TRANSCRIPTOME ANALYSIS IN PRETERM INFANTS DEVELOPING BRONCHOPULMONARY DYSPLASIA

Data processing and statistical analysis of microarray data

Inauguraldissertation
zur Erlangung des Grades eines Doktors der Humanbiologie
des Fachbereichs Medizin
der Justus-Liebig-Universität Gießen

vorgelegt von Windhorst (geb. Höland), Anita Cornelia
aus Dessau

Gießen 2015

Aus dem Institut für Medizinische Mikrobiologie,

Direktor: Prof. Dr. Trinad Chakraborty

Betreuer: Dr. med. Hamid Hossain

1. Gutachter: Prof. Dr. Trinad Chakraborty
2. Gutachter: Prof. Dr. Florian Wagenlehner

Tag der Disputation: 22. Oktober 2015

Writing a book is an adventure. To begin with, it is a toy, and an amusement; then it becomes a mistress, and then it becomes a master, and then a tyrant. The last phase is that just as you are about to be reconciled to your servitude, you kill the monster, and fling him about to the public.

(Churchill, 1949)

Content

1 Introduction

Preterm birth, defined as birth before 37th week of gestational age, accounts for 35% of all neonatal death worldwide. Preterm birth increases the risk for long-term complications such as visual and hearing impairment and chronic lung diseases. Bronchopulmonary dysplasia (BPD) is one of the most common chronic lung diseases and a major contributor to the morbidity of preterm infants, especially in very low birth weight infants (Walsh et al., 2006). Additionally, impairment of the cardiovascular system, neuro-development and behavior have been reported in preterm infants (Blencowe et al., 2013).

Preterm birth is also associated with increased mortality mainly caused by infections of the preterm infant. The reduced innate and adaptive immunity of the immature immune systems of preterm infant weakens its ability to fight bacteria and detect viruses in cells in comparison with term infants. Preterm birth is often caused by intrauterine inflammation due to bacterial infection. This prenatal exposure to inflammation can result in a simplified lung structure, which increases the risk for the BPD development in neonates in addition to the immature lung epithelium resulting from a birth before the lung development could be completed (Melville and Moss, 2013).

Further, the preterm lung faces stretch and oxidant injuries, which may lead to impaired development of the lung and to BPD (Jobe, 2006; Speer, 2003). Not only ventilatory support, and oxygen levels may lead to BPD, but also sepsis as a pro-inflammatory modulator, anti-inflammatory corticosteroids and starvation may lead to BPD (Jobe, 2006). Recent studies show that the extended mechanical ventilation is associated with a decreased number of CD4-T-cells (Ballabh et al., 2003) with changes in cytokine levels (Bose et al., 2013; Köksal et al., 2012), which indicate that sustained systemic inflammation may be a risk factor for developing chronic lung diseases (Melville and Moss, 2013).

Over the last years, research on BPD started to focus on the transcriptome of preterm infants using microarrays. In 2007, Cohen et al. (2007) compared cord blood tissue of preterm infants who later develop BPD or infants who not develop BPD. Kompass et al. (2010) used animal models to research the alterations in gene expression profiles in reaction to ventilation of the lung. Bhattacharya et al. (2012) used tissue of the lungs obtained at autopsy of preterm infants with BPD and control tissue from preterm infants without BPD. And recently, Pietrzyk et al. (2013) investigated the gene expression profiles of preterm infants and their alterations 5, 14, and 28 days after birth and potential pathways associated with the gene profiles. Still missing are gene expression profiles from time of birth, which may give a hint on processes during birth or general susceptibility of the preterm infants that lead to the development of BPD. Even more scarce are studies on the development of mild BPD[1] (as defined by Jobe, 2006), a form of BPD with which neonates are less or not mechanical ventilated and have shorter periods of oxygen treatment.

Transcriptome analysis has become more and more important in the development of diagnostics and the search for biomarkers. Microarray analysis bears the advantage to analyze a great number of transcripts simultaneously and facilitate a better understanding of the role and involvement of specific transcripts. It may provide hints to biological mechanisms resulting from the gene expression profile or underlying mechanisms resulting in the observed gene expression profiles. To detect meaningful gene expression it is important to assure data quality in the preprocessing of microarray data.

In this doctoral thesis methods in microarray transcriptome analysis are presented in a more detailed manner. Especially preprocessing of microarray data obtained from CodeLink™ Bioarrays is reviewed (chapter 2). A workflow for data preprocessing and preparation is developed for the open-source platform in R (r-project.com) with the help of R- and Bioconductor packages. It will be shown, that preprocessing plays an important role in the discovery and correction of handling errors, e.g., mislabeled arrays.

Subsequently, methods in the statistical analysis of microarrays are introduced (chapter 3), especially differential gene expression analysis and cluster analysis methods. But also approaches to predict clinical outcomes with microarray data are suggested. Furthermore, functional gene annotation analysis methods are applied with the help of the Database for Annotation, Visualization, and Integrated Discovery (DAVID) and Ingenuity Pathway Analysis (IPA). These platforms use gene sets and add functional information to give hints

[1] With definition of the new BPD by Jobe (2006) BPD is in very preterm infants (born before the 32 weeks gestational age) diagnosed if the infants are treated with oxygen over 21% for at least 28 days. BPD is then divided into three severity grades: mild BPD, if infants breath rom air at 36 weeks postmenstrual age or discharge; moderate BPD, if the need for oxygen remains, but below 30% oxygen at 36 weeks postmenstrual age or discharge; or severe BPD, if more or equal to 30% oxygen and/or positive pressure, either ventilation or continuous positive airway pressure, is required at 36 weeks postmenstrual age (Jobe, 2006).

in regard to underlying diseases and biological functions, as well as putative upstream regulators.

Combining the described methods of data preprocessing, statistical analysis including hierarchical clustering, differential gene expression, and functional analysis of the obtained set of transcripts a comprehensive evaluation procedure is assembled and applied to a data set of 22 preterm infants (chapter 4). Cord blood of preterm infants born before the 32nd week of gestation was analyzed showing that a differentiation at birth between preterm infants who do not develop BPD and infants who develop different stages of BPD is possible. Here, first a supervised approach, using different BPD groups, was followed and secondly a semi-supervised approach with stratification for gestational age was applied in order to select transcripts whose expression is correlated with mechanical ventilation and/or oxygen support.

2 Preprocessing of microarray data

There is a wide variety of methods to analyze gene expression. One method is based on high-density microarrays; these contain more than 10’000 spots[2] per square centimeter (Lorkowski and Cullen, 2003). In this study, CodeLink^{TM} Bioarrays (GE Healthcare) are used. These microarrays contain probes consisting of specific 30-mer oligonucleotides which represent about 10’000 human genes, their transcripts, and expressed sequence tags.

The microarray analysis technique is based on nucleic acid strand binding between complementary sequences on the microarray and transcript sequences to be sampled. Messenger ribonucleic acid (mRNA) is obtained from tissue under investigation or whole blood. Nucleic acid strand bind strongly and specifically to each other using Watson-Crick base pairing[3]; the specificity and affinity of this binding is reduced by mismatching base pairs until the hybridization is completely prevented. To label mRNA strands, they are usually reverse transcribed to produce complementary deoxyribonucleic acid (cDNA) strands. For detection and quantification purposes they are labeled with fluorescent dye during this process. The cDNA is hybridized on microarrays, where the DNA strands bind complementary to probes attached to the microarray surface. The higher the amount of a specific transcript, the higher is the detection signal obtained from the probe (Lorkowski and Cullen, 2003).

Two types of microarrays can be differentiated: single color microarrays, on which only one sample is hybridized and two color microarrays, on which two samples labeled with two different fluorophores (usually Cy3^{TM} and Cy5^{TM}[4]) are hybridized. Statistical procedures and data transformations differ for these types. In this study, single color arrays were used.

A microarray detection signal consists basically of two parts: (1) the actual signal which indicates the amount of specific transcripts in the sample, and (2) the experimental (or

[2] Spots contain defined sequences of transcripts and genes as probes for the detection of gene expression.

[3] Watson-Crick base pairing is the binding of cytosine to guanine and thymine to adenine in DNA double strands or adenine-uracil and guanine-cytosine in RNA strands by hydrogen bonds.

[4] Cy3^{TM} and Cy5^{TM} are water-soluble cyanine dyes with an absorption wavelength of 552 nm and 667 nm respectively and an emission wavelength of 568 nm and 667 nm respectively (Lorkowski and Cullen, 2003).

background) noise e.g., introduced by microarray handling. The aim of data preprocessing in microarray experiments is the reduction of experimental noise to improve the signal quality.

2.1 Review of methods

The quality of a microarray experiment is determined in every step of the experiment starting with the design of the experiment, selecting and obtaining probe material, extracting RNA, and finally the hybridization of the RNA to microarrays. Each step of the experiment can introduce bias and lead to false positive or false negative results.

Eady and colleagues (2005) investigated healthy donors and observed that gene transcription profiles in the blood of different donors varied with sex, age, body mass index and the presence of varying proportions of different leukocyte proportions. They also found gene transcription profiles within a single donor to be comparatively stable over time. This experiment emphasizes the importance of a careful design of microarray experiments. To solve problems regarding probe material and sampling errors, standardization of experimental conditions, pooling of multiple samples, and multiple replicates of the experiment need to be considered and controlled (Murphy, 2002).

In addition to errors introduced by the experimental setup and the sampling, also the microarray itself can introduce errors, the so called experimental noise (Tu et al., 2002). Variation in gene expression may therefore originate in (1) pre-hybridization steps: probe, target, and sample preparation; in (2) the hybridization process and readout step: background effects, and effects from image processing (Schuchhardt et al., 2000; Tu et al., 2002). These aspects are addressed in data preprocessing in background correction (chapter 2.1.1) and normalization steps (chapter 2.1.2). Furthermore, missing values due to technical problems, the so called “missing completely at random”, or measurements that are not reliable or obtainable in some cases, the so called “missing at random” (Aittokallio, 2010), or due to data handling are addressed in chapter 2.1.3. Low expressed transcripts are very susceptible to data problems introduced by background correction, which is why they need to be addressed in data preprocessing (chapter 2.1.4). Along with missing and low expressed values, also extreme values or outlier values (chapter 2.1.5) must be addressed in the preprocessing of microarray data (Liu et al., 2003). At last, methods are reviewed to detect outliers in samples or technical replicates in order to reduce noise affecting the whole microarray and not only single probes (chapter 2.1.6).

2.1.1 Background correction

In order to compare different microarrays used in a microarray experiment and to reduce experimental noise, three aspects of standardization need to be considered (Reimers, 2010): (1) variance adjustment by binary logarithm transformation (this aspect is not further discussed in this work), (2) adjustment of the central tendency by background correction, and (3) adjustment of distributions of signal intensities by normalization[5]. The positive impact of background correction and data normalization on statistical analyses has been shown, e.g., on clustering by Freyhult et al. (2010).

The term “background correction” refers to the “removal of ambient, non-specific signal from the total intensity […]” (Ritchie et al., 2007) for each microarray. Non-specific hybridization and noise in the optical detection system occurs frequently, e.g., different amount of cDNA were hybridized on different microarrays, background corrections adjusts the signal to get accurate measurements of specific hybridization (Huber et al., 2005).

Subtract. The local background is defined by the periphery of pixel of a spot (Applied Microarrays, 2013). A straightforward approach for background correction is the subtraction of the background from the measured signal intensity. For CodeLink Bioarrays the procedure recommended by the manufacturer is to calculate the mean intensity of pixels within a spot and subtract the median intensity of the local background. According to Ritchie et al. (2007) most image analysis software proceed in this manner. But this method of background correction leads to negative values, whenever the background intensity exceeds the spot intensity; this in turn leads to missing log-transformed intensities. Log-transformation further leads to a higher variability of low intensity values, also known as “fanning” problem. The issues of low intensity and missing values are addressed again in the chapters 2.1.4 and 2.1.3.

[5] The term normalization here refers to the adjustment of distributions of measurements. Sometimes normalization refers to all steps which aim to make microarrays of different samples comparable as seen in Reimers (2010).

Half. To avoid the generation of negative values, it is also possible to replace negative values or values smaller than 0.5 after background subtraction by 0.5. This method is named for its constant, half. Using this method, variance in intensities near the background is reduced, which can be observed in Figure 2-1, presented later in this chapter.

Normexp. A third possibility for background correction for CodeLink Bioarrays as provided in the Bioconductor package *codelink* (Diez et al., 2007) is the method normexp as first introduced by Ritchie et al. (2007) for two-color microarrays. Normexp is an abbreviation for normal exponential convolution model. The observed intensities are modelled by fitting an exponential distribution to the foreground signals and a normal distribution to the background noise where parameter are estimated using a saddle-point approximation. This method is based on background correction by robust multiarray analysis (RMA) developed for Affymetrix GeneChip system by Irizarry et al. (2003), using a global distribution of probe intensities based on empirical observations of global intensities (Bolstad et al., 2005). Normexp was enhanced by Silver (2009) using an exact maximum-likelihood estimation. Normexp can also be combined with a small offset, a constant that shifts the whole distribution to the left and thereby stabilizes the variance of small variables.

Comparison of background correction methods

Ritchie et al. (2007) compared different background correction methods, e.g., standard, normexp and normexp with added offset, for two-color cDNA microarrays. Other methods they compared were: *Kooperberg* based on an empirical Bayes model; *Edwards*, which uses a threshold to decide, whether background is subtracted or estimated by a smooth monotonic function; *vsn*, a variance stabilizing method, which uses an arcsin transformation instead of the logarithm, and is so able to deal with negative values; and *morph* used in the Spot and GenePix software. The focus of the following short review lies on the evaluation of the widely-used normexp and standard methods, as they are further available for oligonucleotide CodeLink Bioarrays.

Ritchie et al. (2007) compared the standard background subtraction method, that may produce negative values, with methods which strictly produce positive corrected intensity values. With the help of three data sets, a spike-in[6], mixture[7], and a quality control study data set, they assessed the precision, bias, and performance in differential expression analysis of the background correction methods.

They estimated precision by using the mixture data set; in short, intensity measures are expected to follow the mixing patterns; a residual standard deviation gives an estimate for the precision of the returned measurements. The fanning effect of background correction can be seen in this setting; for low intensities the precision is generally higher than for high intensities, whereas without background correction this effect is reversed, with very low variances in low and high intensities. Bias is estimated with the spike-in data set which provides an array of true fold changes, which then are compared to the estimated fold changes. The findings are confirmed with the mixture data set. Bias is summarized by the mean absolute deviation of the estimated $\log_2$-fold changes from the true $\log_2$-fold changes.

For the assessment of the ability to detect differentially expressed genes with significance analysis of microarrays (SAM) regularized t-statistic and linear models for microarray analysis (LIMMA) empirical Bayes moderated t-statistics Ritchie et al. (2007) used the mixture experiment data set. With varying mixtures the altitude of the fold changes are expected to differ, but the set of differentially expressed genes should stay constant.

The investigators observed that a trade-off between bias and precision must be made; methods, i.e., normexp with offset or no background correction, performing well in terms of precision, show a higher bias. The standard background subtraction shows a very low precision in low intensities. With normexp this effect is not quite as pronounced, but it shows lower precision in low intensities as well. The method normexp together with a fixed offset value is able to reverse this trend; it shows similar variances in low and high intensities, also middle intensities show lower variances transformed with this method than with other methods. Even lower variances can only be observed without background correction. Standard method of background subtraction performs worst resulting in a high number of false positives in differential gene expression analysis. Again, the alternative of no background correction performs poorly in SAM differential expression analysis. Best

[6] A spike-in data set contains reference control RNA, which is added prior to labelling to produce known fold changes.

[7] In a mixture data set mRNA from different reference samples or cell lines in known relative concentrations are compared.

results in regard to detection of differentially expressed genes can be seen with a variance stabilizing method, e.g., normexp + offset.

Application in CodeLink Bioarrays

For background correction, the Bioconductor package *codelink* by Diez et al. (2007), no background correction, subtraction, half and normexp can be chosen. It also provides graphics, i.e., MA-plots[8], density plots, and image plots, to visually inspect the effects of background correction and normalization.

In Figure 2-1, density of 20 randomly sampled CodeLink human whole genome Bioarrays are shown before and after background correction with the methods subtract, normexp, and half.

It can be seen that raw data (Figure 2-1, black lines) have higher means and lower variance than after background correction. Measures of central tendency of raw data distribution show a high variability and are therefore difficult to compare due to a priori differences in the distributions. Ideally, distributions of signal intensities resemble each other closely. Differences in gene expression between the different treatment groups can be expected to be compensated by the great number of unchanged gene expressions.

After background corrections measurements of central tendency, e.g., modal value resemble each other more closely. As expected an overall shift of the distribution to the left can be observed. With regard to dispersion it can be observed, that variance of the methods subtract (Figure 2-1, red lines) and half (Figure 2-1, blue lines) are higher; after half background correction a second peak, where the constant 0.5 was set can be observed. In subtract these values are smaller than 0.5 or missing values, as they are negative before log-transformation. The method normexp (Figure 2-1, green dotted lines) results in distributions with lower variance and more comparable means.

[8] MA-plots plot the difference in measured signal intensity of each probe between two arrays versus the average of the two arrays. For a pair of arrays i and j, and the k-th probe these are calculated as follows: $M = \log_2(x_{ki}/x_{kj})$ or $M= \log_2(x_{ki}) - \log_2(x_{kj})$, $A = ½ \log_2(x_{ki}*x_{kj})$ or $A= (\log_2(x_{ki}) + \log_2(x_{kj}))/2$ (Bolstad et al., 2003). M and A are mnemiotic for “minus” and “add”. For more than two arrays M stands for the difference between the median intensity of this probe minus the value for this probe and array. It then becomes a “Median (difference) versus Average”-plot.

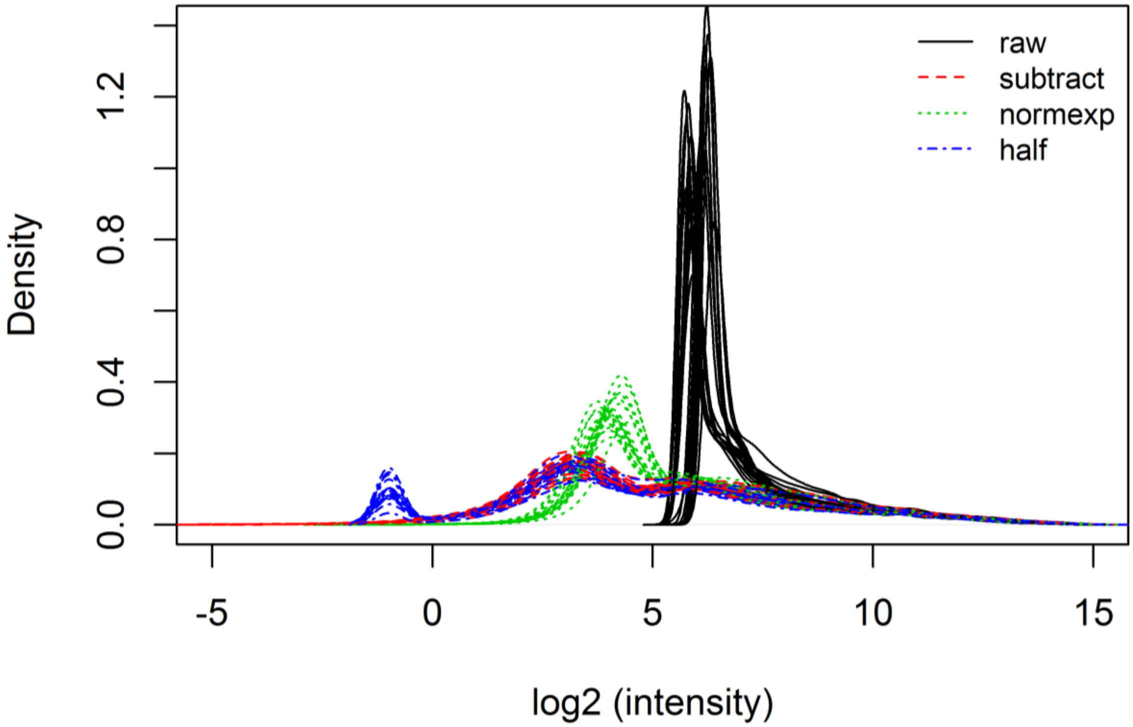

Figure 2-1 Background correction effect on density distributions of 20 CodeLink human whole genome Bioarrays.

In black solid lines raw data distributions are shown, they have a higher mean and lower dispersion in data; after background subtraction (in red), all arrays show a similar distribution, with the mean shifted to the left and a higher variance. The low density at all intensities can be explained by missing values due to values below zero before log-transformation. The same can be observed with an offset of 0.5 for values below zero, but a second mode, where the offset was defined. In normexp background correction (in green) a smaller variance and smaller shift to the left can be seen.

It can be seen that background correction in microarray experiments is necessary; however, background subtraction bears some difficulties, i.e., the generation of missing values and a high variance in expression of low expressed genes. Variance stabilizing method for background correction, e.g., “normexp”, lead to a higher accuracy in returned intensities.

2.1.2 Normalization

The term “normalization” refers to the data transformation step that “[…] adjusts the individual hybridization intensities to balance them appropriately so that meaningful biological comparisons can be made” (Quackenbush, 2002). In short, normalization is the step of data transformation that makes different arrays comparable (Bolstad et al., 2005; Schuchhardt et al., 2000).

With the background correction step, it is possible to reduce the measured intensities of the noise, which leads to shifts in central tendency of the signals of a microarray. But still differences in the distribution of signal intensities remain, making it still impossible to interpret the data. Normalization strategies are proposed to minimize the influence of noise on the signal due to changes in measured intensity between microarrays (Bolstad et al., 2003; Smyth and Speed, 2003; Wu et al., 2005). To achieve similar distributions, now distributions of a microarray are either compared to a baseline-array or are compared in a pairwise-manner. Normalization can only be achieved by comparing with another array, while background correction is based on the microarray distribution itself. Introduced and discussed are normalization methods based on (1) median, (2) local weighted regression (loess), (3) quantiles, and (4) briefly mentioned are two methods which are based on a subset of genes, e.g., housekeeping genes (Iset and Qspline).

Median. Median normalization refers to all methods, which transform measured intensities in a manner that all microarrays have the same median. Median normalization is one of the most common normalization methods for one-color arrays (Edwards, 2003) and is recommended by the manufacturer of CodeLink Bioarrays(Applied Microarrays, 2013). Therefore measured intensities of each microarray are divided by the median intensity, which results in a median of one in all microarrays (Wu et al., 2005). In this case the proposed normalization method is independent of other arrays and allows a comparison of arrays not preprocessed at the same time. The manufacturer of Affymetrix microarrays uses a different normalization method, but also proposes to normalize data in a manner that all arrays have the same median (Bolstad et al., 2003). Intensities are transformed using a baseline array, which usually is the array with the median that equals the median of medians of all microarrays.

Loess. Cyclic loess normalization is based on the idea of MA-plots, where the average log intensity A= $(\log_2(x_{ki}) + \log_2(x_{kj}))/2$ of two arrays or colors is compared with the log intensity ratio for the microarray i and j, and the k-th probe (Bolstad et al., 2003). It is an inter-microarray variant of locally weighted regression (loess) (Cleveland and Devlin, 1988) based normalization method, a local regression method. It estimates the intensity-dependent differences in a pair of microarrays, then uses loess smoothing to center the loess line to zero, and thus removes the pair wise differences. The cyclic loess normalization algorithm is applied in a pairwise manner to all microarrays in one or two iterative steps (Bolstad et al., 2003; Wu et al., 2005).

Quantile. Smyth and Speed (2003) propose a scale-normalization. Scale-normalization describes scaling[9] of a series of arrays, so that the spread of values and the median absolute deviation (MAD) of each array are the same (Smyth and Speed, 2003). Quantile normalization is based on a similar idea, particularly on the idea of Q-Q-plots, where two distributions are compared. The distributions are the same if the plot shows a straight diagonal line. This idea is projected on a higher dimensional level. For the example of two arrays are the intensities adjusted in a manner that generates a straight diagonal line. For each array, ranks are assigned to raw intensity values, the value for each intensity with the same rank are then substituted by the median value of intensities with this rank (Bolstad et al., 2003; Wu et al., 2005). Quantile normalization is nowadays adapted and also used for RNA sequencing technologies (Dillies et al. 2012). It then matches the distributions of gene counts across lanes.

Iset and Qspline. Invariant-set- (Iset) and quantile-spline-normalization (Qspline) by Workman et al. (2002) both are methods based on a baseline array approach and spline smoothing technique with a subset of genes of the array to reduce intensity-dependent differences of the arrays. Iset uses rank-invariant or so-called house-keeping genes with respect to the baseline-array. Qspline uses quantiles of ranked genes to estimate smoothing curves (Bolstad et al., 2003).

[9] Scaling refers to the division of a vector by its standard deviation. It can be combined with a centering step, where from the vector an average is subtracted. When scaling and centering by the arithmetic mean in normal distributed data is done, this process is called z-transformation and the vector then is standard normal distributed.

Assessing normalization methods

To evaluate normalization methods, Wu et al. (2005) examine the ability to reduce noise in a dataset and the ability to retain signal. They use a number of possibilities to examine the effectiveness of normalization methods: by noise reduction via MA-plot, spatial plot, coefficient of variation, correlation and variances in replicate arrays; by signal retention via the ability to predict a fixed number of known differentially regulated genes, or to reveal spike-in genes, overabundance of differentially express genes, or the cross-validation KNN classification error (Wu et al., 2005).

The investigators compared Median, Loess, Quantile, Iset, and Qspline normalization methods specifically for CodeLink Bioarrays by using replicate microarrays and/or positive control probes which are redundantly, i.e., six times, spotted on each CodeLink Bioarray and comparing the numbers of differentially expressed genes. For signal detection a minimum detection threshold is determined by the 80%-trimmed mean of negative control probes and the standard deviation of trimmed negative control probes. Signal reproducibility was assessed by using the number of differentially regulated genes detected by parametric and non-parametric statistical significance tests. They used two data sets: (1) a time course data set using 5 different durations of treatment and a control group on the CodeLink Uniset Rat I Bioarray, and (2) a control versus disease setup with patients with idiopathic pulmonary fibrosis on CodeLink Human I Bioarrays (Wu et al., 2005).

Bolstad and colleagues (2003) used data of a previously described study from Irizarry and colleagues (2003) where datasets are created using a dilution/mixture experiment and a spike-in experiment. The dilution and mixture experiment involved 5 RNA dilution levels and 3 proportions of mixtures of two tissue lines. In the spike-in experiment 11 different cRNA fragments were added at various concentrations. They compared Loess, Quantile, Median (of a baseline array) and Iset normalization methods (in addition to the not further discussed contrast based normalization method, which is similar to Loess normalization, but applies a smoothing transformation in addition) with respect to performance in reduction of obscuring variance without increasing bias (Bolstad et al., 2003). For the evaluation of variance reduction in Cyclic loess, Quantile, and the contrast based normalization method the dilution/mixture experiment is used, where the RNA for the arrays stems from a single source. The expression for each probe-set is calculated and variance and mean of this probe-set expression is calculated across all arrays of a dilution set and for each normalization method. The common source bears the advantage, that normalization methods can directly be compared by variance; the smaller the variance the better the method.

