promote stratification and differentiation of mammalian skin. *Nature*, **437**(7056), 275–280.

Lécuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T. R., Tomancak, P., and Krause, H. M. (2007). Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*, **131**(1), 174–187.

Lemercier, C., Brocard, M.-P., Puvion-Dutilleul, F., Kao, H.-Y., Albagli, O., and Khochbin, S. (2002). Class II histone deacetylases are directly recruited by BCL6 transcriptional repressor. *The Journal of biological chemistry*, **277**(24), 22045–52.

Lin, H. (2008). Cell biology of stem cells: an enigma of asymmetry and self-renewal. *The Journal of cell biology*, **180**(2), 257–260.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology*, **14**(13), 1675–80.

Löffler, G. (2009). *Basiswissen Biochemie: mit Pathobiochemie*. Springer, Berlin, 7th edition.

Lottaz, C., Yang, X., Scheid, S., and Spang, R. (2006). OrderedList–a bioconductor package for detecting similarity in ordered gene lists. *Bioinformatics*, **22**(18), 2315–6.

Lu, L. J., Xia, Y., Paccanaro, A., Yu, H., and Gerstein, M. (2005). Assessing the limits of genomic data integration for predicting protein networks. *Genome Research*, **15**(7), 945–53.

Macara, I. G. and Mili, S. (2008). Polarity and differential inheritance–universal attributes of life? *Cell*, **135**(5), 801–812.

Maneck, M., Schrader, A., Kube, D., and Spang, R. (2011). Genomic data integration using guided clustering. *Bioinformatics (Oxford, England)*, **27**(16), 2231–2238.

Mili, S. and Macara, I. (2009). RNA localization and polarity: from A (PC) to Z (BP). *Trends in cell biology*, **19**(4), 156–164.

Mili, S., Moissoglu, K., and Macara, I. G. (2008). Genome-wide screen reveals APC-associated RNAs enriched in cell protrusions. *Nature*, **453**(7191), 115–119.

Motosugi, N., Bauer, T., Polanski, Z., Solter, D., and Hiiragi, T. (2005). Polarity of the mouse embryo is established at blastocyst and is not prepatterned. *Genes & development*, **19**(9), 1081–92.

Mount, D. W. (2004). *Bioinformatics - Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New Yorck, 2nd edition.

Mowry, K. L. and Cote, C. a. (1999). RNA sorting in Xenopus oocytes and embryos. *FASEB J*, **13**(3), 435–445.

Mullighan, C., Su, X., Zhang, J., Radtke, I., Phillips, L., Miller, C., Ma, J., Liu, W., Cheng, C., Schulman, B., and Others (2009). Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. *New England Journal of Medicine*, **360**(5), 470–480.

Nelson, N. (1999). Metal ion transporters and homeostasis. *The EMBO Journal*, **18**(16), 4361–4371.

Nieuwkoop, P. D. (1985). Inductive interactions in early amphibian development and their general nature. *J. Embryol. exp. Morph.*, **89 Suppl**, 333–347.

Niu, H., Cattoretti, G., and Dalla-Favera, R. (2003). BCL6 controls the expression of the B7-1/CD80 costimulatory receptor in germinal center B cells. *The Journal of experimental medicine*, **198**(2), 211–21.

Park, B. and Marron, J. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**(409), 66–72.

Pavlidis, P., Weston, J., Cai, J., and Noble, W. S. (2002). Learning gene functional classifications from multiple data types. *J Comput Biol*, **9**(2), 401–11.

Pease, a. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P., and Fodor, S. P. (1994). Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *PNAS*, **91**(11), 5022–6.

Pepling, M. E., Wilhelm, J. E., O'Hara, A. L., Gephardt, G. W., and Spradling, A. C. (2007). Mouse oocytes within germ cell cysts and primordial follicles contain a Balbiani body. *PNAS*, **104**(1), 187–192.

Polo, J. M., Juszczynski, P., Monti, S., Cerchietti, L., Ye, K., Greally, J. M., Shipp, M., and Melnick, A. (2007). Transcriptional signature with differential expression of BCL6 target genes accurately identifies BCL6-dependent diffuse large B cell lymphomas. *PNAS*, **104**(9), 3207–12.

Raj, A. and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, **135**(2), 216–26.

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, **4**(10), e309.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. a. (2000). Genome-wide location and function of DNA binding proteins. *Science*, **290**(5500), 2306–9.

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*, **23**(20), 2700–7.

Rosenblatt, M. (1971). Curve estimates. *The Annals of Mathematical Statistics*, **42**(6), 1815–42.

Rosenwald, A., Wright, G., Chan, W., Connors, J., Campo, E., Fisher, R., Gascoyne, R., Muller-Hermelink, H., Smeland, E., Giltnane, J., and Others (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med*, **346**(25), 1937–1947.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*, **20**(1), 53–65.

Rudemo, M. (1982). Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, **9**(2), 65–78.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**(5235), 467–70.