The Median normalization method recommended by the manufacturer shows poor results in the assessment of Wu et al. (2005), i.e., noise reduction in both examined datasets is improved considerably and consistently with Loess, and Qspline normalization. Quantile normalization also shows improvement in the control versus disease – data set, but not in the time course data set. Iset shows no improvement over Median normalization in both data sets. The authors come to the conclusion to best use the Loess or the Qspline normalization method for CodeLink Bioarrays. However, Loess normalization has the disadvantage that it cannot deal with missing values, which are e.g. created by subtract background correction, and thus cannot be paired with background subtraction.

Bolstad et al. (2003) show that normalization outperforms non-normalization in regard to intensity dependent differences between two arrays; distributions of differences between the same probe of two arrays vary around a median of 0 in normalized data, but are shifted in non-normalized data indicating a higher overall intensity level in one of the arrays. The normalization methods are able to reduce the variance across a single detection probe and variances of the distributions across microarrays of the signal intensities. Quantile normalization performs slightly better than Loess or Median normalization in terms of bias correction as assessed with a dilution data set. Ideally slopes near one would be reached; deviations from one give information about the bias or bias correction after normalization respectively. Again Quantile normalization together with Median normalization demonstrates good results.

To illustrate the effect of different normalization methods together with subtract and normexp background correction methods, densities of 20 randomly selected CodeLink human whole genome Bioarrays are presented (Figure 2-2). The normalization methods Median, Quantile, and Loess are incorporated in the *codelink* Bioconductor package by Diez et al. (2007) designed for preprocessing CodeLink Bioarrays. The effect of these methods on variances of single distributions and between arrays can be assessed in the following figure. Subtract and normexp background correction together with Median, Quantile, and in case of normexp Loess normalization is compared as well.

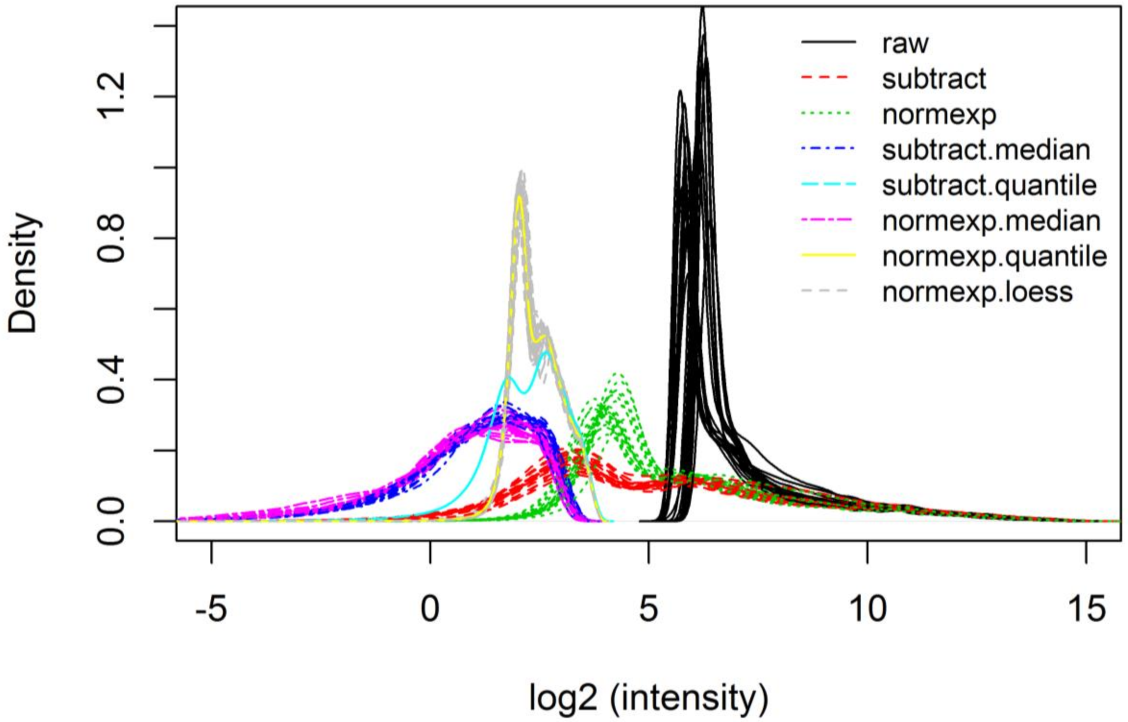

Figure 2-2 Normalization and background correction effects on density distributions of 20 CodeLink human whole genome Bioarrays.

Each color represents a set of microarrays processed with different background and normalization methods. Quantile normalization produces identical distributions; after subtract background correction (in magenta) variance is smaller than variance of distributions after median normalization (in blue); after normexp densities in loess (in grey) and quantile (yellow) normalizations are higher but resemble each other. Median normalization shows the same results after both background correction methods.

All normalization methods achieve a reduction of the in-between microarray variance when distributions of normalized data are compared with either the un-preprocessed raw data set (depicted in black) or the background corrected data set (depicted in red for background subtraction and in green for normexp).

Median normalization achieves similar distributions after both background correction methods (see Figure 2-2, magenta for normexp, and blue for subtract), with high variances between and within microarrays compared to Quantile and Loess, although variance between microarrays is still highly reduced compared with the raw data set (depicted in black lines). Quantile normalization (see Figure 2-2, cyan and yellow) after subtract produces per definition identical distributions and thus a very low variance between microarrays, additionally within microarray variances are reduced compared to the after background correction distributions (see in red and green respectively). Loess normalization produces distributions with a very small variance between microarrays (see grey lines), which equals the Quantile solution.

2.1.3 Missing values

The next steps filter probe sets, which cannot be interpreted due to (1) a high proportion of missing values, (2) expression values near or below the background (see 2.1.4), or (3) a high variability of the probe in a treatment group (see 2.1.4).

Missing values in microarray experiments occur due to various experimental reasons:

(1) Scratches, dust or other incidences may compromise parts of the microarray glass slide. (2) The hybridization or the signal spotting may be ineffective, leading to misshaped spots and then to the removal of the measurement. (3) The user manually removes probes after visual inspection of the hybridization image (Troyanskaya et al., 2001; Tuikkala et al., 2006). Values are also discarded because of (4) low expressed values, i.e., signals below background noise (see chapter 2.1.4), or if (5) outlier values occur within a treatment group (see chapter 2.1.5).

Dealing with missing values is important because their influence has been reported for multiple downstream analyses; e.g. clustering of transcripts is influenced by missing values (Brevern et al., 2004; Celton et al., 2010; Tuikkala et al., 2008), the removal of missing values improves the controlling of false positives and true positives in transcript prioritization or the detection of differentially expressed transcripts (Hua and Lai, 2007; Scheel et al., 2005). Additionally, the impact on the outcome of biological downstream analyses depends on the filtering and imputation methods used, and will therefore be discussed in more detail in the following chapters.

2.1.3.1 <u>Filtering</u>

To deal with missing values three typical approaches exist: (1) transcript filtering, (2) imputation, (3) replacement with a constant (Brevern et al., 2004; Tuikkala et al., 2008). In our analyses, we included two of these three approaches. The third method, replacement with a constant, is inferior to the other methods as demonstrated by the reviews described in the following sections.

So first, transcripts with high rate of missing values were filtered and then all remaining missing values were imputed. Troyanskaya et al. (2001), Tuikkalla et al. (2008), as well as Celton et al. (2010) estimate the imputation accuracy of different imputation methods with percentages of missing values ranging from 5% to 50%. They show that the higher the percentage of missing values is, the lower is the accuracy of the imputation method used. Scheel et al. (2005) also find an sample dependent impact on the performance of imputation methods in the detection of differentially expressed transcripts when using the statistical methods Student's t-test and SAM; in studies with sample sizes of 5 or 10 samples in each group, an elimination of transcripts with a high missing rate is as important as the in the following chapter evaluated imputation methods (Scheel et al., 2005).

Additionally, the accuracy of an imputation of missing values depends on the number of arrays used for the analysis: a high number of arrays rather tolerates a high missing rate in the data set (Celton et al., 2010; Scheel et al., 2005; Troyanskaya et al., 2001; Tuikkala et al., 2008).

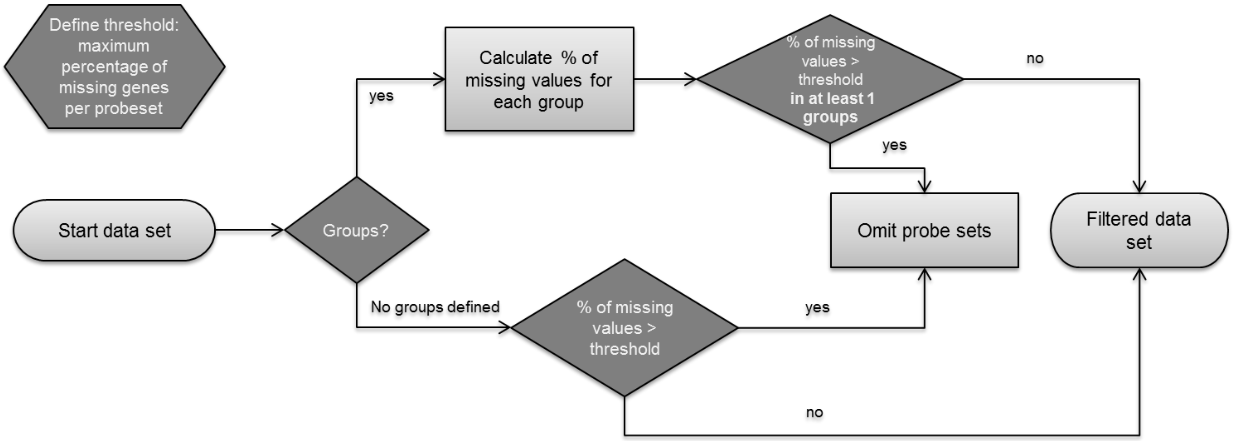
Define threshold: maximum percentage of missing genes per probeset
Calculate % of missing values for each group
% of missing values > threshold in at least 1 groups
no
yes
yes
Start data set
Groups?
Omit probe sets
Filtered data set
No groups defined
% of missing values > threshold
yes
no

Figure 2-3 Missing values filtering algorithm divided into analyses with and without prior consideration of treatment groups.
To start filtering a threshold for the maximum percentage of missing values per group or for all arrays must be given. Probe sets exceeding a certain amount of missing values in at least one group are discarded and not further analyzed.

As mentioned before, a certain percentage of missing values cannot be avoided in a data set of microarrays. To assure the accuracy of the data, all probe sets with a high amount of missing values are removed from further analyses. This step is sensible to *a priori* defined treatment groups[10], so it can be chosen whether to accept a certain amount of missing in the data set containing all microarrays, the *no-groups-scenario*, or if, in the case of the *groups-scenario*, a too high amount of missing values in one of the groups leads to an omission of the detection probe.

In the no-groups-scenario it is assured that a sufficient amount of data exists to interpret and analyze the gene expression. In the groups-scenario, one group is sufficient to eliminate this probe set from the scenario. With a high percentage of missing values in one group, an interpretation of the gene expression in this group in comparison to other groups is not adequate. The gene expression of this probe may not be estimated with sufficient accuracy.

2.1.3.2 Imputation

Imputation is the statistical estimation and substitution of missing values. Imputation is a partial step in data preparation as principal component analysis (PCA), hierarchical clustering, and other downstream analyses require complete data sets. Over the years a number of imputation methods have been developed and used.

In the beginnings of gene expression analyses, missing values were substituted using standard statistical procedures, such as the replacement with 0, in $\log_2$-transformed data, or the row- or gene expression-average of the remaining values of a transcript (Troyanskaya et al., 2001). Several workgroups describe and show the inferiority of these approaches, and more sophisticated, data structure and correlation structure considering, methods were developed (Aittokallio, 2010; Kim et al., 2004; Troyanskaya et al., 2001; Tuikkala et al., 2008). These methods include strategies based on Singular Value Decomposition (SVD) (Troyanskaya et al., 2001), weighted K-nearest neighbors (KNN) (Troyanskaya et al., 2001), the re-use of imputed transcript data sequentially in K-nearest neighbor method (SeqKNN) (Kim et al., 2004), expectation maximization (EM) or least square methods (Bø et al., 2004), local least squares (Kim et al., 2006), linear model based imputation (LinImp) (Scheel et al., 2005), semantic similarity in gene-ontology of

[10] The term “treatment group” refers to the factor analyzed in the study at hand; it can either refer to control vs disease, or different classes of disease, or different treatments.

transcripts (Tuikkala et al., 2006), or based on a Bayesian Principal Components Analysis (BPCA) (Oba et al., 2003; Tuikkala et al., 2008).

Principles of imputation methods

This study focuses on methods available in R or Bioconductor. In Table 2-1 an overview of these methods and their package availability is given.

According to Aittokallio (2010), imputation strategies can be separated into two major classes, (I) the generic statistical methods, and (II) application specific modifications, that use for example spot quality weights (Johansson and Häkkinen, 2006). The first group can further be divided into 6 subclasses (Aittokallio, 2010): (1) Mean imputation, e.g., column or row averages of the non-missing values, (2) hot deck imputation, that includes all methods which use similar non-missing cases, where the similarity is defined by using distance measures. KNN-imputation as a more application specific method derived from this principle; (3) model based imputation describes all methods that use a statistical model, typically linear regression, to predict the missing values of non-missing values of the same case, e.g., expectation maximization (EM), or least squares imputation (LSI); (4) multiple imputation methods estimate multiple values for one missing value; (5) cold deck imputation methods use external sources of information; and (6) composite methods, which use of combination of the aforementioned methods.

Table 2-1 Availability of imputation methods in R

Imputation method	Author	Year	in R	Package	Package citation
K-nearest Neighbors (KNN)	Troyanskaya et al.	2001	BioC	impute	Hastie et al., 2013
Sequential KNN (SeqKNN)	Kim et al.	2004	BioC	SeqKNN	Kim et al., 2008
Least squares imputation (LSI)	Bø et al.	2004	no[1]		Bø et al., 2004
Expectation maximization (EM)	Bø et al.	2004	no[1]		Bø et al., 2004
Local least squares (LLS)	Kim et al.	2006	BioC	pcaMethods	Stacklies et al., 2007
Bayesian Principal Component Analysis (BPCA)	Oba et al.	2003	BioC	pcaMethods	Stacklies et al., 2007
Linear model based imputation (LinImp)	Scheel et al.	2005	BioC	linimp	Scheel, 2007

[1] authors provide an Java-application

KNN. One of the first developed methods for substituting missing values is based K nearest neighbors (KNN) algorithm by Troyanskaya et al. (2001). For a given transcript with a missing value in one array the algorithm would find k transcripts with a similar gene expression, which have a value present in the other arrays. Then transcripts are weighted by their expression similarity to the transcript in question and a weighted average is calculated. This method is available in the Bioconductor-package *impute* (Hastie et al., 2013a).

SeqKNN. An improvement of the KNN imputation methods is the sequential KNN algorithm by Kim et al. (2004). It is designed to improve efficiency in data sets with high rates of missing values. Transcripts are sequentially imputed starting with the transcripts with the least number of arrays with missing values. Starting with the transcripts with only one missing value, k similar transcripts are selected from the set of complete transcripts. The weighted column average is calculated to substitute the missing value. This transcript is then considered as complete transcript and can be selected for subsequent imputation steps for transcripts with more missing values. This methods is available for R until version 2.14.0 in the Bioconductor-package *SeqKnn* (Kim et al., 2008).

LSI. Imputation methods based on the principle of least squares (LS), least squares imputation (LSI), as introduced by Bø et al. (2004) take the correlation structure of arrays and transcripts into account. There are different approaches to estimate missing values with LS.

The gene-based approach, LSI_gene, is based on the correlation of transcript intensities. For a transcript with missing values, the k most correlated transcripts are selected. Then in a single regression that predicts the values of the missing transcript with the complete transcripts. The so obtained k times two regression coefficients are weighed by correlation, the highest correlation of a transcript with missing values is assigned the highest weight, and a weighted average of the regression coefficients is calculated.

The array-based method, LSI_array, is based on the covariance structure between arrays and applies a multiple regression model to estimate missing values. A transcript has missing values in certain arrays then in a multiple regression step these missing values are estimated by using the profiles of the arrays with values in this transcript. To get a first estimate of the missing values and to be able to proceed with the multiple regressions, missing values are first substituted by the LSI_gene approach as explained.

Bø et al. (2004) also developed two combinations of gene and array approach. In both approaches 5% of the missing values are re-estimated. In the “combined” approach, LSI_combined, these estimates are used to determine a global mixing or weighing factor for the estimates of the array and gene approaches. It is called global because this mixing factor is applied to all genes or transcripts respectively. The mixing factor is determined by minimizing the sum of squared errors between known and estimated values. In the “adaptive” approach, LSI_adaptive the correlation structure of the data is considered. The mixing factor is calculated for a set of transcripts with a given maximum absolute correlation coefficient.

Expectation maximization. Bø and colleagues (2004) also implemented two methods based on expectation maxim ization (EM). They are comparable with the gene and array-based LS methods, but use instead of an empirical covariance matrix the maximum likelihood estimate of the covariance matrix. In an iterative algorithm the estimates of missing values and the covariance matrix is updated until the estimates stabilize. Methods are available as Java application from the authors via a supplementary website.

LLSI. In local least squares imputation (LLSI) by Kim et al. (2006) missing values of a transcript are estimated using a linear combination of k similar transcripts. Similar transcripts are selected on the base of Euclidean distance or Pearson’s correlation coefficient. Subsequently a linear regression model of the k most similar gene expression patterns is used to predict the gene expression of the transcript with missing values. LLS imputation is available for R language in the Bioconductor package *pcaMethods* (Stacklies et al., 2007).

BPCA. The Bayesian principal component analysis (BPCA) by Oba et al. (2003) is based on a principal component regression, a Bayesian estimation and expectation-maximization-like repetitive algorithm. In the first step a PCA, i.e., a covariance matrix of gene expression is calculated. The factors extracted from the PCA are then used within a principal component regression to explain the covariance matrix. Missing values within gene expression vectors are estimated by non-missing values of the vector using the principal component regression results, namely the factor scores. Finally Bayesian estimation is used to improve the accuracy of the obtained parameter set. A variational Bayes algorithm is used to execute Bayesian estimation for model parameters and missing values in a repetitive algorithm until the parameters converge. With this method redundant principal component axes are shrunk toward zero and so only relevant axes remain to be used. It is important to note for workflow considerations that the authors suggest using initially even transcripts with high rates of missing values for the BPCA imputation. It improves the estimation ability of the method as these transcripts yield additional information, but they also strongly suggest eliminating those transcripts afterwards from further analyses. BPCA imputation is available in the Bioconductor package *pcaMethods* (Stacklies et al., 2007).

LinImp. Linear model-based imputation (LinImp) by Scheel and co-workers (2005) estimates the gene expression on a certain array with a certain array of a certain variety and gene-based on a linear regression model. Therefore at first all missing values are imputed using for example KNN imputation, then parameters for the regression model are estimated and missing values are replaced by the outcome. This step is iterated until the parameters converge. The authors provided their imputation method as R-package *linImp* (Scheel, 2007).

Influence of imputation methods on biological downstream analyses

Although the different biological downstream analyses, i.e., clustering or differential expression analysis, are discussed later in this work, systematic reviews of the influence of different imputation methods on these analyses are discussed here. The influence of missing values on hierarchical microarray clustering methods are discussed by Brevern et al. (2004) and are later re-evaluated incorporating more imputation methods by the same workgroup (Celton et al., 2010). The influence of missing values on the detection of differentially expressed transcripts is discussed by Scheel et al. (2005) and Oh et al. (2011) who also discuss the influence of imputation methods on sample classification and transcript clustering.

Brevern et al. (2004) compared the very common imputation method KNN with the substitution of 0 on cluster stability of hierarchical clustering methods (see chapter 3.1 for further detail on clustering methods). Therefore a reference dataset was constructed in which all transcripts with missing values are removed. Then missing values were created randomly. The authors developed a Conserved Pair Proportions (CPP) index to assess the cluster stability, which corresponds to the number of pairs found in the reference cluster and after missing values imputation. The effects of missing values differs between different hierarchical clustering algorithms; single linkage method is the most stable method, followed by centroid and average linkage, and Ward’s linkage and complete linkage are the most sensitive to missing values. Substitution of missing values with zero, or with a KNN estimation method improves the cluster stability of all methods. The KNN imputation method thereby outperforms the imputation with zero.

Celton and co-workers (2010) re-evaluated the finding of Brevern et al. (2004) using more imputation methods and clustering methods to compare. For the general performance evaluation they used five different published data sets and discarded all transcripts with missing values, then missing values were simulated and missing rates ranging from 0.5% to 50% by step of 0.5% and then imputed. The Root Mean Squared Error (RMSE) was computed. For the evaluation of the influence on clustering, first the original data set was clustered as reference cluster. Then hierarchical clustering with Euclidean distance measure, several clustering algorithms, and k-means clustering for each imputed data set was performed. The resulting clusters were compared with CPP (Brevern et al., 2004) and a Clustering Agreement Ratio that describes the proportion of pairs of transcripts belonging to the same cluster in the reference clustering after imputation.

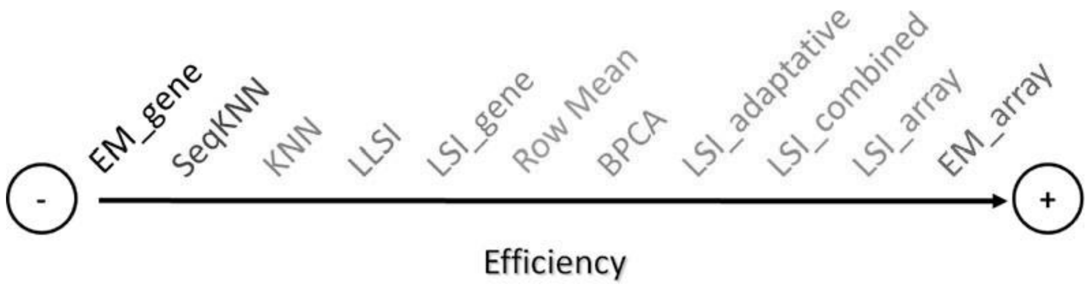

Figure 2-4 Efficiency of different imputation methods with regard to imputation performance and cluster stability (Celton et al., 2010).

Combined results of imputation performance and cluster stability are summarized in Figure 2-4. EM_array, LSI_array, LSI_combined, LSI_adaptive by Bø et al. (2004) perform best regarding the efficiency of imputation yielding the lowest RMSE and efficiency in cluster conservation having the highest CPP values. The widely used methods KNN and its improvement SeqKNN, each with optimized k value as defined by Troyanskaya et al. (2001), perform similarly. In terms of imputation efficiency and clustering they perform worst of the methods compared having the highest RMSE and lowest CPP values.

For the analysis of the impact of imputation methods on detecting differentiated expression Oh and colleagues (2011) examined SAM, LIMMA, and t-test with Benjamini-Hochberg correction (Benjamini and Hochberg, 1995). Therefore they used 8 different microarray data sets with binary clinical outcome. For the impact analysis on differential gene expression they defined a biomarker list concordance index (BLCI) that compares the lists of biomarkers obtained by the reference data set and the imputed data set. They compared the imputation methods: KNN based on correlation and based on Euclidean distance, LS_gene in the article referred to as ordinary LSI, LS_adaptive, BPCA, as well as the not further discussed SVD (Troyanskaya et al., 2001) and partial LS (Nguyen et al., 2004). They find that among the general evaluation of the impact of imputation on differential gene expression detection, sample classification, and transcript clustering, the detection of differential gene expression is the most affected by imputation methods. The best performances in missing values imputation and in detection consistency of differential gene expression show BPCA and LS_adaptive resulting in highest BLCI values.

Brock et al. (2008) introduce entropy for data sets and show that the estimation performance of imputation methods depends on the entropy of data sets. In context of microarrays, entropy describes a data set with aspect to complexity of a gene expression matrix. The more complex a data set is, the more difficult is it to map data to a lower-dimensional subspace, where only a few principal components would be generated in a principal component analysis. In more complex data sets, with then high entropy value, data cannot be reduced to only few components. When the entropy measure of a data set is high, neighbor-based methods, such as KNN or methods based on LS perform better. When complexity is low methods like BPCA, which use information of the whole microarray, perform better. Aittokallio (2010) reviews various studies on imputation methods and recommends choosing robust methods such as LSI for data with local substructures and BPCA for microarray data.

2.1.4 Low expression values

As seen in the previous chapters through background subtraction and log-transformation of low expressed data highly variable intensities for a transcript occur more frequently, also known as fanning (Ritchie et al., 2007). Through the use of more complex background correction methods this problem can be diminished, but not altogether erased. Additionally filtering probes, that have low amount of information due to low gene expression in most arrays, improves the power of the experiment.

To determine whether probe signals are low expressed a detection threshold must be defined. Probes in CodeLink Bioarrays, as used in this work, are determined by a detection threshold calculated by the analysis software or using quality flags of the manufacturer. The detection threshold, also referred to as negative control threshold, was calculated as a global threshold using the 80% trimmed mean of negative control probes as suggested by CodeLink Bioarrays (as referenced in Wu et al., 2005).

Global threshold calculation was replaced in 2007 by a local estimation method for lower detection limit (Applied Microarrays, 2007). The signal-to-noise ratio (SNR) is calculated for every spot, using spot mean intensity and local noise. Every SNR below 1, meaning every probe with higher background than signal value, is flagged as “L” (limit signal), values with SNR ≥ can be flagged as G (good). Local noise is calculated as:

Equation 1 Calculation of local noise for Codelink Bioarrays (Applied Microarrays, 2007)

Local noise = local background median + 1.5 standard deviations of local background

Not only low signals are flagged, but also S (saturated signal), I (irregular shape), M (missed set rate), C (background contaminated) or X for user excludes spots (for further detail see Diez, 2013). These flags are blanked for the analysis and are therefore treated as missing values (as discussed in chapter 2.1.3).

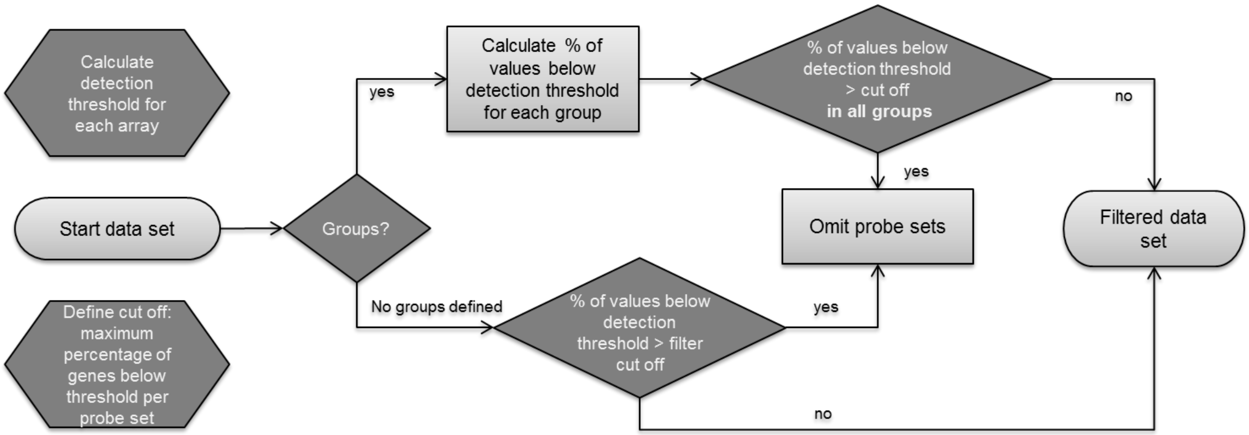
Calculate detection threshold for each array
Start data set
Define cut off: maximum percentage of genes below threshold per probe set
Groups?
yes
Calculate % of values below detection threshold for each group
% of values below detection threshold > cut off in all groups
no
yes
Omit probe sets
Filtered data set
No groups defined
% of values below detection threshold > filter cut off
yes
no

Figure 2-5 Low expression - Algorithm for filtering transcripts with low expression divided into analyses with and without prior consideration of groups of arrays.
To start filtering a cutoff for the maximum percentage of values below detection threshold per group or for all arrays must be given. Transcripts exceeding a certain amount of values below detection threshold in all groups are discarded and not further analyzed.

To improve data quality a filtering step is incorporated into the data quality workflow (see Figure 2-5). Aim of this filtering step is to eliminate all transcripts that are consistently lowly expressed. Therefore, a cutoff percentage of values below detection threshold is set. It can also be distinguished between the analysis of treatment groups in an supervised approach or the analysis of all arrays in an unsupervised approach. In a grouped analysis, only transcripts that exceed the maximum rate of low expressed values in all groups are discarded, to preserve transcripts that are low expressed due to biological reasons. In an ungrouped approach values of a transcript of all arrays are considered.

Spot with L-flag can also be considered in the quality assessment of microarrays. Instead of calculating the percentage of values below the negative-control threshold, the percentage of values flagged with L is calculated. The rest of the procedure remains the same.

2.1.5 Outlier in expression values

Through filtering low expressed probe signals a large part of the highly variable low expressed probes are eliminated from the analysis, but still probes with high variation might occur. Especially problematic are probe signals with only single values deviating from signals of one probe in all microarrays, the so called “outliers”, which can either be extremely large or small in comparison with the other expression values. Outliers can derive from errors in methods or be of biological nature (Pearson et al., 2003). These extreme values can alter the results of the microarray analysis. Especially mean and standard deviations (SD) are influenced by outliers and thus in statistical analyses based on these statistics, i.e., LIMMA, moderated t-test, Pearson regression, outliers may lead to problems. Outliers can also have a severe effect on imputation (Aittokallio, 2010).

For predicting purposes of the gene expression data, only transcripts with a stable gene expression are desired, although outliers or extreme expression values also bear the chance to examine individual effects or to identify subgroups in treatment groups and account for heterogeneity in samples (Ernst et al., 2008). Ernst et al. (2008) use extreme expression values for psychiatric research on an individual level to detect individual differences across a sample set. This shows that outliers are not altogether undesirable, but for the analysis of treatment groups they could lead to false positive or false negative results. Therefore these values are removed and are then handled as missing values, i.e., filtered and imputed together with the missing values from previous data preparation steps (see chapter 2.1.3).

Outlier detection methods

Outlier detection can be based on (1) z-score, (2) median, or (3) median absolute deviation (MAD).

Z-score. A common method defines outlier values as values outside a 2 SD distance from the arithmetic mean. Based on the z-score circa 5% of all values of a probe are detected as outliers with a 2 SD distance. The z-score for the interval [arithmetic mean_{ij} ± z SD_{ij}] for every probe i and treatment group is chosen according to the percentage of expected outliers in an interval. A major drawback of this criterion is that it is only feasible for normal distributed data with outliers included, but the z-score is not robust to outliers, and it is unlikely to detect outliers in small samples (Cousineau and Chartier, 2010; Leys et al., 2013). For this method is it very important to have symmetrically distributed expression values. Values of a probe in different microarrays tend to be skewed to the right, which is why data need to be log-transformed.

Median. Another criterion is based on the median of the probe signals over the samples of a group. The median is a more robust statistic as it is not influenced by the existence of outliers and the overall distribution of the values. After log-transformation outliers are defined as values outside the interval[11] [median_{ij} ± x] for probe i and treatment group j if defined. The value x is selected as a fixed constant.

MAD. Recently also a robust estimate for the SD, the median absolute deviation[12] (MAD) is used (Leys et al., 2013) combining both z-score and median based methods for outlier detection. Outliers are then defined as values outside the interval [median_{ij} ± x MAD] for every probe i and treatment group j.

[11] The interval for log-transformed data equals the interval [median_{ij}/x; x * median_{ij}] for untransformed data.

[12] $\text{MAD}_{ij} = b\ \text{median}_{ij}\ (\ |x_{ijk} - \text{median}_{ij}(x_{ijk})|\ )$, with b=1.4826 for patient k, group j, and transcript I (Rousseeuw and Croux, 1993).

Table 2-2 Example z-scores for considering outliers by a z-score based criterion with arithmetic mean and standard deviation.
Bonferroni correction takes the sample size into account. Every value outside the interval [arithmetic mean$_{ij}$ ± z SD$_{ij}$] for every transcripts i and group j is considered as outlier. The decision criterion α gives the percentage of values expected to be outliers in normal distributed values (Cousineau and Chartier, 2010).

Decision criterion α	no correction	Sample size with Bonferroni correction						
		3	5	10	20	30	50	100
0.1	1.64	2.13	2.33	2.58	2.81	2.94	3.09	3.29
0.05	1.96	2.39	2.58	2.81	3.02	3.14	3.29	3.48
0.01	2.58	2.94	3.09	3.29	3.48	3.59	3.72	3.89
0.001	3.29	3.59	3.72	3.89	4.06	4.15	4.26	4.42

For the preprocessing workflow used in this work z-score based and median based outlier detection methods are implemented. The constants z and x determine the length of the interval. The z-score is derived from the standard normal distribution and the percentage of values expected to lie within the interval (see Table 2-2). A more conservative approach takes the sample size n into account, so that the decision criterion α is corrected for the sample size: $\alpha_c = \alpha/(2n)$ (Cousineau and Chartier, 2010). For choosing x for the MAD based method Table gives a rough estimate of the conservativeness of the intervals.

Outlier detection is repeated at least once to account for shifts in mean or median, which results from the deletion of single values. If no further outliers are detected probes are assessed for the amount of missing values and probes are eliminated if they exceed a certain percentage of missing values as described in chapter 2.1.3. Outlier detection is the last step of data preparation, where missing values are added to the data set, so missing values are imputed after filtering steps are completed.

Assessment of outlier detection methods

Visual inspection is an important tool to assess whether outliers are truly detected (Cousineau and Chartier, 2010). For this purpose we use MA-plots (see Figure 2-6). Background corrected, intra-slide normalized, filtered for missing or low expression values are used. Log-transformed data is mandatory for a more symmetric graphic. The group median of a probe is plotted against the difference of this group median and the expression of this probe in the respective array.

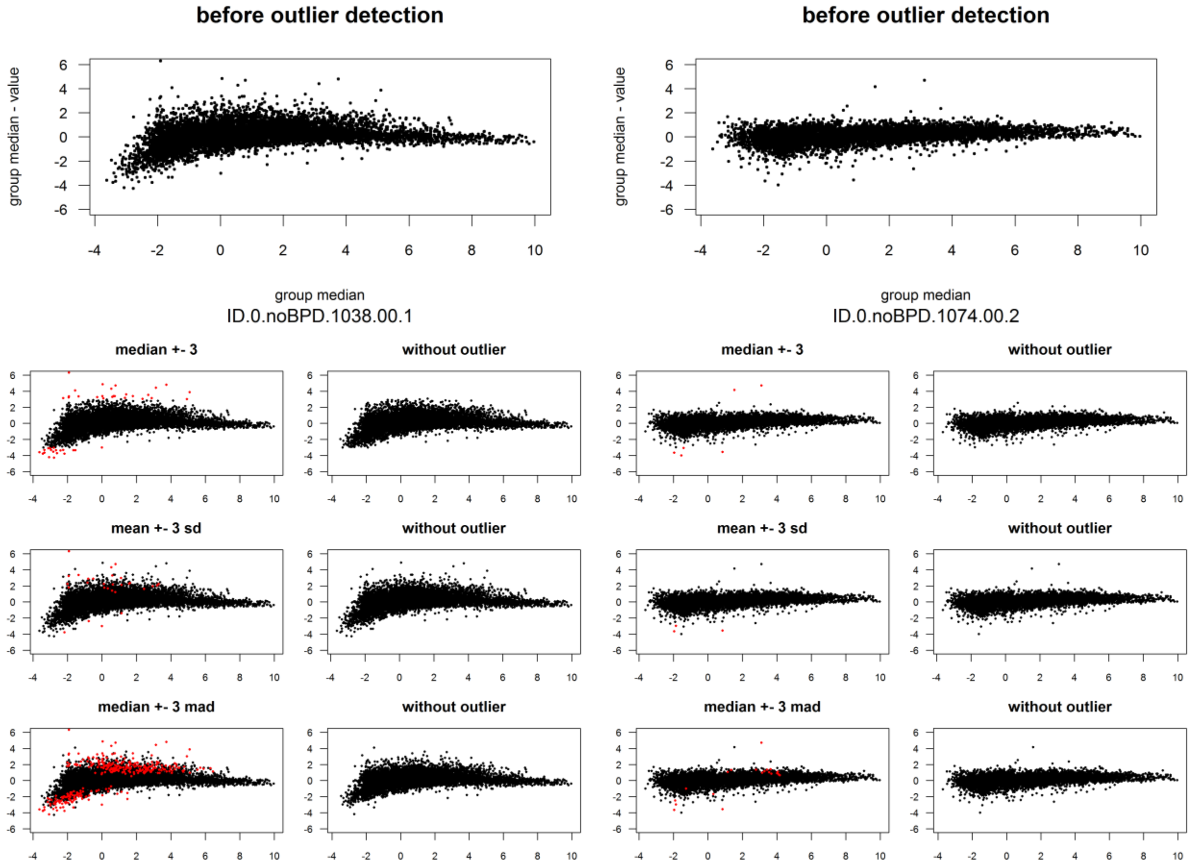

Figure 2-6 Effect of different outlier detection methods on two example arrays. Depicted in red are the detected outliers with different methods based on median, z-score, and median absolute deviation (mad) with constant = 3. The panel on top shows the variation of the values around the group median. The right side represents an array with little variation; the left side shows an example with high variation, especially in low expressed transcripts. In the lower panels on the respective left hand side, detected outliers are depicted; on the respective right hand side is the resulting distribution without outliers shown. The depicted MA-plots are constructed on log-transformed data after filtering for missing or low expression values.

In the upper panel of Figure 2-6, MA-plots of data before outlier detection can be seen. Two examples are given, on the left panel an array with higher variability is shown, due to high deviations from the median especially in low and middle expressed transcripts, a higher amount of outliers can be expected than in the right hand side panel, were the variability around the median is lower. In the lower panels the effect of outlier detection, without the effect of imputation, can be evaluated. Outlier detection based solely on the median results in a clear cut of outliers. Detection based on the z-score takes also SD of transcript expression into account and therefore a clear line cannot be observed. Compared to detection via median another set of values is detected as outliers. The third introduced method of outlier detection using median and MAD identifies a higher number of outliers. The MAD is a more robust estimator for the standard deviation, which is smaller than SD used for the previous two methods and therefore more values lie outside the interval [median_{ij} ± 3 MAD] than the interval [mean_{ij} ± 3 SD]. The MAD in most cases smaller than 1 and therefore the interval [median_{ij} ± 3 MAD] is smaller than the interval [median_{ij} ± 3]. This allows us to increase the constant in this approach to reduce the probability for false positive detection of outliers.

The visual inspection of outliers allows to find arrays that have a high amount of outliers and therefore to identify interesting subjects deviating from the group or arrays that may be corrupted and need to be excluded from the analysis.

2.1.6 Detection of outlier samples and array quality assessment

Outlier samples in contrast to outlier values are arrays whiceviate greatly from the expression profile of the other samples. That can either be the result of hybridization errors or RNA problems or sample mislabeling or sample mistreatment (Kauffmann and Huber, 2010).

The step of controlling outlier samples or outlier arrays, including technical replicates of the samples, in microarray analysis serves to eliminate all greatly compromised arrays (Reimers, 2010). This analysis step should be done as first step of the analysis using raw data, but often outlier arrays are not apparent at the beginning of the analysis. In every step of the analysis, statistics of the arrays should be monitored carefully. In this chapter different approaches to detect outlying arrays are introduced and described.

Kauffmann and Huber (2010) examine different methods for the detection of outlier arrays and show that the use of outlier removal procedures leads to an improvement in the analysis. It can be expected that through the elimination of arrays the power of the analysis decreases, but it actually strengthens the power to discover differentially regulated genes and improves significance levels of biological relevant pathways detected by gene set enrichment analysis. The use of outlier removal procedures leads to an improvement in the biological relevance of the analysis. This description focuses on relative outlier detection; outlier arrays are discovered in comparison to other arrays. Absolute quality metrics use internal controls, or spike-in, or the variability of replicate probe sets on the array (Kauffmann and Huber, 2010). A careful examination of the quality metrics provided by the manufacturer gives an impression of the absolute quality of this array.

Detection of outlier arrays in comparison to other microarrays can be based on robust principal component analysis (Shieh and Hung, 2009) or the correlation coefficient between arrays and number of outlier values per array (Yang et al., 2007). In general the distance between arrays, and the distance between technical replicates with subsequent cluster analysis can be used to detect compromised or mislabeled arrays. For the examples, introduced in this chapter, of outlier detection we used hierarchical clustering with Euclidean distance and complete linkage clustering (see Figure 2-8) or Ward’s linkage clustering (see Figure 2-9).

Visualization tools facilitate the rapid identification of possible outlier arrays. For an overview of all distances a cluster analysis and a dendrogram is helpful. A dendrogram shows not only the distance between single arrays, but also the distances between groups of arrays (see Figure 2-8) as further discussed in chapter 2.1.6.1. MA-plots were again used to identify outliers. For the purpose of identifying whole outlier arrays the relative amount of identified outliers in a technical replicate compared to other technical replicates can be assessed (see Figure 2-7). Lastly, heatmaps are suitable to highlight pairwise distances or gene expression profiles as demonstrated in Figure 2-9. It gives a visual impression of the similarity or distance between the examined expression profiles.

The following examples demonstrate how outlier detection in the presented data set was used to detect outlier in technical replicates by examining outlier values and with the use of cluster analysis at the end of the quality control workflow.

2.1.6.1 Detecting outlier in technical replicates

For our investigation 63 CodeLink Human 10k Bioarrays of 22 preterm infants were examined in an unsupervised analysis approach with no prior definition of groups as described in chapter 4. For each sample two to four technical replicates were prepared. Raw data was background corrected using subtract and intra-slide normalized using median normalization. Negative values were removed, and transcripts were filtered, when more than 50% of the values were missing or below negative control detection threshold. Outlier values were defined as values of a probe i outside the interval [median$_i$ ± 3]. To visually assess whether outlier arrays were present MA-plot were used (see Figure 2-7).

Technical replicates with high number of outlier values in comparison to the other replicates can be considered corrupted and should be removed from the analysis. In this data set the arrays “unsup.896.00.3” and “unsup.1073.00.3” show a great number more outlier values than the respective replicates.

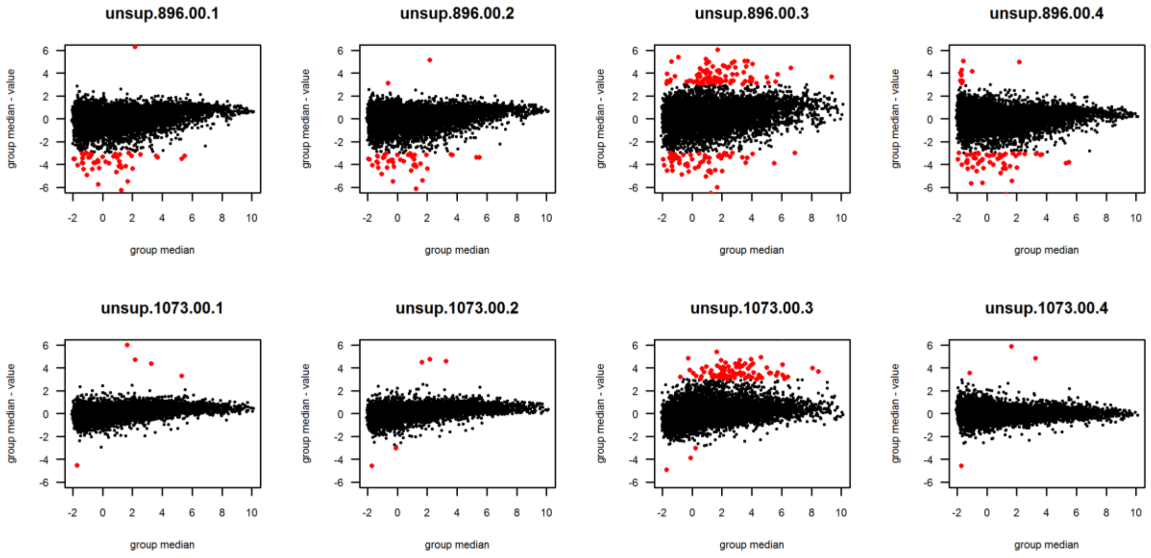

Figure 2-7 Detection of outliers in technical replicates via MA-plots.
Each row shows a set of technical replicates of two samples with outlier arrays. In each row the third plot shows a higher number of outlier values (indicated in red) and can be considered outlier arrays. Median of a transcript of all arrays was plotted against the difference of an expression value of a transcript and array and the respective median. As outlier value were values with an absolute difference to the median greater 3 considered.

For demonstration purposes these arrays were left in the analysis and after filtering probes with high rates of missing values due to outlier values, all missing values were imputed, and technical replicates were averaged. Figure 2-8 demonstrates how outlier in arrays can be detected through the distance of technical replicates to each corresponding replicate and through the distance of a single array to all other arrays.

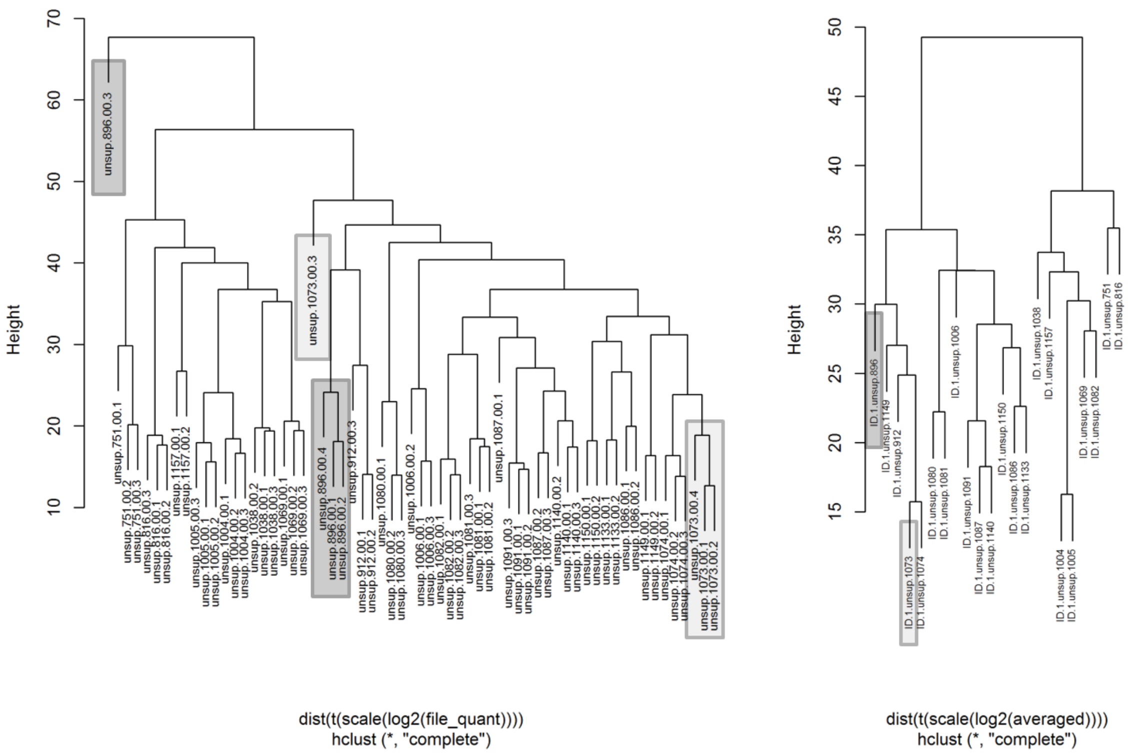

Figure 2-8 Detection of outliers in technical replicates through clustering.
Dendrograms of hierarchical clustering of arrays before, on the left side, and after averaging, on the right side, of technical replicates indicate that two arrays can be classified as outlier as they show a high distance to the other replicates. Contrary to the expectation of a high similarity of technical replicates and thus an early clustering of these arrays, the outlier arrays are added late to the establishing clusters. The two sets of replicates are indicated in lighter and darker grey boxes.

Two dendrograms are shown; the left shows the hierarchical clustering with Euclidean distance and complete linkage clustering of arrays before technical replicates are averaged, while the right shows the hierarchical clustering result after averaging. All genes that remained after quality control are taken into account. Between technical replicates a short Euclidean distance is expected. In complete linkage clustering this translates into an early clustering of technical replicates. In the example however, it can be seen that array “unsup.896.00.3”, indicated in darker grey in the upper left corner of the left dendrogram, not only has a greater distance to its replicates, but also to the rest of the arrays, strongly suggesting that it is an outlier array and should be removed from the analysis. Replicate “unsup.1073.00.3”, indicated in lighter grey, also shows a great distance to its replicates, also marked in lighter grey. The same phenomenon can be seen for the supervised microarray analysis approach (data not shown), consideration of groups has no apparent effect on the distance between outlier-arrays and their replicates.

2.1.6.2 Detecting mislabeled arrays through microarray analysis

Microarray analysis may also be able to detect grouping errors or may be able to generate hypothesis concerning single outlying microarrays. In the later described analysis of cord blood of preterm infants later developing BPD it became apparent, that arrays of preterm infant displayed a gene expression similar to the group it was not assigned to.

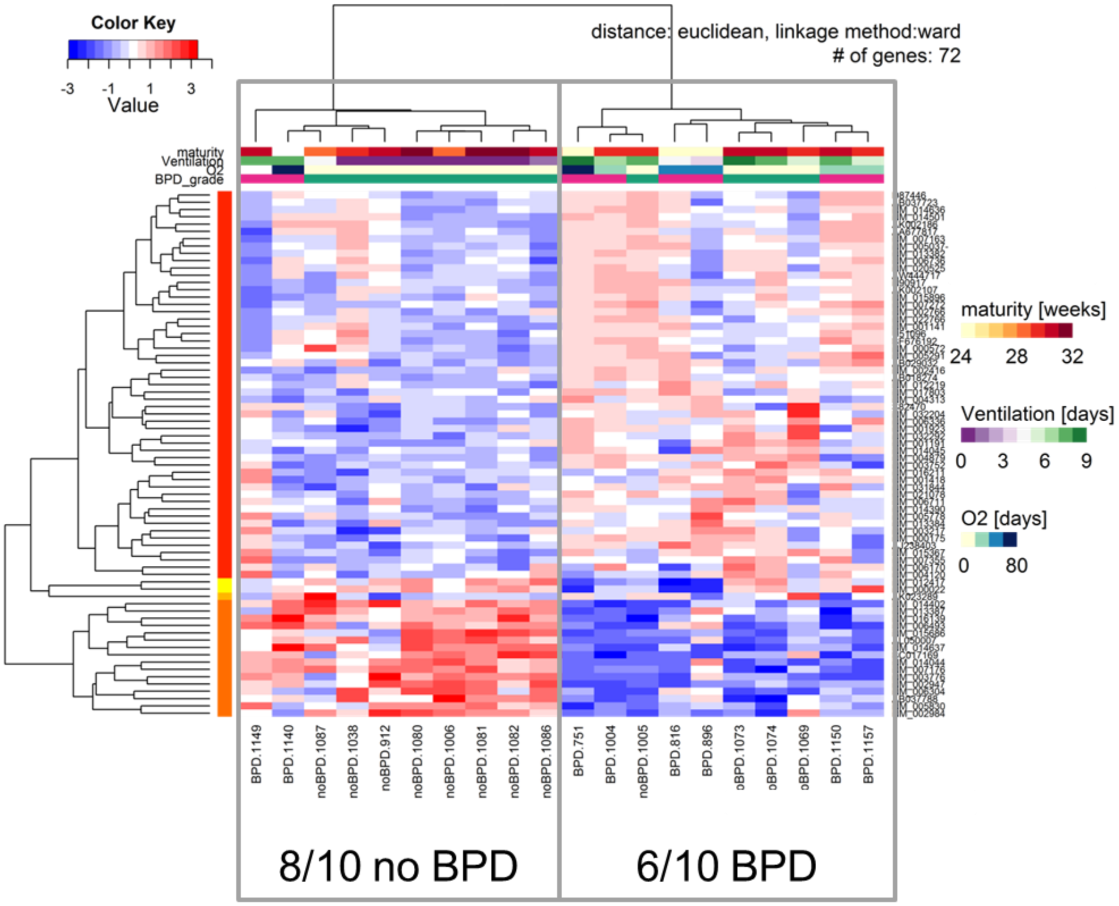

Figure 2-9 Cluster analysis of preterm infants with and without BPD can also be used to detect mislabelled arrays.
In the cluster (on the left hand side) of eight preterm infants without BPD two arrays of BPD preterm infants are clustered; of these two clusters one array (BPD.1149, on the far left) belongs in fact to the group of no BPD preterm infants; it displays a profile similar to the no BPD preterm infants than to the BPD infants.

To detect this, an unsupervised quality control, i.e., without the prior assignment of groups, was conducted. Afterwards transcripts were selected using a linear model of known risk factors to describe the variation in gene expression (see detailed description in material and methods of the study itself in chapter 4.2). If at least one regression-coefficient of the model displayed an adjusted significance level of ≥ 90%, the transcript was selected and all transcripts were used for a hierarchical cluster analysis.