Scher, H. I. and Pantel, K. (2009). Bone marrow aspiration for disseminated tumor cell detection: a must-have test or is the jury still out? *Journal of clinical oncology*, **27**(10), 1531–3.

Schwandner, R., Dziarski, R., Wesche, H., Rothe, M., and Kirsching, C. J. (1999). Peptidoglycan- and Lipoteichoic Acid-induced Cell Activation Is Mediated by Toll-like Receptor 2. *J Biol Chem*, **274**(25), 17406–17409.

Shaffer, a. L., Yu, X., He, Y., Boldrick, J., Chan, E. P., and Staudt, L. M. (2000). BCL-6 represses genes that function in lymphocyte differentiation, inflammation, and cell cycle control. *Immunity*, **13**(2), 199–212.

Sheather, S. and Jones, M. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, **53**(3), 683–690.

Smolkin, M. and Ghosh, D. (2003). Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics*, **4**(1), 36.

Smyth, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**(1).

Smyth, G. K. and Speed, T. (2003). Normalization of cDNA microarray data. *Methods*, **31**(4), 265–73.

Sotiriou, C., Neo, S., McShane, L., Korn, E., Long, P., Jazaeri, A., Martiat, P., Fox, S., Harris, A., and Liu, E. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *PNAS*, **100**(18), 10393–10398.

Staunton, J. E., Slonim, D. K., Coller, H. a., Tamayo, P., Angelo, M. J., Park, J., Scherf, U., Lee, J. K., Reinhold, W. O., Weinstein, J. N., Mesirov, J. P., Lander, E. S., and Golub, T. R. (2001). Chemosensitivity prediction by transcriptional profiling. *PNAS*, **98**(19), 10787–10792.

Stryer, L. (1996). *Biochemie*. Spektrum Akad. Verl., Heidelberg, Berlin, Oxford, 4th edition.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. a., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, **102**(43), 15545–50.

Tarkowski, A. (1959). Experiments on the development of isolated blastomeres of mouse eggs. *Nature*, **184**, 1286–1287.

Thomas, G. (2000). An encore for ribosome biogenesis in the control of cell proliferation. *Nature cell biology*, **2**(5), E71–E72.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS*, **99**(10), 6567–72.

Torres-Padilla, M.-E., Parfitt, D.-E., Kouzarides, T., and Zernicka-Goetz, M. (2007). Histone arginine methylation regulates pluripotency in the early mouse embryo. *Nature*, **445**(7124), 214–218.

Tseng, G. C. and Wong, W. H. (2005). Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, **61**(1), 10–6.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. a. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(6871), 530–6.

Vermilyea, M. D., Maneck, M., Yoshida, N., Blochberger, I., Suzuki, E., Suzuki, T., Spang, R., Klein, C. A., and Perry, A. C. F. (2011). Transcriptome asymmetry within mouse zygotes but not between early embryonic sister blastomeres. *The EMBO journal*, **9**(4), 1841–51.

Vockerodt, M., Haier, B., Buttgereit, P., Tesch, H., and Kube, D. (2001). The Epstein-Barr virus latent membrane protein 1 induces interleukin-10 in Burkitt's lymphoma cells but not in Hodgkin's cells involving the p38/SAPK2 pathway. *Virology*, **280**(2), 183–98.

von Heydebreck, A., Huber, W., Poustka, A., and Vingron, M. (2001). Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics*, **17 Suppl 1**, S107–14.

Wang, A. M., Doyle, M. V., and Mark, D. F. (1989). Quantitation of mRNA by the Polymerase Chain Reaction. *PNAS*, **86**(24), 9717–9721.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lipshutz, R., Chee, M., and Lander, E. S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**(5366), 1077–82.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. a., Marks, J. R., and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS*, **98**(20), 11462–7.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, **30**(4), e15.

Ye, B., Lista, F., Lo Coco, F., Knowles, D., Offit, K., Chaganti, R., and Dalla-Favera, R. (1993a). Alterations of a zinc finger-encoding gene, BCL-6, in diffuse large-cell lymphoma. *Science*, **262**(5134), 747.

Ye, B., Cattoretti, G., Shen, Q., Zhang, J., Hawe, N., de Waard, R., Leung, C., Nouri-Shirazi, M., Orazi, A., Chaganti, R., and Others (1997). The BCL-6 proto-oncogene controls germinal-centre formation and Th2-type inflammation. *nature genetics*, **16**(2), 161–170.

Ye, B. H., Rao, P. H., Chaganti, R. S., and Dalla-Favera, R. (1993b). Cloning of bcl-6, the locus involved in chromosome translocations affecting band 3q27 in B-cell lymphoma. *Cancer research*, **53**(12), 2732–5.

Statement of Authorship

I herewith declare in lieu of an oath that I have produced this thesis independently and without using any other than the aids listed. Any thoughts directly or indirectly taken from somebody else's sources are made discernible as such.