Then a closer look was taken at the cluster formation in the gene expression patterns that fit to the pattern of risk factors for BPD that were established later, i.e., days of mechanical ventilation and oxygen dependence of the preterm infants, as well as a priori risk factors, i.e., the gestational age or maturity of the preterm infants. In Figure 2-9, a heatmap of the expression profiles of 20 preterm infants at time of birth is displayed. Hierarchical cluster analysis (Euclidean distance, Ward's linkage clustering) was performed on samples and on transcripts. In transcript clustering roughly two main clusters can be distinguished, a cluster of up-regulated and a cluster of down-regulated transcripts between the two clusters identified on sample level. One of the sample cluster mainly, i.e., 8 of 10, contains preterm infants without BPD (see Figure 2-9, top left cluster) and one cluster mainly contains BPD preterm infants (see Figure 2-9, top left cluster). Examining cluster 1 with only two no BPD preterm infants it becomes apparent, that these two expression profiles rather seem like profiles of preterm infants without BPD than with BPD. In the process of examining this phenomenon it becomes apparent that one of these preterm infants ("BPD.1149") had a short need of oxygen support and therefore should not be grouped in the BPD-group. After double-checking with the clinical data of these preterm infants it becomes clear, that this infant in fact did not develop BPD.

2.2 Preprocessing workflow

The preprocessing workflow (Figure 2-10) described in this chapter will summarize the methodology and steps used in the BDP study to be presented in chapter 4, which are necessary to make data obtained from different microarrays comparable through noise reduction by background correction and normalization, as well as examination and preparation of signals for statistical analysis. The presented workflow thereby allows to easily evaluate the order of those steps and to allocate their respective place within the workflow.

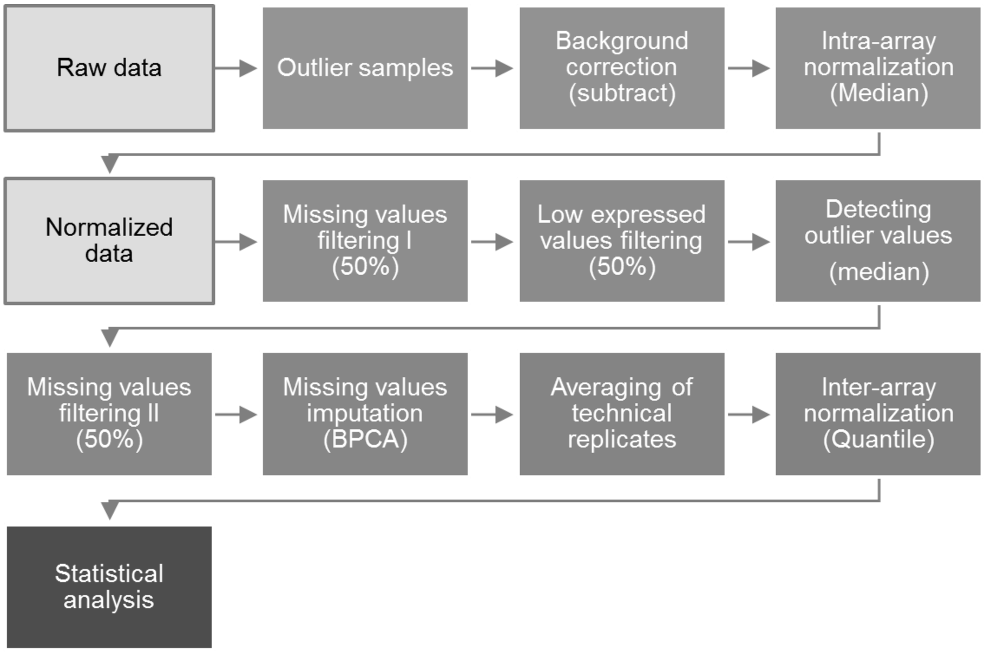
Raw data
Outlier samples
Background correction (subtract)
Intra-array normalization (Median)
Normalized data
Missing values filtering I (50%)
Low expressed values filtering (50%)
Detecting outlier values (median)
Missing values filtering II (50%)
Missing values imputation (BPCA)
Averaging of technical replicates
Inter-array normalization (Quantile)
Statistical analysis

Figure 2-10 Preprocessing workflow as established for this investigation.
It begins with the raw data format. Microarrays are constantly checked for abnormalities affecting whole arrays. Outlier samples are then removed from the data set. Background is subtracted and intra-slide normalized using median normalization provided by the manufacturer software. With the normalized data begins data preparation in R. Values are filtered if more than 50% in at least one group are missing, or more than 50% in each group are low expressed. Then outlier values are identified by the group median and set to missing. Again values were filtered for missing values. Missing values are then imputed by Bayesian Principal Component Analysis (BPCA). Technical replicates are averaged using the arithmetic mean of the transcripts

In short, the workflow aims at the (1) removal of uninterpretable signals due to high rates of missing values, or rates of values below a detection threshold, (2) reduction of variances of gene expression across microarrays of a treatment group due to high rates of outlier values in a treatment group, (3) imputation of missing values to meet requirements and improve results of statistical analyses. The following workflow is incorporated in R (R Core Team, 2014) or prepared to be incorporated into a R-routine.

Outlier samples. Abnormalities in the data with regard to overall distributions of microarrays, e.g., in the number of missing values or the number of outlier values is constantly monitored. In addition, clustering techniques (see chapter 3.1) are applied and correlation measures between technical replicates used to detect microarrays which deviate strongly from technical replicates or from the treatment group. Detected outliers are then examined separately for possible defects or handling errors, e.g., mislabeling and are finally discarded.

Background correction As stated in section 2.1.1 it is necessary in microarray experiments to improve the signal of gene expression and to reduce the noise in the measured intensities due to technical issues, e.g., different amount of hybridized mRNA. To ensure the ability to compare the recent study with other previously done studies based on this array type we followed the recommendation of the manufacturer of CodeLink Bioarrays and used background subtraction. But it later has been shown that background subtractions bears some difficulties and a variance stabilizing method for background correction, i.e., “normexp” in addition to an offset, lead to a higher accuracy in returned signals (Ritchie et al., 2007). In future studies it should be considered to switch to another background correction method. Resulting negative values were blanked.

Normalization. For the investigation presented in this work, we normalized the data twice. After the preparation of the microarray data was first background corrected using the subtract method and then normalized using Median normalization as an *intra*-microarray normalization step. As previously stated in section 2.1.2, Median normalization has a great advantage in microarray analysis, because it is applicable when microarrays are processed at different times and it is, as well as background correction, independent of the expression of other microarrays and still achieves comparable distributions in signal intensities.

After Median normalization data is filtered as discussed below affecting once more the distribution. Therefore we added a second normalization step. Here, Quantile normalization as an *inter*-microarray method is used, which then compares the distributions of the microarrays actually used in the study. For Quantile normalization the Bioconductor *limma* (Ritchie et al., 2015) was used.

Missing values filtering. Filtering for missing values was done twice. First, data was filtered for missing values accounting for missing values due to preparation steps and background correction. If at least 50% of the values of a transcript in at least one group were missing the transcript was excluded. A second step of missing value filtering is conducted after outliers are identified and removed from the data set, again rejecting transcripts with missing rates > 50%.

Low expressed values. In contrast to filtering for missing values low expressed value (missing values are counted as low expressed values) filtering is only done, if in each group a certain amount of low expressed values are found. Filtering for low expressed transcripts was conducted using the negative control threshold as quality flag data was not available for our data set. In the case that flags are available, they should be considered as they bear even more information than the negative control threshold. Transcripts were excluded if in each groups at least 50% of the values were below detection threshold. The order of the missing value or low expressed value filtering does not play an important role. We decided to place this step second because of computation time constraints. The removal of transcripts with high amounts of missing values in a treatment group may speed up the process of low expression filtering.

Outlier values. Outlier detection is conducted in order to avoid bias due to sporadic extreme values in a treatment group and transcript. Data was analyzed for outlier values defining log-transformed values as outlier using a median based method. We used a method based on median plus offset for outlier detection as it is a more robust method to detect outliers than using a z-transformation based method. In the future using and implementing a MAD based method would be more conservative as it considers a robust estimation for the standard deviation of a transcript over all arrays. The procedure was performed twice to adjust for new medians after the first step of outlier detection. Afterwards data was again filtered for missing values using the same settings as described above.

Imputation. The remaining missing values, values were log-transformed, were imputed using an BPCA imputation described by Oba et al. (2003) via the Bioconductor-package *pcaMethods* (Stacklies et al., 2007). BPCA is recommended as robust method suitable for microarray data by various reviews and imputation comparison studies as elucidated in chapter 2.1.3.2.

Log-transformation. After inter-microarray normalization, technical replicates are averaged using the arithmetic mean of a transcript. One procedural question remains: when to log-transform the data. In the current workflow data is log-transformed as the last step before statistical analysis, but for other case studies it may have advantages to log-transform at an earlier point in data preparation. The background correction methods subtract and half are not affected by log-transformation. For normexp background correction Ritchie et al. (2007) recommend to perform a started log-transformation afterwards. Median and Quantile normalization are independent of the prior distribution of the transcripts. Procedures filtering transcripts with missing values or values below detection threshold are not affected whether data is log-transformed or not, but imputation procedures are affected. Evaluation of literature sources (Aittokallio, 2010; Oba et al., 2003; Oh et al., 2011; Troyanskaya et al., 2001) indicate a preference for log-transformation of data before the imputation step. Troyanskaya and colleagues (2001) also point out that log-transformation reduces the effect of outliers present in the data set on transcript similarity detection. For averaging technical replicates it is also beneficial to use log-transformed data, as this ensures normal distributed data where the mean is the most accurate statistic.

3 Statistical analysis of microarray experiments

Statistical analysis in microarray data serves two main purposes: (1) find similar expression profiles caused by a certain disease at a certain time point and identify the underlying biological processes, and (2) find a gene expression profile that is able to predict a certain disease or outcome.

Based on the prior information about the subjects, statistical methods can be classified as supervised, when information about the subjects is considered, and unsupervised, when no information about the subjects is given (Bair and Tibshirani, 2004; Boutros and Okey, 2005). Unsupervised methods perform well in recognizing underlying patterns. The most common unsupervised methods are hierarchical clustering and partitioning clustering (i.e. k-means clustering and self-organizing maps (SOM)) (Boutros and Okey, 2005).

To gain more information about a certain disease, treatment groups are defined prior to the microarray analysis in most studies. In these supervised approaches, differentially expressed genes or transcripts between those subtypes are determined; this so called gene prioritization generates a small set of relevant transcripts to analyze further. Two of the main problems in the analysis of microarray experiments are (1) the larger number of predictors, here gene transcripts, which exceeds the number of observations, here patients or samples and (2) multicollinearity of the microarray experiment itself as the expression of different transcripts are highly correlated (Pérez-Enciso and Tenenhaus, 2003). Therefore adaptions to common statistical tests, e.g., Student's t-test for two-sample problems have to be made.

A third class of methods is defined by Bair and Tibshirani (2004): the semi-supervised methods, that combines gene expression data and clinical data to predict disease subtypes. Analyses for prediction purposes try to prioritize the gene sets even more stringent. Only the transcripts which contribute most to the classification of a certain disease are selected.

To gain more information about the underlying biological processes, the set of transcripts is annotated and analyzed for common functional themes, e.g., biological functions or regulators. Therefore overrepresentation analyses were developed.

In the following chapter different methods for classification of disease types, differential gene expression between disease types, prediction of disease types, and functional annotation are introduced.

3.1 Clustering approaches

Clustering or unsupervised pattern recognition tries to identify small subset of transcripts or samples that have a similar expression pattern (Boutros and Okey, 2005; Chipman et al., 2003; Modlich and Munnes, 2007). In subjects or patients different types of diseases e.g. cancer subtypes (Bair and Tibshirani, 2004; Yeoh et al., 2002) shall be identified with clustering approaches. In gene expression subsets of genes with similar expressions, the so called co-expression, shall be identified to discover possible functional relationships as specific functions tend to be enriched in gene clusters (Boutros and Okey, 2005; D'haeseleer, 2005). These co-expression patterns then can, e.g., be used to identify common regulators or common transcription factor binding sites (Boutros and Okey, 2005).

Two of the most important classes of clustering methods are the unsupervised methods hierarchical clustering and partitioning (D'haeseleer, 2005). Hierarchical clustering subdivides clusters into smaller clusters, a hierarchical structure of clusters is established; similar genes or patients are successively grouped together. In partitioning clustering, as for example k-means or SOM clustering, a number of clusters is predetermined (Bair and Tibshirani, 2004; D'haeseleer, 2005).

Before clustering techniques can be applied a standardization step should be considered, so every transcripts has the same weight in clustering or classification (Dudoit and Fridlyand, 2003). Standardization, also referred as scaling, is a z-transformation that achieves that every variable has the mean zero and standard deviation of one, by subtracting the mean of the variable from every value and dividing by SD of the variable; also possible are more robust estimators: median and MAD (Dudoit and Fridlyand, 2003).

All methods have in common that the similarity between patients or transcripts needs to be determined. Most commonly used are Euclidean distance, which is sensitive to scaling and differences in average expression levels and the Pearson correlation coefficient (D'haeseleer, 2005).

Hierarchical clustering. In hierarchical clustering first the two objects with the shortest distance are clustered together, and then depending on the linkage function a new distance matrix is calculated and the next object is added to the cluster. If two unclustered object have the shortest distance, then first those two objects are clustered (Quackenbush, 2001). The added object can be:

- the object with the shortest distance to one of the clustered object (single linkage),
- the object with the shortest distance to the arithmetic average or median distance of both objects or the centroid of two clusters (average or median or centroid[13] linkage respectively),
- the shortest distance to the furthest object in the cluster (complete linkage) (Boutros and Okey, 2005; Brevern et al., 2004; D'haeseleer, 2005; Quackenbush, 2001).
- or the object that minimizes the sum of squared deviations from the mean of a cluster of two clusters (Ward's clustering) (Quackenbush, 2001; Ward, 1963)

This way a tree-like structure, a dendrogram, is built (see Figure 2-8). The number of clusters and their members are then determined by using a maximal distance between to clusters as cut-off (Boutros and Okey, 2005; D'haeseleer, 2005).

K-means clustering. Partitioning methods use an iterative approach to minimize the within-group dissimilarity for a predefined number (k) clusters (Chipman et al., 2003). In k-means clustering objects are randomly assigned to one of the previously defined k clusters, an average expression is calculated for each of the clusters; the inter- and intra-cluster distances are calculated. In an iterative step objects are shuffled until moving an object would lead to higher intra-cluster variability and lower inter-cluster variability (Quackenbush, 2001).

Self-organizing-maps. In self-organizing-maps (SOM) a set of k reference vectors is assigned for each partition. Objects are then clustered in accordance to their similarity to these vectors. The reference vectors are first derived from a geometric grid, predefined by the user, laid over all objects to analyze. A random expression is then used as reference vector, this expression vector is iteratively adjusted when new objects are assigned to the cluster (Quackenbush, 2001; Tamayo et al., 1999).

Principal component analysis. A problem of partitioning methods is to identify the number of clusters in the data set. Here, a PCA can help. PCA reduces the dimensionality of a data set by identifying principal components in highly correlated data that are no longer correlated (Chipman et al., 2003). The resulting principal components should still explain the same amount of variance in the data that the single variables in the analysis

[13] The centroid of a cluster can be understood as center of gravity of a cluster in a three-dimensional space (Quackenbush, 2001)

(Raychaudhuri et al., 2000). PCA can be used to inspect variables and their distance to each other in a three-dimensional space or to estimate the number of clusters to be analyzed using k-means clustering or SOM. Eigenvalues of principal components indicate how much of the total variance can be explained by the component. The number of k clusters can be determined by inspecting the Eigenvalues of the principal components: the higher the Eigenvalue the more variance is explained. The number of components with high Eigenvalues estimates the number of clusters (Raychaudhuri et al., 2000).

Common visualization techniques for hierarchical clustering in microarray data analysis are heatmaps accompanied by the respective transcript and sample dendrograms (see for example Figure 2-8). Heatmaps provide a quick overview of similarities and dissimilarities of samples and transcripts; hereby, the most effective way is ordering the transcripts and samples. Dendrograms or other clusters methods provide the possibility to order the heatmap in an effective way (Chipman et al., 2003). For co-expression analyses often profile-plots are used.

Freyhult et al. (2010) examined seven cancer data sets and different clustering methods for their performance in detecting known classes. They examined hierarchical clustering with Euclidean distance, Manhattan distance, and Pearson correlation coefficient distance together with average linkage and Ward’s linkage, as well as k-means, SOM, prediction analysis of microarrays (PAM, see 3.3.1), and model-based clustering (Mclust) with the help of the Rand index, which measures the similarity of two portioning methods. They find that in hierarchical clustering Ward’s linkage outperforms average linkage clustering with every distance measure investigated. They report, that hierarchical clustering with Ward’s linkage and correlation distance as well as k-means performs better than PAM and SOM. Also Brevern et al. (2004) show that hierarchical clustering and especially complete linkage and Ward’s linkage methods are the most suitable clustering methods for microarray experiment and therefore these methods will be pursued for the presented BPD study.

3.2 Prioritizing genes or differential gene expression analysis

In comparative microarray experiments gene expression is measured for defined groups, the so called treatment groups, of samples. These treatment groups may be for example a group of healthy subjects versus a group of patients, a control versus disease set up, or different grades of a single disease. Also in a microarray experiment samples taken at different time points and different stadiums of a disease can be compared. Is a transcript or gene over- or underexpressed, in one or more groups this transcript is differentially expressed (Li and Tibshirani, 2013). Differential expression analysis also can be used to identify transcripts that are correlated with a quantitative feature or the survival of patients.

In the following chapter some common methods to identify differential expression in two- or multiple-group set ups are introduced. But first the problem of testing multiple hypotheses at once needs to be addressed: Multiple testing problems occur when multiple parameters are tested in a single population of probes. In microarrays thousands of genes or probe sets are monitored simultaneously. According to the multiple testing problem the probability of a false discovery or a false positive detection of statistically significant differentially expressed genes increases dramatically (Reiner et al., 2003). Therefore, the *“use of ordinary t-tests or other traditional univariate statistics to assess differential expression be disastrous”* [(Ritchie et al., 2007) citing (Smyth, 2004)].

Jeanmougin and colleagues (2010) used gene list analysis, simulations, spike-in data sets, and re-sampling to compare the Welch’s t-test, analysis of variance (ANOVA), Wilcoxon rank sum test as classical methods to analyze differences in groups, and SAM and LIMMA among others for their efficiency to detect differentially regulated genes. They show that LIMMA and SAM tends to increase power in the spike-in data set; ANOVA, SAM, and LIMMA have no deviation in the actual percentage of false-positives compared to the expected percentage of false positives in the simulation study, while Wilcoxon and Welch’s t-test tend to be more conservative, especially in small data sets (Jeanmougin et al., 2010).

This shows that the use of methods designed to address the problems of microarray data analysis are better suited than using conventional statistical analyses as, e.g., the t-test.

3.2.1 Fold change

Differential gene expression analysis started out by simply comparing expression levels a ratio of the mean expression of a group in a transcript (see DeRisi, 1997), the so called fold change (FC).

The FC is calculated as shown in

Equation 2 for an example microarray design with control probes and disease probes. But the calculation holds true for every pairwise comparison of groups of microarrays.

Equation 2 Calculation of the fold change (FC) for a disease/control design

$$\text{FC}=2^{\text{expression}\frac{\text{disease}}{\text{control}}}\text{, for ratios > 1(up regulated transcripts)}$$

$$\text{FC}=-\frac{1}{2^{\text{expression}\frac{\text{disease}}{\text{control}}}}\text{, for ratios < 1(down – regulated transcripts).}$$

As gene expression is log-normal distributed, the expression ratio is calculated with log-transformed expression values. The expression ratio is then calculated using Equation 3.

Equation 3 Calculation of the log-ratio for a disease/control design in single-channel arrays

$$\text{log-ratio= ld}\left(\text{expression}_{\text{disease}}\right)\text{-ld}\left(\text{expression}_{\text{control}}\right)$$

The fold-change as single criteria does not consider the variance of gene expression levels and is thus prone to identify a differential gene expression that are not truly differential expressed, but have a higher measurement error as e.g. in low expressed transcripts. They then meet the minimum FC request, but do not represent a truly differentially expressed transcript. On the other hand it is also possible that transcripts with a low FC are very stable in a certain condition but are truly differentially expressed; they are not detected due to the set cut-off FC (Mutch et al., 2002).

Today the fold-change is often used as an additional criterion to define a minimum change in expression to be relevant in biological networks.

3.2.2 Significance analysis of microarrays (SAM)

Comparing normal distributed gene expressions with homogeneous variances would lead to a series of conventional Student's t-tests in an experiment with a small number of genes to test. This would not only consider the FC of the transcript in question, but also the variances in the two different groups. Thereby stable transcripts are selected to be significantly differentially regulated from a statistical point of view. But a large number of tests would lead to a large number of transcripts identified as differentially regulated by chance alone.

Tusher and colleagues (2001) adapted the Student's t-test specifically for microarray experiments, which would require a large number of t-tests. The test-statistic is here calculated as shown in Equation 4, where s is the pooled standard error of the difference in expression of both groups resulting in an adapted t-statistic for d(i).

Equation 4 Test-statistic in SAM (Tusher et al., 2001)

$$d(i)=\frac{\overline{x_{i\ \text{disease}}}-\overline{x_{i\ \text{control}}}}{s+s_0}$$

The authors added a small positive constant s_0 to the denominator to make the variance of d(i) independent of the gene expression level. Then all balanced permutations, all permutations with an equal amount of control and disease probes, were computed. From these permutations the expected average expression is calculated and compared to the calculated d(i). A cut-off is set as minimum difference between observed and expected expression levels (Tusher et al., 2001).

The number of false positives is determined by using the permutations, where no regulation is expected; the average number of transcripts identified as significantly differentially regulated in permutations indicates the number of false positives. Compared with the number of significantly differentially regulated genes the percentage of false positives is computed (Tusher et al., 2001). The percentage of false positive transcripts is also called false discovery rate (FDR), first introduced by Benjamini and Hochberg (1995) later adapted by Storey (2002), whose version is used for SAM.

This method can also be extended to multivariate or paired analysis by re-defining parameters used to calculate d(i) (see Equation 4) except s_0. E.g., for three or more groups an experiment d(i) is defined using an adaption of Fisher's linear discriminant. For survival time analysis Cox's proportional hazard function is adapted, and for a quantitative parameter the Pearson correlation coefficient is adapted (Tusher et al., 2001).

The function was prepared for the Bioconductor platform and can be found in the packages *samr* and can be used for differential expression analysis in microarray and sequencing data (Chu et al., 2014; Tibshirani et al., 2011).

3.2.3 Two-sample Bayesian t-test

While in SAM the variance to calculate the test-statistic for a two-sample comparison is adapted by adding a fixed parameter s_0, empirical Bayes approaches estimate this parameter from the data to analyze. To do this a prior probability for the expression means and variances are estimated using a probabilistic Bayesian framework. Thereby information about dependent gene expression profiles can be used.

Fox and Dimmic (2006) developed a two-sample t-test for microarrays based on an empirical Bayes approach developed by Baldi and Long (2001) called CyberT (see also Hatfield et al., 2003), which uses the dependency of the gene expressions on a microarray to obtain prior variances and degrees of freedom and so gain additional information about the data. In short, they combine empirical variances of the genes on the microarray with local background variances of the neighboring genes (Baldi and Long, 2001).

Having m genes or transcripts in n replicates, the prior degrees of freedom v_0 are calculated by $v_0 = m(n-1)$, which are then used to calculate the prior variances from the sum of all sums of squared deviations (SSD) of an expression from the mean expression per gene (see Equation 5) (Fox and Dimmic, 2006).

Equation 5 Calculation of the prior variance σ_0^2 for a two-sample Bayesian t-test (Fox and Dimmic, 2006)

$$\sigma_0^2 = \frac{\sum_{k=1}^{m}\sum_{i=1}^{n}(y_{k,i}-\bar{y}_k)^2}{v_0} = \frac{\sum_{k=1}^{m}\sum_{i=1}^{n}(y_{k,i}-\bar{y}_k)^2}{m(n-1)}$$

The posterior SSD from the mean is then obtained by adding to the prior the total SSDs from the samples. Together with the posterior degrees of freedom $\mathrm{v_n}$ as sum of the degrees of freedom from the samples and the prior degrees of freedom a posterior variance σ_n^2 can be calculated as shown in Equation 6, which in turn is used to calculate the t-statistic (Fox and Dimmic, 2006).

Equation 6 Posterior variance for the two-sample Bayesian t-test (Fox and Dimmic, 2006)

$$\sigma_n^2 = \frac{\mathrm{v}_0\sigma_0^2 + (n_1 - 1)\mathrm{s}_1^2 + (n_2 - 1)\mathrm{s}_2^2}{\mathrm{v_n}} = \frac{\mathrm{SSD}_0 + \mathrm{SSD}_1 + \mathrm{SSD}_2}{n_1 + n_2 + \mathrm{v}_0 - 2}$$

3.2.4 Linear model for microarray analysis (LIMMA)

LIMMA is the microarray analysis equivalent of an ANOVA or a multiple regression (Smyth, 2004). For every gene or transcript a linear model is fitted to explain the variance of every gene by the specified predictors. The method improves the test statistic by testing if the null hypothesis, the coefficients equal 0, in similar fashion as the two-sample Bayes t-test for microarrays previously described. As gene expressions in a microarray experiment are not independent from each other a hierarchical model is defined which describes how the coefficients and expression variances vary across genes.

Therefore first, prior distributions for the coefficient and expression variances are estimated using an empirical Bayes approach by estimating the parameters from the data taking advantage of the dependency of gene expression in a microarray. The posterior values are then obtained by adjusting the prior values with the actual observed values. For the testing of every coefficient, information is borrowed from the microarray itself, thereby degrees of freedom are saved and thus more genes with differential expression can be detected (Smyth, 2004).

The ordinary test-statistic is calculated as follows (Smyth, 2004):

Equation 7 Calculation of the ordinary t-test test-statistic for testing regression coefficients (Smyth, 2004)

$$t_{gj} = \frac{\widehat{\beta}_{gj}}{s_g \sqrt{v_{gj}}}$$

For every gene g and contrast j, $\hat{\beta}_{gj}$ is the observed FC, s_g is the residual SD of gene g, and results together with $\sqrt{v_{gj}}$ in a standard error for $\hat{\beta}_{gj}$. For the moderated t-statistic the posterior SD $\tilde{s}_g$ is used for which information is taken from an empirical Bayes approach with s_0^2 as a priori variance i.e. a common variance for all genes and d_0 as a priori degrees of freedom and the residual SD (Smyth, 2004):

Equation 8 Calculation of the moderated t-statistic in LIMMA (Smyth, 2004)

$$\tilde{t}_{gj} = \frac{\widehat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}} = \frac{\widehat{\beta}_{gj}}{\sqrt{\frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}} \sqrt{v_{gj}}}$$

To apply the LIMMA analysis to various research questions an extensive Bioconductor package *limma* was created (Ritchie et al., 2015; Smyth, 2005). Here the overall moderated F-value of a general linear model is calculated as well as the moderated t-statistics for the comparisons specified together with the respective p-values and adjusted p-values are computed. Also a B-statistic is given, which specifies the log-odds for a transcript to be differentially regulated; using the log-odds a probability for the differential expression can be calculated (Smyth, 2005).

3.2.5 Rank Product Analysis

Based on the test-statistic of Smyth (2004) further methods have been developed; e.g., intensity-based moderated t-statistic by Sartor et al. (2006), which estimates posterior variances by local regression of the expression levels, or a fully moderated t-statistic developed by Yu et al. (2011), which adapts the prior variances by local regression of every expression.

A different approach was taken by Breitling et al. (2004) with the development of the calculation of rank products (RP). Gene expression in all arrays was ordered and ranked for every array. If a gene is randomly expressed, every gene has the same probability to be ranked first ($p = 1/n^k$, for k replicates and n genes). If a gene is up regulated in a group compared to the other group, it is likely to be ranked higher in one group than the other. This principle is the foundation for the RP method. Genes are ranked across an array and then across all ranks. To account for a possibly unbalanced number of genes per array the ranks are then divided by the number of genes in the array. In this manner a probability for up or down regulation can be obtained for single sample designs. For two samples design first all pairwise log-ratios or FCs are calculated and ranked. The so obtained RP value is the probability to observe a certain expression or FC at a certain rank. A permutation step allows the calculation of an expected RP value for the experiment. The average expected value is then used to calculate the percentage of false-positives (Equation 9), where rank(g) is the rank of the gene g by their RP values. The percentage of false-positives is an estimator for the FDR (Breitling et al., 2004).

Equation 9 Percentage of false positives q for gene g (Breitling et al., 2004).

$$q_g = \frac{E(RP_g)}{\mathrm{rank}(g)}$$

The RP method is a robust method as it is based on ranks and not influenced by outliers, and makes only weak assumptions. So equal variances in the arrays are assumed, which is an assumption met after normalization of the arrays. The RP methods does not require to calculate the variance for each gene or transcript and is therefore also applicable in experiments with a low number of replicates, where e.g. SAM fails to calculate significance levels (Breitling et al., 2004).

The RP method is prepared as Bioconductor package *RankProd* (Hong et al., 2006). The RP method is extended for the package to compare different microarray platforms or different origins of microarrays for meta-analysis purposes as described by Hong and Breitling (2008).

Here, some of the most common approaches in microarray analysis to detect differentially regulated genes have been presented. Starting with the definition of fold changes to to the comparison of expression levels at different states, continuing with the presentation of approaches adapting a common t-test with hierarchical Bayesian models, and concluding with approaches based on the formulation of ranks for which a variety of methods have been developed. All tests have in common that they are adapted for multiple testing while trying to improve power of the test by reducing the number of estimated parameters.

3.3 Prediction of clinical outcome with microarray data

RNA microarray data can be used to explain differences in phenotype by the use of differential gene expression. This information is the foundation for in-depth analysis of the gene transcripts involved via function gene expression tools, described hereafter. But gene expression analysis also can be used to facilitate diagnosis or predicting clinical outcome, for which a list of a few transcripts or genes may be helpful.

The viewpoint of the statistical analysis hereby differs from the differential gene expression analysis: differential gene expression analysis determines genes that are differential expressed in probes with different conditions. Therefore, differential gene expression is a result of different pre-existing conditions; in other words, the clinical condition explains the gene expression. In the prediction of a clinical outcome with microarray data; gene expression profiles are used to predict a clinical outcome. The classification approach with microarrays is challenging due to the large number of predictors (transcripts or genes) and small numbers of samples, which makes it difficult to identify the best classifying predictors (Tibshirani et al., 2002).

Different methods have been developed and different approaches were followed to use expression profiling for the diagnosis of diseases. Khan et al. (2001), e.g., used artificial neural networks to classify cancers to specific diagnostic categories. Gruvberger-Saal et al. (2004) adapted the use of artificial neural networks to predict a continuous outcome, i.e., they predicted estrogen-receptor protein values instead of just the estrogen-receptor status in breast cancer samples. Geveart et al. (2006) integrate clinical data with microarray data by using Bayesian networks. Tibshirani and colleagues (2002) developed a predictive analysis of microarrays (PAM) based on an microarray adaption of the t-test SAM to predict tumor types. A third approach is the use of partial least square (PLS) regression. Pérez-Enciso and Tenenhaus (2003) used PLS discriminant analysis to predict the clinical outcome of breast cancer.

In the following sections PAM and PLS are further described as both methods are available for the integration into R and are commonly used to evaluate clinical outcomes.

3.3.1 Predictive Analysis of Microarrays (PAM)

Tibshirani and colleagues (2002) adapted SAM as introduced in chapter 3.2.2 for the purpose of predicting tumor types by gene expression analysis, called predictive analysis of microarrays (PAM). Here, a similar estimation for d_i is used as in SAM (compare to Equation 4): with the difference that s_0, the median of all s_i, is also multiplied with m_k (Equation 10); s_i is the pooled within-class SD of gene i and results with the multiplication of m_k in the pooled standard error for the difference in gene expression from class k to the overall centroid.

Equation 10 Test-statistic in PAM (Tibshirani et al., 2002)

$$d_{ik}=\frac{\overline{x_{ik}} - \bar{x}_i}{m_k\ (s_i+s_0)}$$

It uses then a nearest shrunken centroids approach to identify the class centroid a gene belongs to. The t-statistic d_i is rewritten to explain the class centroid $\overline{x_{ik}}$, and then a threshold is set by which amount d_{ik} is reduced (“shrunk”). If a gene does not contribute to the nearest-centroid computation the shrunken centroid for all classes is the same. The higher the chosen threshold the lower is the number of genes near the class centroids or the lower the number of predicting genes. The threshold is determined by cross-validation and the misclassification error. Thus a minimal list of genes for predictive purposes can be obtained (Tibshirani et al., 2002).

PAM is prepared for microarray analysis as Bioconductor package *pamr* (Hastie et al., 2013b).

3.3.2 Partial least squares based prediction analysis

Partial least square (PLS) is a dimension reduction method coupled with regression suitable for high-dimensional, noisy data with a higher number of predictors than samples and high level of multicollinearity as often found in transcriptomics and proteomics data (Boulesteix et al., 2008; Boulesteix and Strimmer, 2007; Pérez-Enciso and Tenenhaus, 2003). Pérez-Enciso and Tenenhaus (2003) used PLS discriminant analysis (PLSDA) to predict the clinical outcome of breast cancer patients. Boulesteix and Strimmer (2007) review different PLS methods and give examples for the application of PLS in regression, classification, feature selection, survival analysis problems. PLS regression can be applied in cases of univariate or multivariate responses (Boulesteix and Strimmer, 2007).

In PLS regression first all predictors are summarized in PLS components, thus reducing the dimension of the predictors. In the second step, these PLS components are used to explain the response in an ordinary least squares regression (Boulesteix et al., 2008). For univariate responses the covariance of the predictors and response is calculated directly, but with multivariate responses first a common latent variable which explains the response variables simultaneously needs to be determined (Boulesteix and Strimmer, 2007).

The latent components explaining the covariation of the different transcripts are obtained by linear transformation of the expression matrix with a matrix of weights (Boulesteix and Strimmer, 2007). For the interpretation of the variable importance often the weight vector of the first component is taken; it maximizes the estimated covariance between predictors and response variables (Johansson et al., 2003).

The R-package *plsgenomics* (Boulesteix et al., 2012) implements PLS regression for regression problems, a combination of PLS and linear discriminant analysis for classification problems, and an algorithm to select variables for binary classification problems.

3.4 Functional gene annotation analysis

Microarray analysis could end with a list of genes, which are differentially expressed in different conditions, or are able to predict a certain outcome, but as an exploratory tool it is important for microarray analysis to end with a research question. Databases and exploratory pathway analysis facilitate the search for the right question for confirmatory approaches rather than the right answer (Kelder et al., 2010).

The challenge is now to gain insight into the biological mechanism of the studied conditions or diseases in order to understand what lead to the observed gene expression pattern or to what an observed expression pattern could lead (Hatfield et al., 2003; Subramanian et al., 2005).

3.4.1 Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) developed by Subramanian and colleagues (2005) is a method to determine whether genes at the top or bottom of a ranked list of genes or transcripts are involved in e.g. a common pathway, biological function, or a chromosome.

Therefore, first all genes are ranked according to their correlation with the phenotype. Gene sets are defined using different kind of databases, so gene sets can be defined by e.g. Gene Ontology (GO) categories or pathways of the Kyoto Encyclopedia of Genes and Genomes (KEGG). Then, it is determined whether genes of a certain gene set are enriched, i.e. present more often than by chance alone. The enrichment score (ES) in GSEA is determined by walking down the list of genes and calculation a running-sum statistic, where the sum is increased if the encountered gene is in the list and decreased if not. The statistical significance is determined by an empirical phenotype-based permutation test procedure, i.e. the phenotype labels are permuted, and so the ES is recalculated until a null-distribution for the ES is determined. In a final step results are adjusted for multiple hypothesis testing by using the false discovery rate (FDR) (Subramanian et al., 2005).

3.4.2 Database for Annotation, Visualization, and Integrated Discovery (DAVID)

The Database for Annotation, Visualization, and Integrated Discovery (DAVID, http://david.abcc.ncifcrf.gov/) is an open-source set of data-mining tools (Dennis et al., 2003; Huang et al., 2009). It uses different gene identification and annotation databases, e.g. GenBank, UniGene, RefSeq, LocusLink, KEGG, Online Mendelian Inheritance in Man (OMIM), or GO as referenced in Dennis et al. (2003) to link the genes in question to the data stored within those databases to a common DAVID identification number (Sherman et al., 2007). Based on this knowledgebase different exploration tools are implemented in DAVID (Huang et al., 2009):

Gene name batch viewer examines a list of genes, and finds genes with similar functions in the list by examining related genes.

Gene functional classification facilitates the identification of gene groups via ES and the function associated with these groups. The ES is calculated by the geometric mean of all enrichment p-values. The enrichment p-values in turn are calculated by a modified Fisher exact score (EASE score).

The **functional annotation chart** is more focused on the common biological functions of the genes in the analysis. Different functional annotation databases can be chosen by the user of DAVID. Again the EASE score is used to examine the significance of the gene-term enrichment, along with a correction for multiple hypothesis testing. Furthermore a fold enrichment measure is provided, which measures the percentage of genes in the analyzed gene list compared to the background of all gene eligible for selection, the so called background.

Functional annotation clustering focuses more on common themes in functional annotation. Redundant, similar, or hierarchical terms are clustered together by their percentage of common genes involved in these terms. An ES of 1.3 is a p-value of 0.05 on negative log-scale, an ES of 1 stands for a geometric average of p-values of 0.1. Every functional cluster with ES greater 1.3 or also 1 is a reasonable starting point for further investigations.

3.4.3 Ingenuity Pathway Analysis (IPA)

Ingenuity Pathway Analysis (IPA, www.ingenuity.com) is a commercial tool for gene list analysis with a manually curated database for gene annotation and gene functions. For the curation of the IPA knowledgebase various sources are used. The direction of the effect, i.e. upstream and downstream effect, as well as the tissue, species, and cell of the found interaction are included into the knowledgebase and are made accessible (Kramer et al., 2014).

An IPA core analysis consists of several analyses. First a background of genes available for selection is used. The IPA software offers several microarray platforms as implemented backgrounds as well as the option to use the user dataset as background. As the user data set typically does not cover all possible genes it is advisable to select this option. Otherwise significances are overestimated due to the fact that, e.g., a higher percentage of genes attributed to infection than to metabolism are spotted on a microarray. Then the significance level of the analysis is set. With this set of transcripts or genes upstream biological causes and downstream effects on biological and disease functions are analyzed with the help of a Fisher's exact test to test whether genes of a canonical pathway, biological, or toxicological functions are more frequently than expected by chance alone. In the same manner are upstream regulators analyzed, e.g., transcription factors, but also micro-RNA, cytokines, or any gene or small molecule, which affects gene expression (Kramer et al., 2014).

In addition to associating the list of analysis genes with upstream regulators and downstream effects, it is predicted, whether upstream regulators are putatively up or down regulated and whether downstream effects are activated or deactivated inferred by the causal relationship of the molecules in the analysis. For this purpose, along with the enrichment score determined by the Fisher's exact test p-value, an activation z-score is used. The activation z-score takes the consistency of the network of regulated genes and upstream regulators or downstream effects respectively into account and the consistency of the pattern in comparison with a random pattern (Kramer et al., 2014). Upstream and downstream analyses are then combined into regulator effects networks, if they share the same regulated genes. Regulator effects networks are built by using the respective p-value and activation cut-off z-scores. They help to focus research on putative mechanism underlying the observed gene expression pattern.

3.5 Statistical analysis workflow

For the study presented in chapter 4 different approaches were used (Figure 3-1). For each approach different methods that were described in the previous chapters were used. In this chapter the different approaches and the used statistical analyses are summarized. The presented workflow thereby allows to easily evaluate the order of those steps and to allocate their respective place within the workflow.

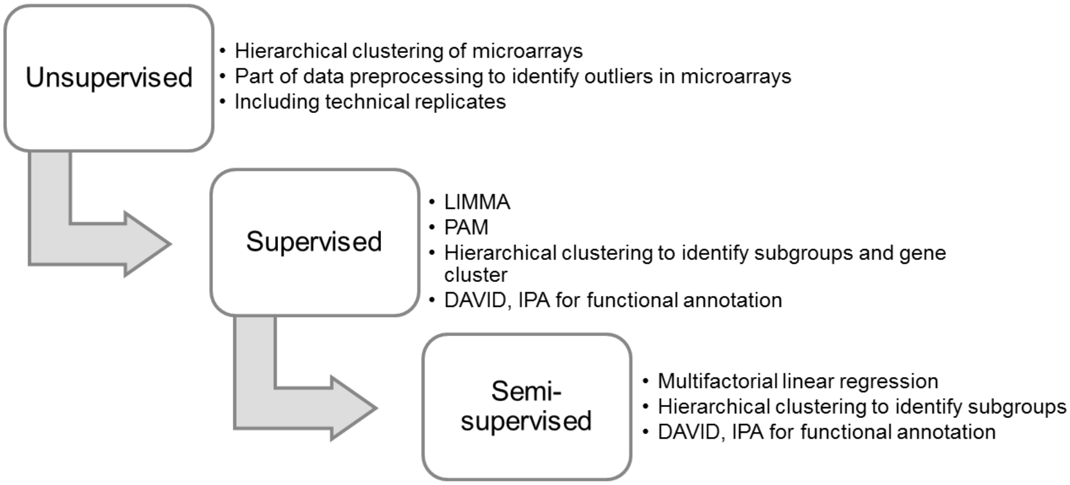
Unsupervised
• Hierarchical clustering of microarrays
• Part of data preprocessing to identify outliers in microarrays
• Including technical replicates
Supervised
• LIMMA
• PAM
• Hierarchical clustering to identify subgroups and gene cluster
• DAVID, IPA for functional annotation
Semi-supervised
• Multifactorial linear regression
• Hierarchical clustering to identify subgroups
• DAVID, IPA for functional annotation

Figure 3-1 Statistical analysis approaches with main statistical methods used in this investigation.

Unsupervised clustering. Starting with an unsupervised clustering approach to identify possible outliers in arrays or technical replicates as described in chapter 2.1.6., as last preprocessing step to check whether preprocessing can be completed. Here a hierarchical clustering method was used. As in an unsupervised approach no prior number of treatment groups was assigned for k-means clustering or SOM. Pearson's correlation coefficient for technical replicates and Euclidean distance for samples were used as measures of similarity between microarrays or samples respectively. We used Ward's linkage in accordance with Brevern et al. (2004) and Freyhult et al. (2010) as one of the most suitable method for analyzing microarray data.

Supervised approach. Then the LIMMA method as a supervised approach followed. The study population of 22 preterm infants was therefore divided into three treatment groups: i) 13 preterm infants with no BPD, ii) 6 with mild BPD, and iii) 3 with moderate/severe BPD. LIMMA was used to identify differentially regulated transcripts in the data set in order to identify processes at birth that are differentially activated between the infants. LIMMA is a suitable method in this context as the study population encompasses more than 2 treatment groups. Subsequently, for all pairwise comparisons the FC was calculated. Subgroups in treatment groups and co-regulated transcripts were analyzed using hierarchical clustering with Euclidean distance and Ward's linkage clustering as described above.

In a second step of the supervised approach, PAM was applied to find a set of transcripts able to predict to which BPD group the neonate belongs at birth. From this set of transcripts it may be possible to define a set of transcripts for diagnostic purposes. Here we used the obtained set of transcripts together with the FC from LIMMA to detect biological processes responsible for the development of mild or moderate/severe BPD.

LIMMA and PAM are especially designed for microarray analysis purposes and try to save degrees of freedom by taking the general correlation structure of gene expressions into account. LIMMA also allows the pairwise comparison of the defined groups based on overall F-test, which tests whether two of the three groups in this analysis show a difference in gene expression. In two-sample test as in the Bayesian t-test or Rank Product analysis this would not have been possible.

Semi-supervised approach. In a third approach, a semi-supervised approach was used. In order to identify transcripts correlated with continuous clinical factors, i.e. duration of assisted ventilation and duration of oxygen supply which show effects despite or independent from gestational age of preterm infants. Here, an unsupervised preprocessing workflow is followed by a selection of transcripts based on a linear regression model, with gene expression as independent variable and gestational age, duration of oxygen support, and duration of mechanical ventilation as explaining variables. This idea is derived from the partial least square (PLS) method. Although, for multivariate responses first a latent variable is calculated making it impossible to separate the effects of GA from the other effects. So for every gene expression a multifactorial regression model with oxygen support, mechanical ventilation and gestational age was fitted and all transcripts are selected that show an effect in oxygen support and/or mechanical ventilation or the interaction between both factors are selected. In further studies it will be interesting to adapt this approach for the correlation structure underlying microarray experiments as demonstrated in LIMMA and thereby save explanatory power.

All obtained sets of transcripts were analyzed for functional relevance using DAVID functional annotation clustering for clusters of differentially regulated transcripts. DAVID was used in order to detect general themes in clusters, as well as in up or down regulated transcripts. As DAVID gives no hint in regard to the activation or deactivation of biological processes and the involvement of upstream regulators, e.g. transcription factors, IPA pathway analysis was used to determine activation or deactivation of upstream regulators and downstream effects.

4 Transcriptional profiling of preterm infants with Bronchopulmonary Dysplasia (BPD) and integration of clinical data

As a proof of concept, the methods for preprocessing and biostatistical analyses of microarrays as described in section 2 and 3 were applied on a data set of preterm infants with bronchopulmonary dysplasia (BPD).

4.1 Background

Prematurity is defined as birth before the completion of 37 weeks of gestation and accounts for 35% of all neonatal death worldwide (Blencowe et al., 2013). Preterm birth increases the risk for acute and long-term complications. The development of chronic lung disease, i.e. bronchopulmonary dysplasia has been identified as a major determinant for both pulmonary and neurologic sequelae (Walsh et al., 2006). BPD occurs mainly in preterm infants and is defined as (1) mild BPD, when oxygen requirement persists at 28 days of life or (2) as moderate BPD when oxygen requirement is below 30% at 36 weeks postmenstrual age (PMA) or (3) as severe BPD when oxygen requirement is ≥ 30% and/or positive pressure, either as ventilation or continuous positive airway pressure (CPAP) is required at 36 weeks postmenstrual age (Jobe, 2006).

The main risk factor for the development of BPD is the degree of immaturity. Thus, preterm infants < 28 weeks gestational age (GA) are at greatest risk due to their functional and structural immature lung (Blencowe et al., 2013). Furthermore, pre- and postnatal infections, nutrition status and the impact of necessary postnatal therapies, i.e., oxygen supplementation and mechanical ventilation (MV) contribute to the development of the disease (Jobe, 2006; Speer, 2006; Thompson and Bhandari, 2008).

Potential downstream results of the indicated risk factors are sustained inflammation due to MV, hyperoxia, chorioamnionitis, infection (Ballabh et al., 2003; Bose et al., 2013; Köksal et al., 2012; Melville and Moss, 2013; Speer, 2006), and oxidative stress (Perrone et al., 2012; Saugstad, 2010). Oxidative stress is thereby defined as the incapacity of the antioxidant defense to bind free radicals, e.g., reactive oxygen species (ROS), which results in an overflow of free radicals. Free radicals occur due to oxygen treatment and/or inflammation. The oxidant damage in turn can be increased by a low calorie intake (Jobe, 2006). The pathophysiological characteristics are impaired alveolarization and vascularization, resulting in a simplified lung structure (Jobe, 2006; Melville and Moss, 2013).

Over the last years, research on BPD has started to focus on the transcriptome of preterm infants using microarrays. In 2007, Cohen et al. (2007) tested umbilical cord tissue of preterm infants with or without the development of BPD. They first studied the influence of GA with preterm infants < 27 weeks GA vs. infants with 27 to 28 weeks GA on the expression profile and found three pathways to be overrepresented, which were related to oxidative phosphorylation, mitochondrial energy metabolism, and DNA repair. In the comparison of preterm infants with and without BPD, defined as persistent oxygen requirement at 36 weeks PMA, they found pathways involved in bioenergy, histone acetyltransferase binding activity, and chromatin remodeling. However, these findings are not based on significantly differentially regulated genes, and can therefore result in a large set of false positives.

Kompass et al. (2010) used animal models, i.e., they used ventilated lung tissue of premature baboons, rats, and mice, to investigate the alterations in gene expression profiles after ventilation of the lung. They identified highly conserved transcriptional responses to mechanical ventilation. Activating transcription factor 3 (ATF3) and FBJ osteosarcoma oncogene (FOS) are differentially expressed across several models of ventilator-induced lung injury. Among the differentially regulated genes they found a set of genes overrepresented in the transforming growth factor-β (TGF-β) receptor signaling pathway.

Bhattacharya et al. (2012) investigated tissues of lungs obtained from autopsies of preterm infants with and without BPD. Significantly differentially regulated genes are involved in biological processes that included cell-cycle regulation, immune-cell regulation, i.e., immunodeficiency signaling and B-cell development, and processes specific to the lung, i.e., sonic hedgehog signaling and retinol metabolism.

Recently, Pietrzyk et al. (2013) examined the gene expression profiles of preterm infants and their alterations 5, 14, and 28 days after birth and identified potential pathways associated with the disease. They found the T-cell-receptor pathway as the most down regulated pathway. In addition to the T-cell-receptor pathway primary immunodeficiency is continuously down regulated on all days of observation. The identified pathways depended on disease severity and immaturity.

Except for Pietrzyk et al. (2013), all studies have in common that they focused on severe cases of BPD, no information can be obtained about the development of mild BPD. Although, Pietrzyk et al. (2013) use samples from preterm infants with different BPD severity grades, they don't distinguish between severity grades. Therefore it is one aim of our study to generate hypotheses in regard to whether and how the development of mild BPD differs from those cases without or severe forms of BPD.

With the use of umbilical cord tissue we want to generate hypotheses concerning mechanisms before and after birth predisposing infants to the development of BPD and diagnostic markers, that can be obtained at birth and differentiate between BPD severity grades.

This study will investigate whether transcriptional profiles can be found at birth of preterm infants that give a clue in regard to the development of BPD and especially in the mild forms of BPD. Cord blood of 22 preterm infants born before the 32nd of gestational age is analyzed using CodeLink Human 10k Bioarrays. The chosen cohort is especially suitable to further investigate the mild form of BPD due to the fact that the major part of preterm infants with BPD has developed a mild rather than a severe form of BPD.

Transcriptional profiles as well as biological processes and upstream regulators will be identified that differ between infant developing no BPD, mild BPD, or severe to moderate BPD. In addition, gene expression profiles explained by a model of the duration of mechanical ventilation, oxygen support, and gestational age is used to identify transcripts, whose gene expression can be explained by mechanical ventilation (MV) and oxygen (O_2) support despite the influence of the GA. Again cytokine upstream regulators and downstream biological effects are identified in order to generate hypotheses on the development of (mild) BPD and on possible biomarkers for BPD obtainable at birth.

4.2 Methods

The study has been approved by the legal ethical committee (File 79/01, University of Giessen, Germany).

4.2.1 Patient characteristics

Newborn infants ≤ 32 weeks GA, were prospectively included in the study. Exclusion criteria were premature rupture of membranes ≥ 3 weeks prior to birth leading to oligo- or anhydramnios, severe congenital malformations and the diagnosis of severe metabolic disorders. Furthermore, prepartum treatment of the mother with cytostatic or immunosuppressive medication other than for lung maturation, as well as postnatal treatment with corticosteroids in a dose ≥ 1 mg/kg body weight, led to exclusion of the neonate. Analysis of C-reactive protein (CRP), whole white blood count and microbiological examination of blood cultures, swabs, urine and stool samples were done in the first 72 hours of life. Patients were clinically re-evaluated in short intervals and continuously monitored for vital signs, i.e. heart rate, blood pressure, microcirculation and breathing pattern.

Patients were allocated to one of the three following groups according to the presence of persistent oxygen requirement or ventilatory support at 36 weeks postmenstrual age: (I) no BPD, (II) mild BPD, and (III) moderate and severe BPD according to the definition of Jobe and Bancalari (2001).

A total of 22 preterm infants were included in the gene expression analysis study. Of these, 13 preterm infants developed no BPD (group I), 9 developed BPD grade 1 (mild BPD, group II) and 3 preterm infants developed BPD grade 2 or 3 (moderate and severe BPD, group III).

Groups are matched for gender, CRP, intrauterine growth restriction (IUGR), antenatal corticosteroids, and incidents of placental chorioamnionitis (see Table 4-2 in the results section). GA could not be perfectly matched for the groups. Preterm infants of group I tend to be more mature at birth than preterm infants, who develop mild BPD, but have a similar GA as the preterm infants of group III. Infants of group I also tend to have more weight than preterm infants who develop BPD.

4.2.2 Microarray analysis

The PAXgene Blood RNA System (PreAnalytiX, Heidelberg, Germany) was used to collect whole blood samples and isolate the RNA according to the manufacturer's recommendations (PreAnalytiX). Total RNA was quantified with Nanodrop (NanoDrop Technologies, Rockland DE, USA) and the quality of RNA was assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies GmbH, Boeblingen, Germany). When the total RNA fulfilled quality criteria such as sufficient yield (> 2 µg), a 260/280-ratio of > 1.9 and electrophoretic profiles showing clear and sharp ribosomal peaks, the RNA was subjected to cRNA synthesis, cRNA fragmentation and hybridization on CodeLink UniSet Human 10 K Bioarrays (GE Healthcare, Freiburg, Germany) using the CodeLink Expression Assay Kit (GE Healthcare) according to the manufacturer's instructions. Each patient sample was hybridized on at least two Bioarrays (technical replicates). Bioarrays were stained with Cy5™-streptavadin (GE Healthcare) and scanned using the GenePix® 4000 B scanner and the GenePix Pro 4.0 Software (Axon Instruments, Arlington, USA). A total of 75 array images were subjected to data analysis.

Spot signals of CodeLink Bioarrays were quantified using CodeLink Expression Software V1.21 (GE Healthcare), as outlined in the user's manual. CodeLink Expression Software V4.1 generated raw data as well as background corrected and median- centered intra-slide normalized data. The intra-slide normalized data were used for further analysis. The software automatically calculated thresholds for intra-slide normalized intensities for each array and flagged probes as TRUE when the signal intensity was higher than the threshold or FALSE when the intensity was lower than the threshold. The present call of a microarray was given as the ratio of probes flagged as TRUE by the total number of probes on the microarray. Microarrays subjected to data analysis showed a mean present call of 81%, indicating a high number of probes above the threshold, i.e. being flagged as TRUE. Furthermore, the software flagged each probe value as GOOD, EMPTY, POOR, NEG or MSR, thus defining different quality measures as outlined in the user's manual.

Table 4-1 Number of replicates per sample amounting to 61 microarrays in the data set

Patient ID	1004	1005	1006	1038	1069	1073	1074	1080	1081	1082	1086	1087	1091	1133	1140	1149	1150	1157	751	816	896	912	**Total**
number of replicates	3	3	3	3	3	3	3	3	3	3	2	3	3	2	3	2	2	2	3	3	3	3	**61**

Preprocessing for microarray analysis was performed as discussed in chapter 2 in consideration of the grouping of the arrays. Outlier arrays and outliers in technical replicates were detected using spearman correlation coefficients and cluster analysis as described in further detail in chapter 2.1.6. For each sample 2 to 3 technical replicates were prepared. The final data set for data preparation consists of 61 microarrays (Table 4-1).

Microarrays were background corrected using background subtract and intra-slide normalized using Median normalization as recommended by the manufacturer. The 9945 transcripts were filtered for high rates of missing values, low expressed values and outlier values per probe over all microarrays as discussed in chapter 2.1 and is summarized in chapter 2.2. Transcripts with a rate > 50% missing values in a probe of all microarrays of a group (chapter 2.1.3.1), as well as transcripts with a rate > 50% of values below the negative control detection threshold (chapter 2.1.4) were discarded. Outliers were defined as values three times greater or smaller than the median (chapter 2.1.5). All transcripts with a rate > 50% with outlier values were also discarded. A total of 2416 probes were filtered. The remaining missing values were estimated using BPCA imputation (chapter 2.1.3.2). Data was subsequently inter-slide normalized using Quantile normalization. A total of 7529 probes were submitted to annotation and statistical analysis.

Transcripts were annotated using the SOURCE database[14] by Diehn et al. (2003) All transcripts are annotated with Genbank Accession number as provided by the manufacturer, a total of 6955 could be annotated with Unigene Cluster ID and a gene name, 6796 with Hugo gene symbol (6594 unique gene symbols) and Entrez gene ID.

For the *semi-supervised microarray analysis* no treatment groups were considered a priori. Background correction, normalization, as well as filtering, and imputation steps remained identical; a total of 2860 transcripts were filtered. For this part of the analysis a total of 7085 transcripts were obtained, which translate into 6537 annotated transcripts with 6386 gene symbols (6202 unique gene symbols).

[14] SOURCE database can be found at: http://smd.princeton.edu/cgi-bin/source/sourceSearch.

4.2.3 Statistical analyses

Analysis of microarray data

For microarray analysis umbilical cord tissue of the 22 preterm infants, further on also referred to as study population or cohort, was used. In chapter 3.5, three approaches to analyze microarray data have been developed: an unsupervised, supervised, and semi-supervised approach. As gene expression profiles associated with BPD or the need for prolonged MV or oxygen support should be identified the supervised and semi-supervised approaches are chosen for the present study.

(1) In the *supervised approach*, first the difference in gene expression between the treatment groups was detected with Linear Models for Microarray Analysis (LIMMA) (Smyth, 2005) using the Bioconductor package *limma* (Ritchie et al., 2015). The significance level was adjusted for multiple testing and the cut-off FDR for differential expression was set at FDR < 0.05 and a minimum absolute FC of 2 between at least two groups. Hierarchical clustering of the gene set and arrays was conducted with Euclidean distance measure and the Ward's linkage clustering method. Methods of this first step of the supervised approach are described in detail in chapter 3.2 - Prioritizing genes or differential gene expression analysis.

(2) The second step is based on the methods described in chapter 3.3 - Prediction of clinical outcome with microarray data: to identify transcripts able to predict to which group a preterm infant belongs as early as time of birth, a predictive analysis of microarrays (PAM) developed by Tibshirani et al. (2002) was performed using the Bioconductor package *pamr* (Hastie et al., 2013b).

(3) In a *semi-supervised approach* linear regression models were fitted in order to stratify preterm infants according the quantitative identifiers for BPD, i.e. their dependency for oxygen supplementation and ventilatory support, under consideration of the degree of immaturity (GA in weeks). Transcripts were selected when the correlation coefficients for O_2 and/or MV and/or interactions between those parameters were statistically significant with $p < 0.01$. Hierarchical clustering with Euclidean distance and Ward's linkage was performed to separate groups of samples and transcripts. Fisher's exact test and pairwise Wilcoxon tests were used to distinguish clusters of preterm infants for their BPD group, GA, O_2, or MV respectively.

Functional annotation

Cluster of co-regulated transcripts, identified by hierarchical clustering, were submitted to Database for Annotation, Visualization, and Integrated Discovery (DAVID, http://david.abcc.ncifcrf.gov/) functional annotation clustering in order to identify common biological processes in transcript clusters in regard to Gene Ontology Biological Processes.

To take the analysis a step further, the sets of transcripts obtained by the supervised and semi-supervised approaches were submitted to the Ingenuity Pathway Analysis (IPA) (Ingenuity®Systems, 2014). IPA can not only analyze the data for the enrichment of certain biological functions downstream of the expression pattern and upstream regulators, but also adds information in regard to activation and deactivation of functions and regulators. Sets of transcripts were obtained by pairwise comparison of BPD groups (supervised analysis) with an absolute FC > 2 and p-value < 0.05, and regression coefficients with p < 0.01. Significance level for enrichment was set at p < 0.05.

Furthermore, upstream regulators and downstream effects in biological functions are combined to identify regulator–effect networks (Kramer et al., 2014). Regulators and biological functions were considered if the showed an absolute activation z-score ≥ 1.5.

4.3 Results

4.3.1 Patient characteristics

Between the three groups of the cohort (no BPD, mild BPD, moderate or severe BPD) differences in regard to the GA, birth weight, frequency of congenital sepsis, duration of CPAP, and also in the duration of oxygen supply could be detected. Differences in oxygen supply were expected from the definition of BPD, and serve as a proof of principle. Statistically significant differences between the cases with moderate/severe BPD and cases with mild BPD could not be detected (Table 4-2).

Table 4-2 Patient characteristics of the study cohort.

		no BPD			mild BPD			moderate/ severe BPD			Sig.
		Mean (SE)	Median	N	Mean (SE)	Median	N	Mean (SE)	Median	N	
Gestational age		29.99 0.28	30.29_{a}	13	26.14 1	24.71_{b}	6	27.52 1.61	$27.57_{a,b}$	3	0.009
Birth weight		1382 74.2	1400_{a}	13	932 145.15	770_{b}	6	960 105.04	1060_{b}	3	0.011
Length / cm		38.11 1.11	40	9	33.6 2.23	33	5	37 1	37	2	0.182
Head circumference / cm		27.06 0.56	27	9	23.8 1.03	23	5	25.5 1	25.5	2	0.071
CRP value / mg/l		9.58 1.41	9.75	4	46.83 23.36	35.7	3	40.25 29.05	40.25	2	0.112
Duration of mechanical ventilation (days)		6 0.93	7	7	5.5 0.76	5.5	6	17.33 13.45	7	3	0.858
CPAP days		3.55 1.03	2.00_{a}	11	22 6.04	21.00_{b}	6	6 1.15	$6.00_{a,b}$	3	0.007
O_2 days		6.08 1.62	4.00_{a}	13	46 6.45	43.50_{b}	6	80.33 30.02	66.00_{b}	3	0
Gender	male	7_{a}			1_{a}			1_{a}			0.353
	female	6_{a}			5_{a}			2_{a}			
Growth (IUGR)	no	13_{a}			5_{a}			3_{a}			0.409
	yes	0_{a}			1_{a}			0_{a}			
Congenital sepsis	no	7_{a}			0_{b}			$0_{a, b}$			0.041
	yes	6_{a}			6_{b}			$3_{a, b}$			
Antenatal corticosteroids (ANCS)	no	2_{a}			2_{a}			1_{a}			0.657
	yes	6_{a}			1_{a}			2_{a}			
ANCS - 7 days to 24 h before birth	no	5_{a}			3_{a}			1_{a}			0.779
	yes	6_{a}			1_{a}			2_{a}			
Clinical suspicion of amniotic infection syndrome (AIS)	no	12_{a}			$4_{a, b}$			0_{b}			0.005
	yes	1_{a}			$2_{a, b}$			3_{b}			

Significances were determined by Kruskal-Wallis-Test for quantitative parameters and Fisher's exact test for qualitative parameters. Every index (a, b) indicates a subset of BPD study groups, which do not differ from each other on a 5%-significance level (pairwise Wilcox-Test with FDR correction for multiple testing).

It becomes apparent that the preterm infants in this study without congenital sepsis do not develop BPD, while the infants with congenital sepsis develop additionally BPD. Differences in the number of preterm infants with congenital sepsis can be observed between no BPD preterm infants and mild BPD preterm infants. Differences in the number of infants with suspicion of an amniotic infection syndrome (AIS) can be observed: no BPD infants tend to have less AIS than infants with moderate/severe BPD. Differences between the groups can also be observed in GA, birth weight, duration of CPAP ventilation, and the duration of O_2. The preterm infants in our study without BPD tend to be older, to have a higher birth weight, to be in less need for CPAP and O_2, and have fewer cases of congenital sepsis than mild BPD preterm infants. This is also true for the comparison with moderate/severe BPD preterm infants in terms of weight and oxygen assistance. Additionally can for all cases of moderate/severe BPD an AIS be suspected.

4.3.2 Differential expression analysis of BPD severity groups using LIMMA

Supervised clustering reveals distinct gene expression pattern discriminating preterm infants with lower and higher BPD grades at birth.

The supervised approach of the microarray analysis revealed 238 differentially regulated transcripts with a FDR < 0.05, and |FC| > 2 out of 7529 transcripts (see Figure 4-2). Hierarchical clustering of samples with these differentially regulated genes reveals two main clusters of expression profiles. One cluster contains all microarray patterns of preterm infants with mild BPD. The other cluster can be divided into a subcluster of no BPD preterm infants and a subcluster, which contains all samples of preterm infant of group 3. The expression patterns indicate that processes at birth in the group of no BPD and moderate to severe BPD are more similar to each other than to the group of patients, who developed a mild BPD.

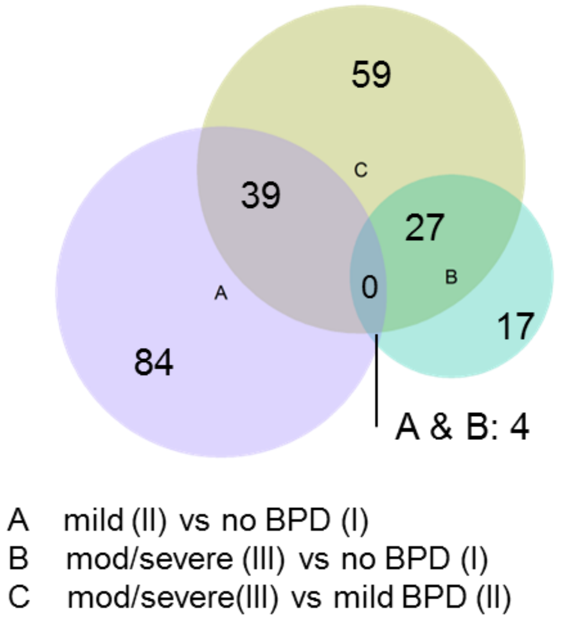

Figure 4-1 Total of 238 differentially expressed genes in comparison of three BPD severity grades (no, mild, and moderate-severe BPD)

Most of the differentially expressed transcripts can be found in the comparison of preterm infants without BPD and with mild BPD (Figure 4-1). In this comparison, 127 transcripts are differentially expressed, 84 are up regulated, and 43 are down regulated in infants with mild BPD compared to preterm infants without BPD. Preterm infants with mild BPD also show a distinct pattern compared to preterm infants with moderate or severe BPD with 125 (62 up, 63 down regulated) significantly differentially regulated transcripts.

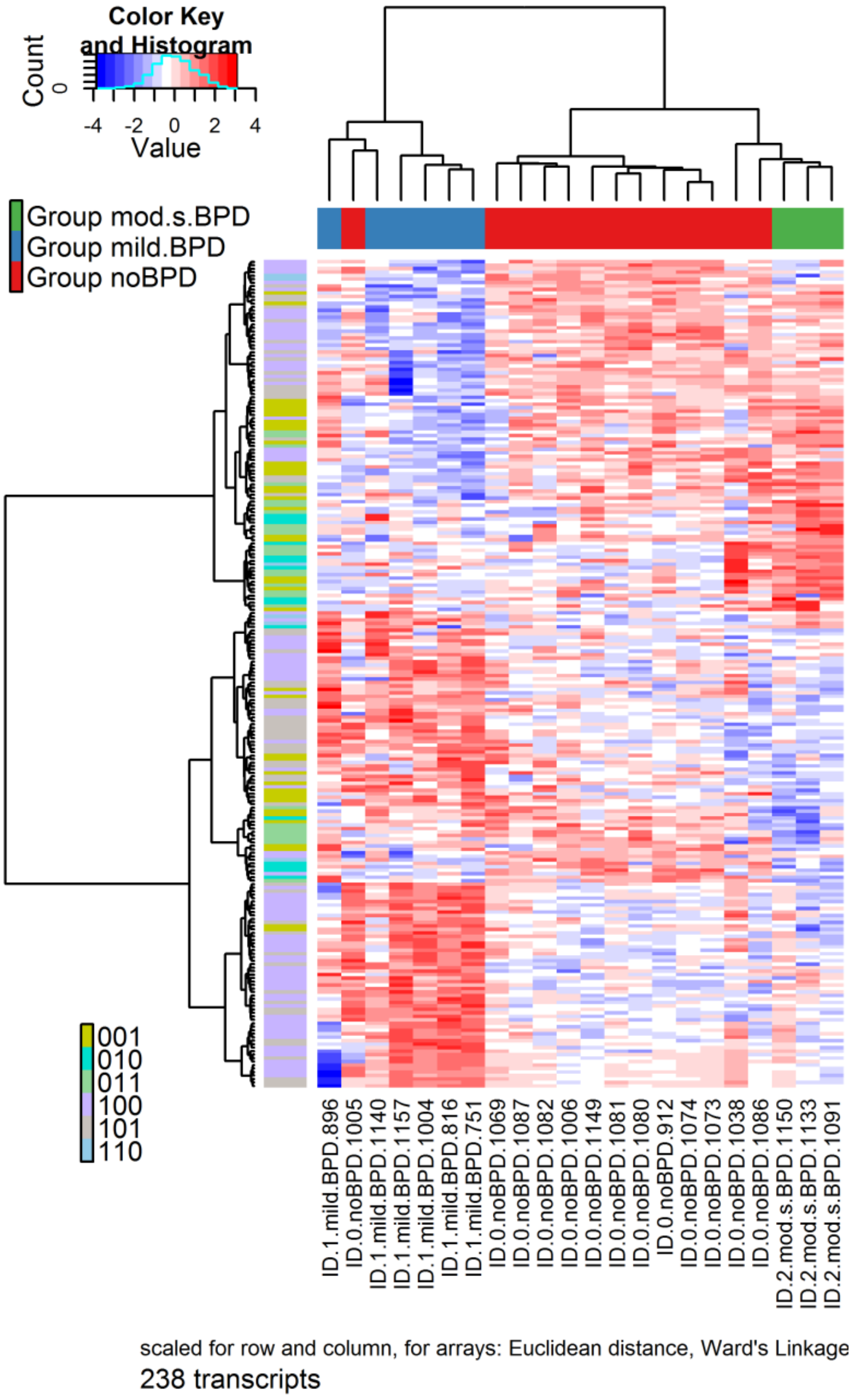

Figure 4-2 Expression profiles of significantly differentially regulated genes in preterm infants reveal two main hierarchical clusters of expression.

Preterm infants with BPD grade 1 are depicted in blue, with BPD grade 2 or 3 in green, and infants without BPD in red in the upper hierarchical clustering dendrogram. Hierarchical clustering was performed on scaled data with Euclidean distance measure and Ward's Linkage clustering method.

A high number of differentially regulated genes in the comparisons of group II vs. I and group II vs. III of 39 transcripts indicates that the gene expression pattern in preterm infants developing mild BPD differs greatly from no BPD and moderate to severe BPD gene expression. The development of mild BPD follows different routes than the development of more severe BPD. The most down regulated transcripts in mild BPD are CRIP1 (Cysteine-rich protein 1 (intestinal)), NM_005129, PTPRCAP (Protein tyrosine phosphatase, receptor type, C-associated protein); the most up regulated transcripts in mild BPD are GAL (Galanin prepropeptide), SLC24A3 (Solute carrier family 24 (sodium/potassium/calcium exchanger), member 3), and CDH13 (Cadherin 13, H-cadherin (heart)) (see Table 8-1 in

Appendix). DAVID functional annotation clustering indicates an overrepresentation of these genes associated with the cytoskeleton and calcium homeostasis. Transcripts which are only significantly differentially regulated in group II (mild BPD) compared to group I (no BPD) preterm infants are putatively overrepresented in the *membrane/glycoprotein*, *transmembrane transport*, and *leukocyte activation*. Transcripts only differentially regulated between group II and group III (moderate/severe BPD) are putatively overrepresented in the *mitochondrion* or *oxidation reduction*, and *regulation of cell migration*.

Group III preterm infants express 27 genes differentially, when compared to preterm infants who develop mild or no BPD (see Table 8-2 in

Appendix). The most up regulated transcripts at birth in preterm infants at birth with moderate/severe BPD are HPR (Haptoglobin-related protein), MAP4K3 (Mitogen-activated protein kinase 3), and PGLYRP1 (Peptidoglycan recognition protein 1), CDA (Cytidine deaminase). The most down regulated genes in moderate or severe BPD are FKBP14 (FK506 binding protein 14), GNG11 (Guanine nucleotide binding protein (G protein), gamma 11), and ETNK1 (Ethanolamine kinase 1). DAVID functional annotation clustering indicates an overrepresentation of genes associated with *transcription* or *transcription factor activity*.

Group III preterm infants differ in the expression of 48 transcripts (32 up regulated, 16 down regulated) from infants without BPD. A total of 4 transcripts, i.e. ACTN2 (Actinin, alpha 2), NM_003832, NM_018104, and SLC2A11 (Solute carrier family 2 (facilitated glucose transporter), member 11), is able to differentiate between mild BPD and no BPD preterm infants, and is also differentially regulated in preterm infants with moderate/severe BPD and infants without BPD (see Table 8-3 in

Appendix). Transcripts only significantly differentially regulated in group III preterm infants compared to no BPD infants are putatively overrepresented in the biological processes: *induction of apoptosis*, *regulation of transcription*. For the transcripts differentially regulated in both group II and group III preterm infants compared to no BPD infants, no functional clustering analysis can be conducted.

Principal component analysis reveals a clear distinction between preterm infants without BPD, mild BPD, and moderate/severe BPD (Figure 4-3). In hierarchical cluster analysis it can be seen that the gene expression pattern of preterm infants without BPD is more similar to gene expression in preterm infants with moderate or severe BPD than with mild BPD.

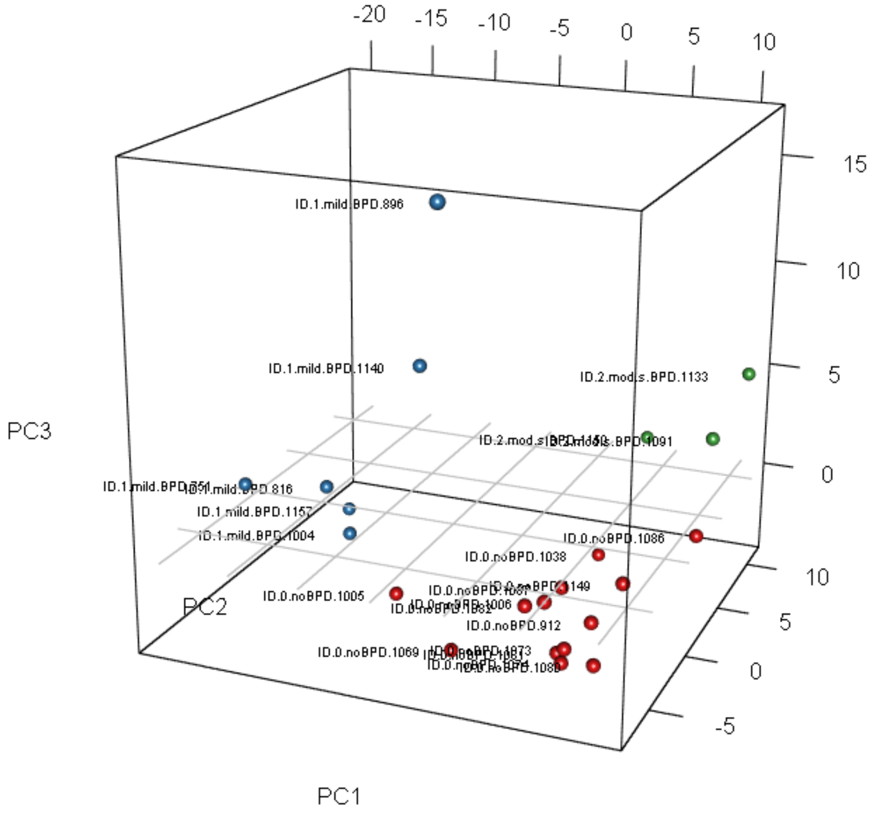

Figure 4-3 **Principal components analysis using 238 differentially regulated genes identified by gene expression analysis of BPD severity groups**

At time of birth differentially regulated genes in preterm infants are involved in chemotaxis and leukocytes.

An IPA comparison analysis for the unsupervised analysis approach predicted (1) activated chemotaxis of cells and (2) decreased biological functions associated with apoptosis and accumulation of leukocytes in preterm infants developing mild BPD (vs. no and moderate/severe BPD) as well as (3) decreased phagocytosis in infants with moderate/severe BPD (vs. no BPD) (see Table 4-3).

Table 4-3 Biological functions predicted to be increased or decreased at time of birth in preterm infants developing BPD

Shown are functions with z-scores ≥ |1.5|, which indicate the direction of activation; processes with positive z-scores are predicted to increased, negative z-scores indicate a decreased function; asterisks indicate significance level of enrichment: * p < 0.05, ** p < 0.01, * p < 0.001.

Diseases and Bio Functions	mild vs. no BPD		mod./s. vs. no BPD		mod./s. vs. mild BPD	
accumulation of cells	-2.42	*				
apoptosis of cancer cells	-2.23	*				
damage of kidney	-2.22	*				
accumulation of granulocytes	-2.22	**				
cell death of cancer cells	-2.22	*				
accumulation of leukocytes	-2.20	**				
accumulation of eosinophils	-1.98	***				
weight loss	-1.71	*				
phagocytosis			-1.73	*		
branching of neurites					1.96	*
cell movement of vascular smooth muscle cells					-2.00	*
chemotaxis of cells	2.66	*			-1.86	*
aggregation of cells	1.95	**			-0.73	*
chemotaxis of leukocytes	2.28	*				
quantity of granulocytes	2.18	*				
chemotaxis of mononuclear leukocytes	2.11	*				
quantity of neutrophils	2.06	*				
airway hyperresponsiveness	1.85	**				
chemotaxis of lymphocytes	1.83	*				

Cytokines as upstream regulators in mild BPD

Cytokines play an important role in the innate immunity of preterm infants (Melville and Moss, 2013). Therefore the upstream regulator analysis (see chapter 3.4.3 functional gene annotation analysis with IPA) is focused on cytokines with the aim to identify potential cytokines as biomarkers present in the cord blood of preterm infants (see Table 4-4).

Table 4-4 Cytokine upstream regulators (IPA) predicted to be activated or deactivated at birth of preterm infants developing BPD

Upstream regulators	mild vs. no BPD		mod./s. vs. no BPD		mod./s. vs. mild BPD	
IL2	**1.80**					
TNF	**1.66**	**				
IL6	**1.52**	*	0.00	*		
IL10	**1.21**	*				
IFN alpha/beta	**1.13**	**				
IFNG	**0.56**	*				
CCL19	0.00	**				
CCL8	0.00	*				
CXCL9	0.00	*				
Ifn gamma	0.00	**				
IFNE	0.00	**				
IFNK	0.00	*				
IFNW1	0.00	**				
Mac	0.00	*				
TSLP	0.00	*				
CSF3			0.00	*		
EBI3			0.00	**		
IFN Beta			0.00	*		
IFNA1/IFNA13			0.00	*		
IFNA10			0.00	*		
IFNA14			0.00	*		
IFNA17			0.00	*		
IFNA21			0.00	*		
IFNA4			0.00	*		
IFNA5			0.00	*		
IFNA6			0.00	*		
IFNA7			0.00	*		
IFNA8			0.00	*		
IL27			0.00	**		
IL5					**2.00**	
IL8			0.00	*		
WNT1					0.00	*

Shown are z-scores, which indicate the direction of activation; regulators with positive z-scores are predicted to activated (red); asterisks indicate significance level for enrichment: * $p < 0.05$, ** $p < 0.01$, * $p < 0.001$.

In mild BPD compared to no BPD TNF-α and interleukins (IL-2, IL-6, and IL-10) are the highest activated cytokines, but also interferons and chemokines are predicted to be activated. Differentially regulated genes in moderate/severe BPD infants on the other hand are mainly regulated by interferons.

Networks combining differentially expressed genes, upstream regulators, and downstream effects, so called regulator effects networks, of differentially regulated genes in preterm infants with mild BPD compared to infants with no BPD firstly demonstrated a relationship between the activation of IL-6, TCR (T cell receptor), TNF-α and the regulation of CXCL9 (chemokine ligand 9), IL-10, LAT (Linker for activation of T cells), LGALS3 (lectin, galactoside-binding, soluble, 3), MMP7 (matrix metallopeptidase 7), TLR3 (toll-like receptor 3), TNFRSF1A (tumor necrosis factor receptor superfamily, member 1a) (Figure 4-4). The regulation of these genes is linked to a predicted activation of the function chemotaxis of leukocytes, and the deactivation of the accumulation of eosinophils and fibrosis.

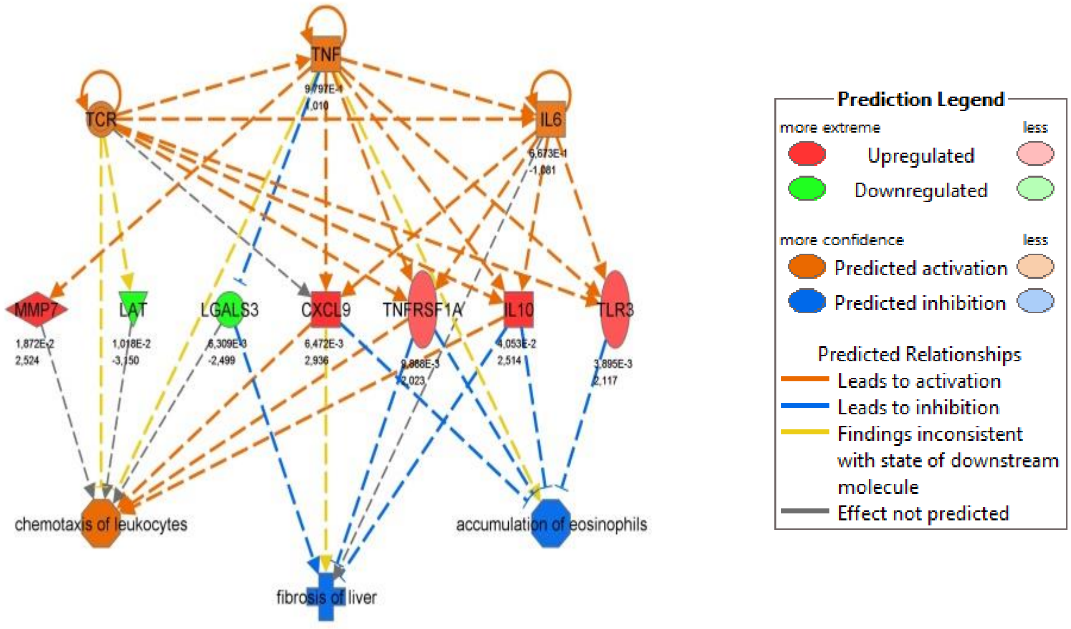

Figure 4-4 **Regulator effector networks linking TCR, TNF-α, IL-6 activation to activation of chemotaxis of leukocytes and deactivation of fibrosis and accumulation of eosinophils in preterm infants with mild BPD compared to preterm infants without BPD**

Secondly, a relationship between the predicted activation of MAPK14 (mitogen-activated protein kinase 14), the differential expression CAT (catalase), IL-10, SREBF1 (sterol regulatory element binding transcription factor 1), ZFP36 (zinc finger protein 36), and the predicted deactivation of disease functions leading to apoptosis and necrosis (damage of kidney, cell death of cancer cells) was found (Figure 4-5).

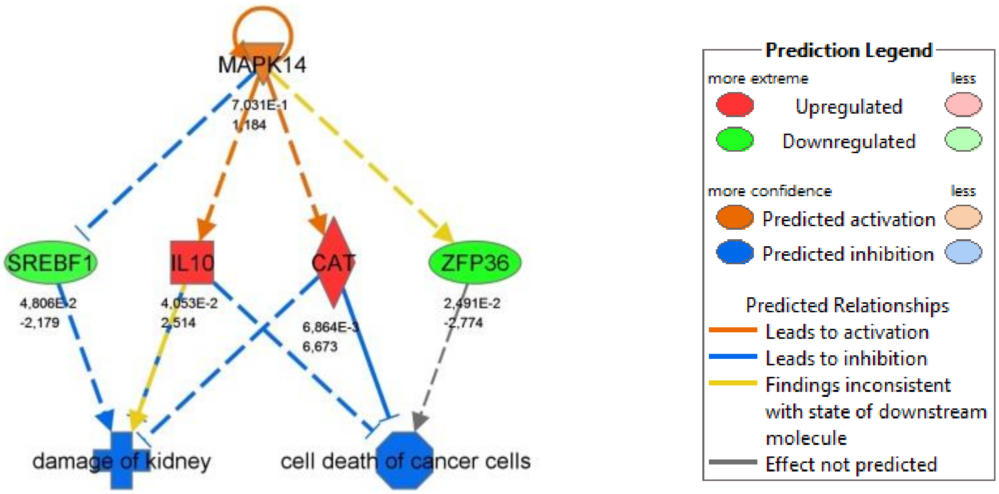

Figure 4-5 **Regulator effector networks linking MAPK14 activation and the inhibition of functions leading to apoptosis and necrosis in preterm infants with mild BPD compared to preterm infants without BPD**

The activation of TNF-α was furthermore predicted by the expression of CAT, CDH13 (cadherin 13), IL-10, LGALS3, TNFRSF1A which in turn had been linked to airway hyperresponsiveness in preterm infants with mild BPD (Figure 4-6).

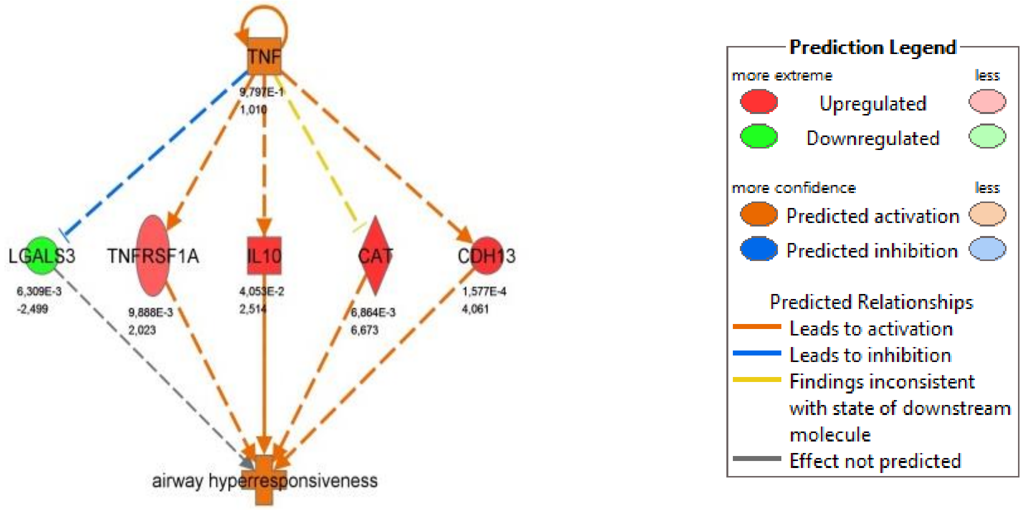

Figure 4-6 Regulator effector networks linking TNF-α activation to the activation of airway hyperresponsiveness in preterm infants with mild BPD compared to preterm infants without BPD

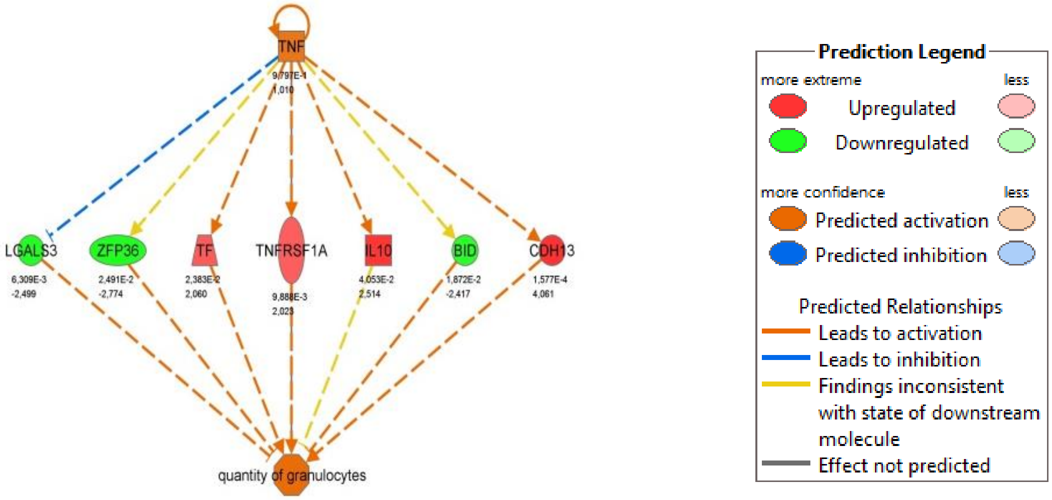

Figure 4-7 Regulator effector networks linking TNF-α activation to the activation of proliferation of granulocytes in preterm infants with mild BPD compared to preterm infants without BPD

An additional prediction analysis showed the activation of TNF-α by BID (BH3 interacting domain death agonist), CDH13, IL-10, LGALS3, TF (transferrin), TNFRSF1A, ZFP36, involved in granulocytes proliferation (Figure 4-7).

Increased neutrophil number is predicted as a consequence of up-regulated IL-10, TNF-α, TNFRSF1A through the regulator IL-6 (Figure 4-8).

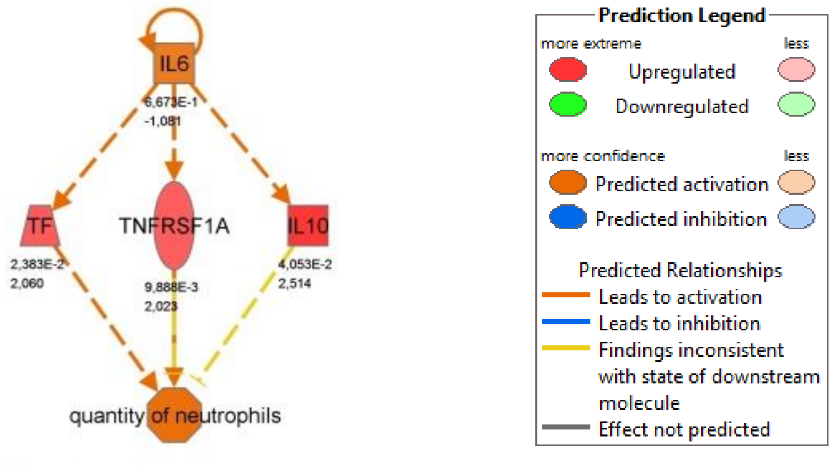

Figure 4-8 Regulator effector networks linking IL-6 activation to the activation of proliferation of neutrophils in preterm infants with mild BPD compared to preterm infants without BPD

4.3.3 Predictive Microarray Analysis (PAM) for preterm infants with and without BPD

In order to be able to predict the severity of BPD from gene expression at the time of birth a PAM was conducted. A set of 71 transcripts is obtained using the methods described in chapter 3.3 especially PAM (chapter 3.3.1), that is able to discriminate between the three groups of preterm infants (Figure 4-9 A; Table 8-4 in supplemental material).

Of these 71 transcripts 58 transcripts also are differentially regulated (see Figure 4-9 C). For DAVID functional annotation and gene functional classification 47 transcripts could be converted into DAVID IDs. Gene classification could not find classes of genes with a suitable enrichment score. DAVID functional annotation clustering (see Table 8-5 in supplemental material) shows an enrichment of genes involved in regulation of leukocyte activation (ES: 1.65, involved genes: GAL, HLX, ZEB1, LAT), and regulation of cell proliferation/ embryonic organ development (ES: 1.19, involved genes: GAL, CDH13, BTG3, ALDH1A2, PRTN3, HLX, KLF5, ZEB1).

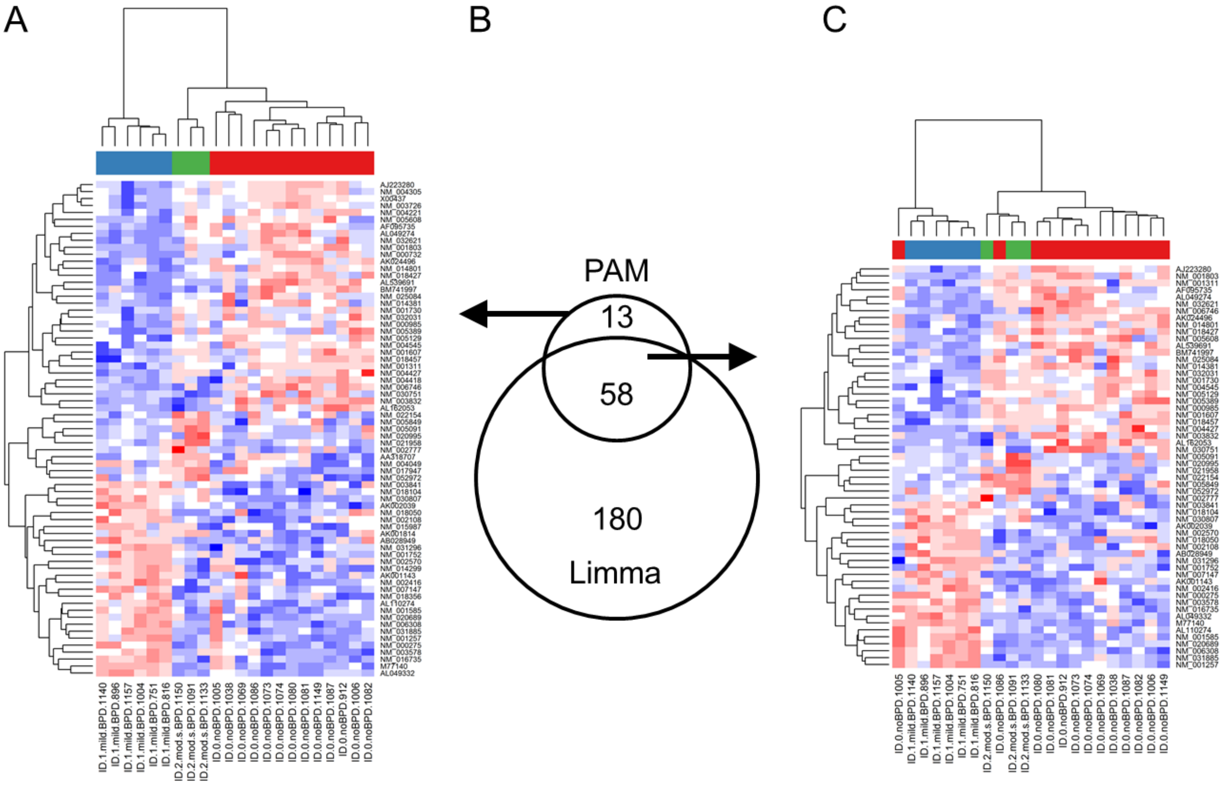

Figure 4-9 **Expression profiles of transcripts that are able to differentiate between groups of BPD preterm infants.**

A: heatmap of all 71 predictive transcripts,

B: Euler Venn diagram shows the overlap of transcripts between PAM analysis for predictive transcripts and LIMMA analysis for differentially regulated transcripts,

C: heatmap of 58 predictive and differentially regulated transcripts.

Cluster analysis revealed a pattern comparable to the one identified by the LIMMA analysis with a high similarity between the transcriptome expression pattern of infants with moderate/severe BPD and no BPD. The clinical data of the study cohort also showed a similarity between moderate/ severe and no BPD and difference to mild BPD in GA, number of congenital sepsis, duration of non-invasive positive pressure ventilation, and CPAP (Table 4-2).

Down regulation of reactive oxygen species prevents the development of higher grade BPD

From the 71 transcripts that are able to discriminate between the different BPD severity states, 66 transcripts could be annotated. IPA toxicity analysis indicates that transcripts associated with oxidative stress are overrepresented in this set of genes. Genes involved in *synthesis of reactive oxygen species (ROS)* and *production of ROS* are overrepresented in this analysis, but differ between the groups in the direction of regulation (Figure 4-10).

Compared to preterm infants with mild BPD, the production of ROS is predicted to be activated in no BPD (activation z-score$_{\text{I vs. II}}$ = 1.6) and in moderate to severe BPD ($z_{\text{III vs. II}}$ = 1.6). No difference in activation state could be found between group I and III. Together with the activation of the inflammatory response and chemotaxis of cells in BPD the activation state of these functions are able to differentiate BPD severity groups. Activated in both group II and group III compared to group I (no BPD) are the biological functions inflammatory response ($z_{\text{II vs. I}}$ = 2.0, $z_{\text{III vs. I}}$ = 1.4) and chemotaxis of cells ($z_{\text{II vs. I}}$ = 1.5, $z_{\text{III vs. I}}$ = 1.7).

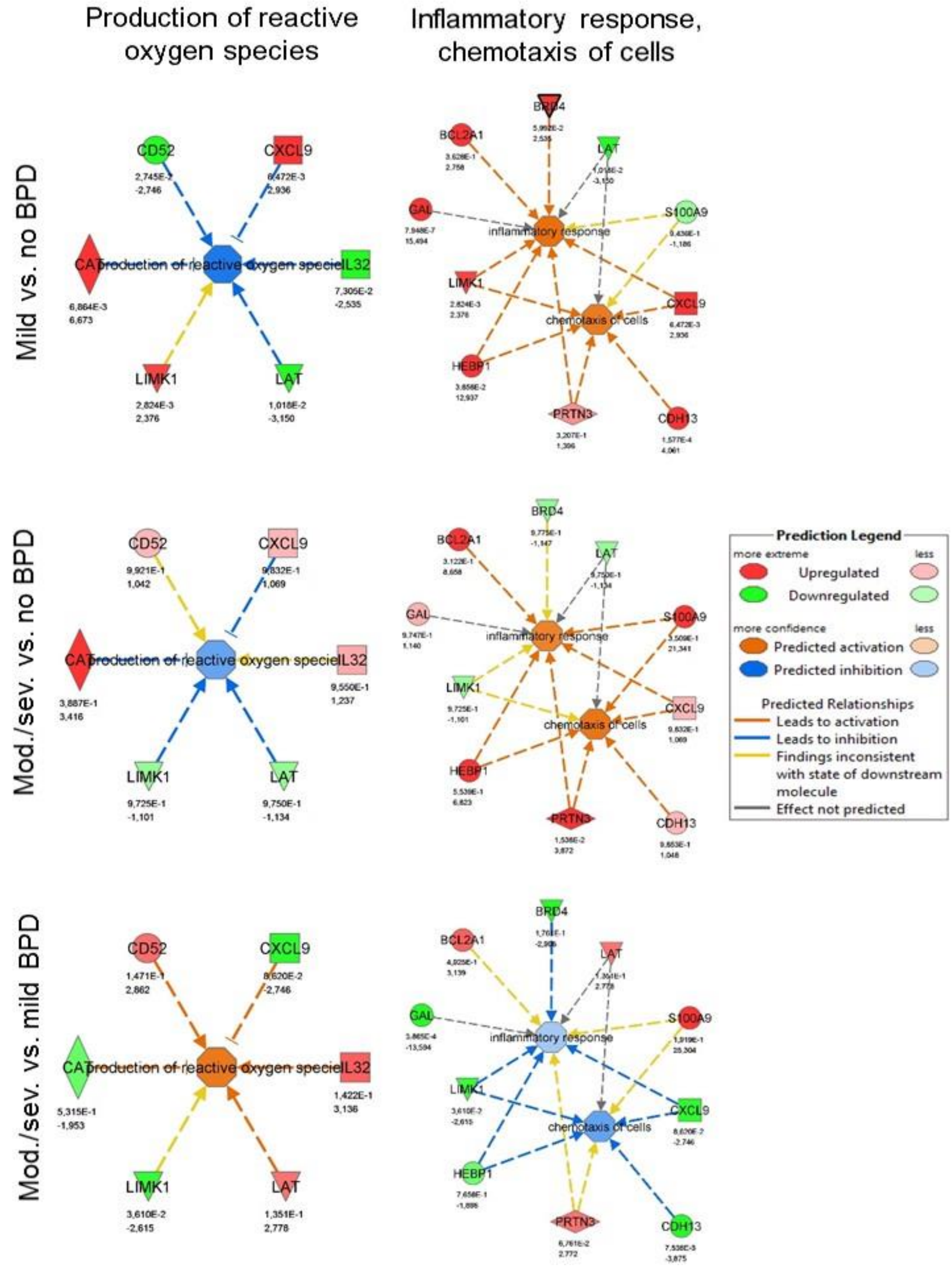

Figure 4-10 **Biological processes *inflammatory response* and *chemotaxis of cells* are predicted to be activated in both BPD groups compared to no BPD and the *production of reactive oxygen species* is predicted to be deactivated in mild BPD compared to no BPD, but is predicted to be active in moderated-severe BPD compared to no BPD.**

In IPA upstream analysis, again TNF-α can be identified as regulator of more genes than could be expected by chance alone. Additionally, interleukins, chemokines, and interferons can be identified as putative cytokine upstream regulators (Table 4-5).

This predictive analysis for microarray analysis emphasizes the role of cytokines and inflammatory processes at birth in preterm infants. To secure that these findings are not solely based on the effect of GA, we correlate gene expression with O2 and duration of MV under consideration of the GA of the infants as outlined in chapter 3.5 in the semi-supervised approach.

Table 4-5 Cytokine upstream regulators (IPA) predicted to be activated or deactivated at birth of preterm infants developing BPD of genes able to discriminate between BPD groups (PAM, threshold=2.2)

			Z-scores			
	Mild BPD vs.		**mod. /sev. BPD**		**mod. /sev. BPD**	
Upstream regulators	**no BPD**		**vs. no BPD**		**vs. mild BPD**	
TNF	**0.29**	**	**1.97**	**	**1.50**	**
Csf	0.00	**	0.00	**	0.00	**
Interferon alpha	**-0.74**	**	**1.23**	**	0.00	**
IL32	0.00	*	0.00	*	0.00	*
CCL19	0.00	*	0.00	*	0.00	*
CCL21	0.00	*	0.00	*	0.00	*
IL4	0.00	*	0.00	*	0.00	*
IL6	**0.45**	*	**1.34**	*	**0.45**	*
IL12 (complex)	0.00	*	0.00	*	0.00	*
CCL3L1/CCL3L3	0.00	*	0.00	*	0.00	*
Ifn gamma	0.00	*	0.00	*	0.00	*
IL3	0.00	*	0.00	*	0.00	*

Shown are z-scores, which indicate the direction of activation; asterisks indicate significance level for enrichment: * $p < 0.05$, ** $p < 0.01$, * $p < 0.001$.

4.3.4 Semi-supervised approach – regression model

In order to further separate preterm infants with different BPD grades and to identify genes associated with prolonged ventilatory support and/or oxygen supply under consideration of the degree of immaturity, an advanced linear regression model was designed. A total of 210 genes with significant correlation to the duration of O_2 and/or to MV or interactions between O_2 and MV were selected at a 1%-significance level (Table 8-4). A total of 17 transcripts were correlated with MV, 83 transcripts with O_2, 1 showed correlation with an additive effect of MV and O_2, and 109 were correlated with the interaction effect of MV and O_2 (Table 4-6). Of these 210 transcripts 55 (7 in MV, 32 in O_2, 16 in interaction) were also correlated with GA.

Table 4-6 Number of transcripts correlated with oxygen supply, mechanical ventilation, and interaction between oxygen and mechanical ventilation on a 1%-significance level.

number of transcripts	with effects by gestational age (GA)		without effects by GA	
	single effects	additive effects	single effects	additive effects
oxygen (O2)	83	1	51	1
mechanical ventilation (MV)	17		10	
interaction O2:MV	109		93	
Total	**210**		**155**	

Unsupervised microarray analysis identifies infants with different BPD grades at birth

By the use of the above indicated, unsupervised analysis, three main clusters of preterm infants were identified, which could be assigned to either ‘BPD’ or ‘no BPD’ (Fisher’s exact test, p-value = 0.032; Figure 4-11).

Cluster 1 comprises transcriptome profiles from preterm infants with no BPD (median GA 30.9 weeks, IQR = 0.5), short duration of O2 (median 1 day, IQR = 0.75), and no history of ventilatory support exceeding 48 hours. Cluster 2 comprises the transcriptome pattern of preterm infants with no BPD (median GA 29.9 weeks, IQR = 1.2), a median duration for

O2 of 16 days (IQR = 17.25) in addition to a history of ventilatory support (median 7 days, IQR = 1.5). Cluster 3 comprises preterm infants with all grades of BPD, but preterm infants in this cluster show a higher degree of immaturity (median GA of 24.7 weeks, IQR = 3), a median duration for assisted ventilation of 4 days (IQR = 4.8) and a median duration of oxygen supplementation of 49.5 days (IQR = 36).

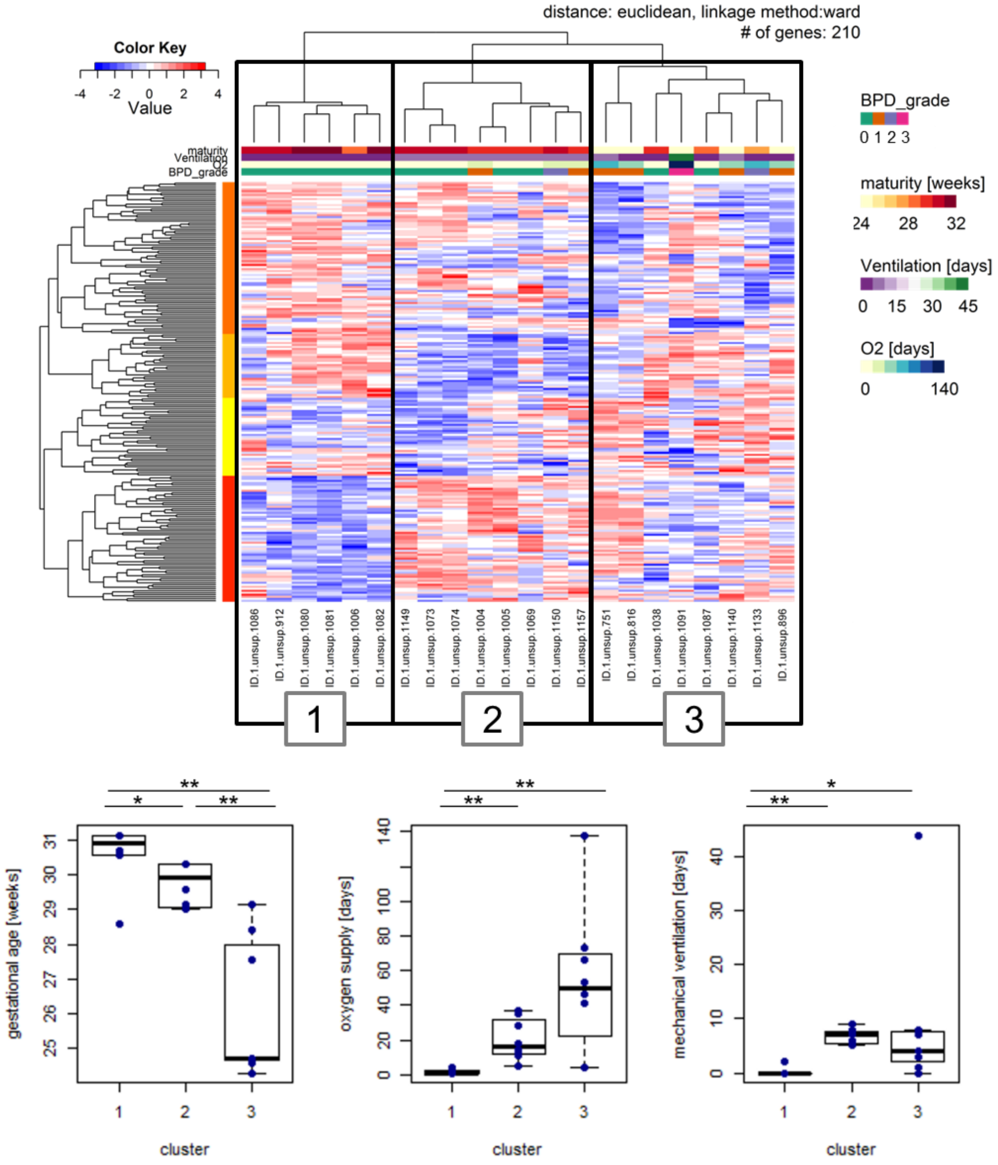

Figure 4-11 Heatmap of genes correlated with the duration of assisted ventilation and/or oxygen support, stratified for maturity (upper plot) and plot of gestational age, oxygen supply, and mechanical ventilation for the clusters obtained by regression analysis. Asterisks indicate the significance levels of pairwise Wilcoxon test with FDR correction: *** FDR < 0.001, ** FDR < 0.01, * FDR < 0.05.

All transcripts are separated in sets of genes correlated with the duration of ventilation or duration of O2 or the interaction of both and submitted to IPA for analysis of regulator effector networks. Transcripts correlating with ventilation indicate a decreased function in the morphology of leukocytes, decreased hypoplasia of thymus glands, and decreased lack of T lymphocytes, but increased quantity of thymocytes with prolonged ventilation (Table 4-7). For transcripts correlating only with the duration of O_2 no increased or decreased functions can be found. For transcripts correlating with the interaction of ventilation and oxygen an increased proliferation of double-negative T lymphocytes can be found.

Table 4-7 IPA de- or increased diseases and biological functions for sets of transcripts correlated with either ventilation, oxygen support or the interaction of both treatments.

By regression coefficients			Z-scores	
Diseases and bio functions		**ventilation**	**oxygen support**	**ventilation: oxygen**
morphology of lymphocytes	*	-2.22	0.51	1.19
morphology of leukocytes	*	-2.22	0.51	1.19
hypoplasia of thymus gland	*	-2.22	0.49	1.48
morphology of T lymphocytes	** **	-1.98	0.06	0.83
lack of T lymphocytes	*	-1.98	0.06	0.83
morphology of blood cells	*	-1.78	0.81	0.81
proliferation of leukemia cell lines	*	1.94	0.28	-1.11
quantity of thymocytes	**	1.58	0.25	-1.19
quantity of double-negative T lymphocyte	*	-0.93	1.19	1.99

All activated or inhibited biological functions showed a p-value < 0.05 based on test for over-representation.

For the transcripts associated with the duration of ventilation, the prediction revealed an activation of IL3, CD40LG, and CSF2 as putative cytokine upstream regulators. For transcripts correlated with O_2, a deactivation of IL-1B and IL-1 can be predicted. Their role is also indicated by the analysis of transcripts showing an effect in correlation to both, ventilator support and O_2, as an activation of IL-1 and deactivation of IL-10 is predicted here (Table 4-8).

Table 4-8 Overrepresentation of IPA upstream regulators for sets of transcripts correlated with either ventilation, oxygen support or the interaction of both. The direction of activation is determined by the direction of the regression coefficients.

	Z-scores		
Upstream regulators	**ventilation**	**oxygen support**	**Interaction of ventilation and oxygen**
IL3	1.95 *	-1.60 *	0.31 *
CD40LG	1.93		
CSF2	1.93		
IL7	1.30 *	-1.49 *	-0.37 *
IL1B		-1.98	
IL1		-1.94	1.94
Interferon alpha	0.06 **	-1.35 **	0.65 **
IL10	*	*	-1.83 *
IFNA2	0.51	-0.11	0.11

All activated or inhibited upstream regulators showed a p-value < 0.05 based on test for over-representation.

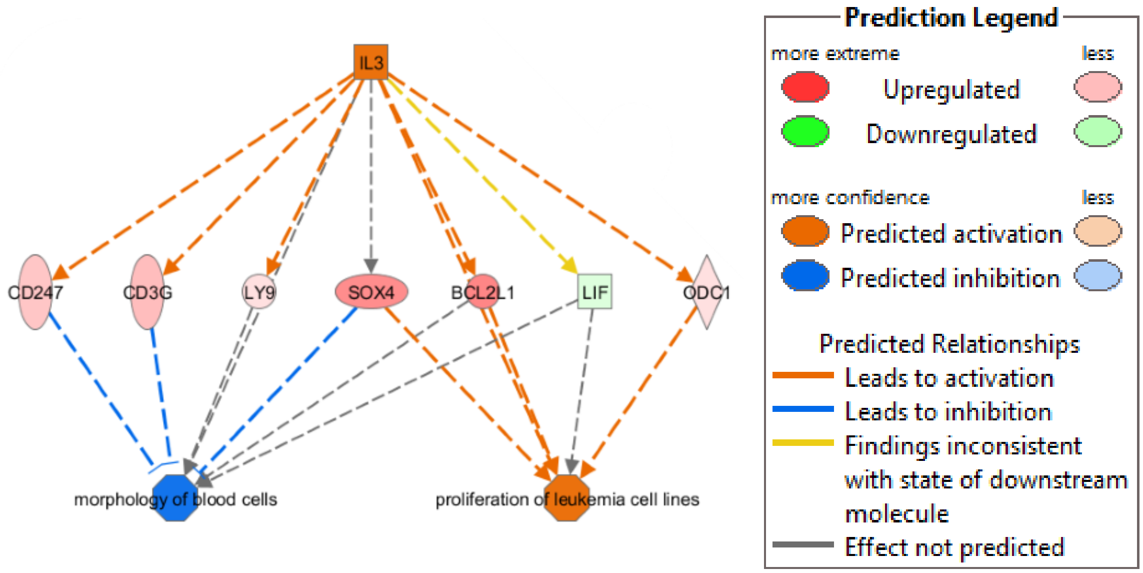

Figure 4-12 Regulator effects network for genes correlated with the need of assisted ventilation. To be able to predict the direction of regulation in upstream regulators or downstream biological processes regression coefficients were used. Upregulated translates then in positively correlated, down regulated in negatively correlated.

For genes correlating with ventilation, a regulator effects network could be identified showing how IL3 and blood cell function are connected (Figure 4-12). Here, activation of IL3 as its regulator can be predicted. The increase in expression of BCL2L1, CD247, CD3G, LY9, ODC1, SOX4 and the decrease in LIF expression with increasing need for assisted ventilation indicates a decrease of biological functions involved in the morphology of blood cells and an increase in proliferation of leukemia cell lines.

4.3.5 Cytokine upstream regulators in microarray analyses

The results indicate that cytokines are overrepresented as putative upstream regulators; they are significantly overrepresented as regulators for genes differentially expressed between the different groups of BPD preterm infants, and in transcripts correlated with the duration of assisted ventilation or O_2 in consideration of the maturity. In the various steps of the microarray analysis a total of 46 cytokines as putative activated or deactivated upstream regulators are identified (Figure 4-13).

A common cytokine in the unsupervised approach, the regression analysis, and the supervised approach using LIMMA is the IL-10, which is predicted to be down-regulated (activation z-score= -1.83) through the interaction effect of MV and O_2, while predicted to be activated in mild BPD compared to no BPD (activation z-score= 1.21). In both the unsupervised approach and the predictive supervised approach using PAM IL-3 and the group of interferon alpha cytokines are predicted to be involved as upstream regulators for the observed gene expression. The activation of ventilation is associated with MV, while O_2 is associated with the deactivation of IL-3. For interferon alpha no indication concerning the direction of activation can be made.

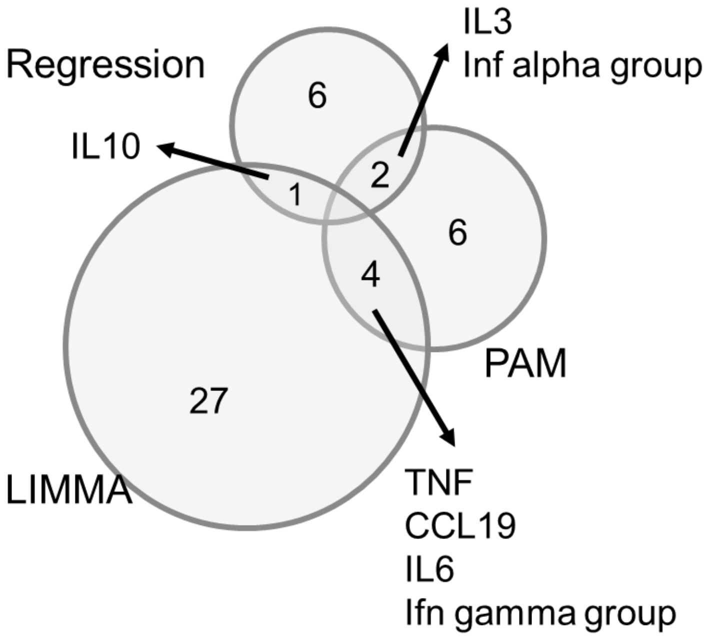

Figure 4-13 **Common cytokine upstream regulators of the gene set of 238 differentially regulated genes (LIMMA), 71 genes suitable for predicting BPD groups (PAM), and of the 210 genes correlated with duration of ventilation and/or oxygen support, corrected for maturity of preterm infants (Regression).**

4.4 Discussion

4.4.1 Discussion of results

The developed workflow for a standardization transcriptome analysis (chapter 2 and 3) has been successfully applied in a proof of concept presented in the current chapter 4 dealing with the development of BPD in preterm infants. The presented tool set thereby proved to be flexible and robust enough to handle the flaws in real world data sets and deliver robust results in accordance with the research question at hand. The results of the proof of concept study will be interpreted in the following discussion, and shall highlight the full potential of a possible clinical study to be conducted based on the presented work.

4.4.1.1 Supervised microarray analysis – differential gene expression analysis

The presented analysis shows that differences in gene expression patterns at the time of birth of preterm infants lead to the development of different BPD grades (Figure 4-2, Figure 4-1, Figure 4-3). Interestingly, preterm infants with mild BPD exhibit a different expression pattern as preterm infants without BPD and preterm infants with moderate/severe BPD. This may indicate that the development of mild BPD follows different routes than the development of more severe BPD. One would expect that preterm infants with any grade of BPD have similar expression patterns at time of birth, if these patterns determine the development of BPD at such an early stadium as birth. But only four transcripts are differentially expressed at birth in both BPD groups compared to the group of no BPD infants. Nevertheless, it is possible to distinguish infants with moderate or severe BPD from preterm infants without BPD at birth. This indicates that certain processes at birth predispose infants for the development of moderate or severe forms of BPD. Three potential underlying causes can be discussed here:

1. Genetic predisposition for higher grade BPD,
2. Immaturity or other clinical factors of infants hindering preterm infants to cope with infection,
3. Prenatal inflammatory processes.

Genetic predisposition. Twin studies indicate that the development of BPD in one twin serves as a risk factor for the other twin. In a careful review of twin studies Shaw and O’Brodovich (2013) come to the conclusion that at least for moderate/severe BPD a genetic factor may be suspected. For mild BPD no evidence for heritability could be found (Shaw and O’Brodovich, 2013). Lavoie et al. (2008) examined clinical data of monozygotic and dizygotic twins born before 30 weeks GA for genetic and environmental susceptibility factors towards BPD. Their findings suggest that the need of oxygen at 36 weeks postmenstrual age is genetically influenced. For mild BPD they found little evidence for a genetic influence, but mild BPD could be attributed to environmental effects, such as gestational age. Bhandari et al. (2006) as well conducted a twin study on preterm infants born with ≤ 32 weeks GA. They also found a genetic association with the need for oxygen supplementation at 36 weeks postmenstrual age. However, in our findings a higher similarity between preterm infants without BPD and with moderate or severe BPD cannot be explained by the findings of the twin studies examined by Shaw and O’Brodovich (2013), Lavoie et al. (2008), or Bhandari et al. (2006). A genome-wide-association study by Wang et al. (2013) could not find an genetic factor that makes preterm infants susceptible to the development of BPD. Wang et al. (2013) however, examined a putatively different ethnicity. While the twin studies of Shaw and O’Brodovich (2013) probably examined Caucasian patients, one from Canada (Lavoie et al., 2008) and one from central USA (Bhandari et al., 2006), the examined study population of Wang et al. (2013) is of Mexican-Hispanic origin.

Clinical data. In comparison with the clinical data of the preterm infants in this study (Table 4-2), we noticed that GA, congenital sepsis, and the duration of CPAP follow a similar pattern as the microarray cluster analysis (Figure 4-2). Here differences could be found between neonates with mild BPD and without BPD, but not in comparison of either group to moderate/severe BPD. Infants with mild BPD were born younger (Median GA: 24.7 weeks in mild BPD, 30.3 weeks in no BPD, 27.6 days in cases with moderate or severe BPD), received longer CPAP treatment (Median: 21 days in mild BPD, 2 days in no BPD, 6 days in moderate or severe BPD infants) and developed more often than expected a congenital sepsis. In our study population, all neonates who later developed BPD had congenital sepsis, but only half of the no BPD infants. Immaturity of preterm infants is a well-known risk factor for the development of BPD and it is difficult to eliminate its influence entirely from the analysis. It also becomes apparent that the incidence of BPD and congenital sepsis cannot be clearly separated in our study group. Both diseases are strongly associated with the birth weight. This explains why it was not possible to match BPD groups by the factors gestational age despite our best efforts. Nevertheless the observed unique expression pattern for preterm infants developing mild BPD can be seen as a starting point for future research.

Inflammatory processes. Findings based on DAVID functional annotation clustering indicate that differentially expressed transcripts in preterm infants with mild BPD compared to moderate/severe BPD are overrepresented in ontology terms associated with *cytoskeleton* and *calcium homeostasis*. Transcript differentially expressed in mild BPD compared to no BPD are putatively overrepresented in the gene ontology terms *transmembrane transport* and *leukocyte activation*. Moderate/severe BPD infants express transcripts differentially that are putatively overrepresented in induction of *apoptosis* and *transcription*. The differences in the profiles of moderate/severe BPD infants compared to both other groups of infants may be lead back to transcription factor activity. The differences compared to only mild BPD indicate differences in *cell migration*, *oxidation reduction*, and the *mitochondrion*.

As the overrepresentation analysis by DAVID gives no clues to the direction of the processes, an additional IPA analysis was conducted to predict the increased or decreased activation of processes, as well as finding upstream regulators responsible for the observed gene expression patterns. In summary, it can be seen that the regulators TNF, IL6, TCR, and MAPK14 play an important role in processes at birth of children developing mild BPD and are frequently activated for those infants. Processes to be activated are involved in chemotaxis; processes predicted to be deactivated are involved in the accumulation of eosinophils and cell death and fibrosis. In preterm infants with moderate or severe BPD is the chemotaxis of cells predicted to be deactivated in comparison with mild BPD infants (see Table 4-3).

It becomes apparent that in preterm infants developing mild BPD cytokines indeed play an important role (see Figure 4-4, Figure 4-6, Figure 4-7, Figure 4-8); the upstream regulators TNF-α and IL-6 are linked to an activation of airway hyperresponsiveness, chemotaxis, and an increased proliferation of neutrophils and granulocytes. The observed regulator effect networks indicate the presence of inflammatory processes in preterm infants at birth. Melville and Moss (2013) review the effect of MV on preterm infants and find that sustained systemic inflammation may be a risk factor for chronic lung diseases. This is also based on findings from Köksal et al. (2012), who find in tracheal aspirate and serum from the umbilical cord of preterm infants increased levels of TNF-α, IL-1β, IL-6, but decreased levels of anti-inflammatory IL-10. They find no differences in cytokine levels by

BPD severity like in our study. The expression levels in our study also indicate activation of TNF-α and IL-6 of for preterm infants with mild BPD (see Table 4-4), although not for higher severities, which may be due to the rather low sample size. Köksal et al. (2012) also find differences in GA in the observed groups, which makes the study population comparable to the results in our study.

4.4.1.2 **Supervised microarray analysis – predictive microarray analysis**

The IPA comparison analysis for transcripts identified by PAM s further empathizes the importance of cytokines, here especially TNF-α, and the importance of inflammatory processes (Figure 4-10). For these genes the process *production of ROS*, which is putatively deactivated in preterm infants with mild BPD, and the processes *chemotaxis of cells* and *inflammatory response*, which are predicted to be activated in both BPD groups, either mild or moderate/severe BPD, stand out as they represent biological processes which may be able to differentiate between no BPD, mild BPD and moderate/severe BPD.

Similar mechanism were also detected by DAVID functional clustering for the set of transcripts obtained by LIMMA, which identified the mitochondrion or the gene ontology term of *oxidation reduction* as well as *cell migration* as overrepresented in transcripts differentially regulated in the mild BPD group and *induction of apoptosis* in the moderate/severe BPD group, each compared to no BPD.

The synthesis of ROS or free radicals can have beneficial effects when it is used to fight infection as a first line defense system, but together with an immature antioxidant defense system and oxygen treatment of the preterm infants it can lead to severe oxidant stress and cause lung damage (Perrone et al., 2012; Speer, 2006). ROS can also serve as second messengers to transcription factor activation and induce apoptosis, further radical formation, and inflammation; it amplifies the inflammatory response (Auten and Davis, 2009).

Our finding show that the transcripts able to predict BPD severity grades lead to an activation of inflammatory response, and chemotaxis processes in preterm infants with BPD. However, in preterm infants with mild BPD a deactivated synthesis of ROS is predicted (Figure 4-10, Figure 4-14). So it can be assumed that the deactivation of ROS synthesis has a beneficial effect on preterm infants dealing with inflammation. This may be due to the reduced oxidative stress and reduced inflammatory response.

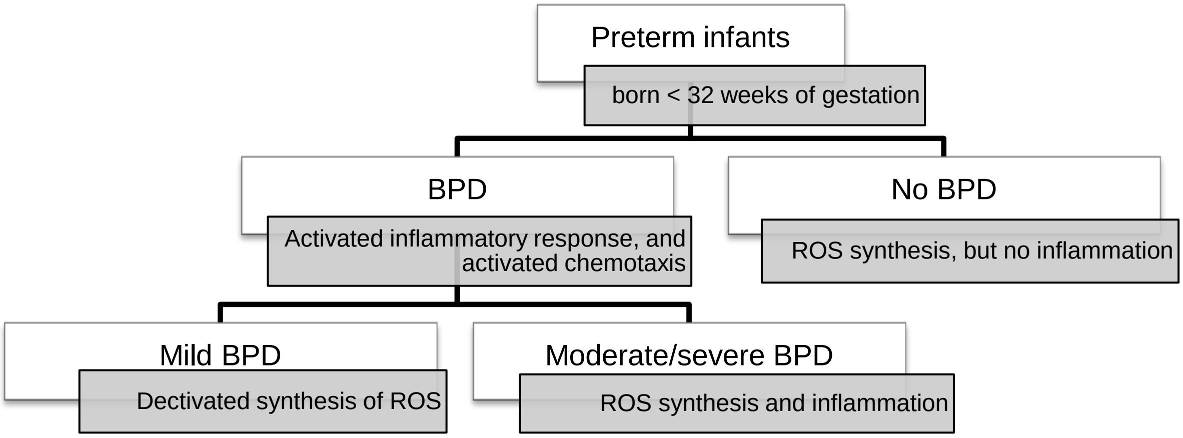
Preterm infants
born < 32 weeks of gestation
BPD
Activated inflammatory response, and activated chemotaxis
No BPD
ROS synthesis, but no inflammation
Mild BPD
Dectivated synthesis of ROS
Moderate/severe BPD
ROS synthesis and inflammation

Figure 4-14 In BPD are processes in inflammatory response and chemotaxis activated, but the deactivation of ROS synthesis leads to an only mild form of BPD.

4.4.1.3 Semi-supervised microarray analysis – linear regression models

To assure that only the risk factors, which are able to discriminate between BPD severity grades, are taken into account we fitted multiple linear regression models with duration of ventilation, duration of O_2, and GA to explain the observed gene expression patterns (Figure 4-2). A total of 210 transcripts were either positive or negative regression coefficients for ventilation and/or oxygen or the interaction between those two parameters were selected. Despite the fact that the effect of GA cannot be eliminated, the differences between the clusters of preterm infants become less pronounced than in the initial analysis. Of the 210 transcripts 55 showed regression coefficients also significant in GA at a 90%-significance level. Despite differences in GA, clusters obtained by this set of transcripts also show different levels of O_2 and MV, which serves as proof that this analysis indeed, selects groups of preterm infants with greater variation in these parameters.

Again, it can be observed that cytokines and inflammatory processes at birth are connected with the development of BPD. The correlation structure with the duration of assisted ventilation and O_2 indicates that processes, concerning for example the morphology of blood cells, influence the time a preterm infants needs to be ventilated thereby influencing the risk for the development of BPD. Transcripts which correlate with the interaction of O_2 and MV and are regulated by IL-1 indicating an activation of these cytokines, while IL-10 is deactivated (Table 4-8). These findings are also supported by findings of Köksal et al. (2012), who report an increase in IL-1β and an decrease in IL-10 in the tracheal secret and cord blood of preterm infants developing BPD. It also shows that the gene expression of some transcripts leads to a prolonged treatment with O_2 and MV due to the inflammatory response they trigger.

4.4.2 Limitations of the study

It is important to notice that despite applying appropriate statistical methods, the sample size of this analysis remains small. It is thus difficult to detect significant differences between the BPD groups and generate hypotheses based on these findings. Especially preterm infants developing moderate or severe BPD are sparse. Owing to this our finding can and should not be generalized to deduce new principles valid beyond the scope of this study. While they serve as a profound basis to initiate further research and hint towards possible systematic in the development of BDP in preterm infants one needs to be aware of this limitation when evaluating the results.

In the regression model based analysis a rather simplistic approach was used to connect clinical data from time of birth and later stages in the life of preterm infants. This analysis is based on the assumption that a linear relationship between clinical data and gene expression exists as well as on the assumption that residuals are normally distributed with and expected mean of zero. This simplistic approach was necessary because we wanted to separate effects associated with the continuous factors GA, MV, and O_2. The same holds true for LIMMA models which are also able to consider quantitative data. LIMMA would be an interesting alternative to the here presented approach, as it saves degrees of freedoms and thus increases the probability to detect associated transcripts. However, the selection of transcripts associated with certain factors and not with other factors remains difficult.

Another alternative multivariate approach would have been the partial least squares based prediction analysis. It uses gene expression data to predict quantitative clinical parameters. But to accomplish this, first a factor analysis to reduce the multivariate problem to a univariate problem has to be conducted. Then this factor is predicted by gene expression. Factor analysis bears the advantage of eliminating multicollinearity between the factors, which can bias the estimation of regression coefficients. But on the other hand factor analysis makes it impossible to separate the influence of GA from the other factors. Due to these circumstances together with the very low sample size it is not possible to create a complete picture. This is a first step to understand the need for MV and O_2 determined at birth, but certainly needs further investigations to create more suitable models.

4.5 Conclusion

This study demonstrates that the development of mild BPD follows different routes than the development of moderate/severe BPD in preterm infants. Transcriptome analysis indicates a high similarity between the preterm infants with no BPD and the preterm infants with moderate/severe BPD investigated in the study. Preterm infants with mild BPD seem to follow a different route in the development of BPD. The following possible drivers for BPD have been discussed: 1) genetic predisposition, 2) maturity or other clinical factors of the preterm infants, 3) or inflammation prior to birth.

The effect of genetic factors and the role they play in the development of mild BPD was not analyzed in this study, even though the performed hierarchical clustering hints toward a possible genetic component influencing the transcriptome. For future genetic studies, it may be interesting to separate preterm infants with different grades of BPD and set a focus on the development of mild BPD as it has not been covert in recent literature.

The clinical data of our patient cohort revealed that GA, duration of CPAP, and congenital sepsis follow the same pattern as obtained by the supervised microarray analysis approach. Maturity in our study is higher in preterm infants with no BPD and in preterm infants with moderate/severe BPD, while CPAP treatment is shorter. From the supervised analysis approach we learn that in these infants transcripts are differentially expressed than in mild BPD infants that are not only associated with oxidation reduction, but also lead to a predicted activation of ROS synthesis. In addition with transcripts leading to an increased inflammatory response and chemotaxis of cells in preterm infants with either mild or moderate/severe BPD it may be possible to diagnose BPD at the birth of preterm infants with the help of specialized microarrays.

These findings point at a connection between the maturity of the preterm infants, which is associated with the ability of ROS synthesis, and an increased inflammatory response in the development of different BPD severity grades. In short, preterm infants with moderate/severe BPD show a much stronger inflammatory response than preterm infants with mild BPD. The importance of oxidative stress (Auten and Davis, 2009; Perrone et al., 2012; Saugstad, 2010) and inflammation (Melville and Moss, 2013; Speer, 2006, 2003) for the development of BPD has been discussed in various publications, but up to now no publication has successfully separated the different severity grades of BPD based on microarray data.

When only the duration of MV or oxygen support is used to select transcripts, it can be seen that a correlation between a gene expression of transcripts that predict an increased development of T cells and the duration of MV and oxygen support exists.

In accordance with Jobe (2006) the BPD severity grades used in this thesis were assigned to the preterm infants based on the required time of O_2 and MV treatment. Gene expression patterns and pathways leading to a prolonged need for MV and O_2 under consideration of the GA were investigated. This analysis allowed us to filter genes that are associated with MV and/or O_2, or the interaction of both factors, but showed at most an additional association with GA. It became apparent that prolonged ventilation is correlated with gene expression at birth leading to an increase in T-cell development. In the studied cohort it can be seen that the cornerstone for MV and prolonged O_2 is laid at birth, possibly through inflammatory processes and oxidative stress starting at the time of birth.

Not only the downstream effects hint at inflammation at birth, but also the analysis of upstream regulators shows that especially cytokines are involved in the gene expression observed mediating the inflammatory response. A total of 46 cytokines were identified to be overrepresented as regulators of the transcripts identified by the different microarray analysis approaches.

The findings of the microarray analysis have yet to be validated by a second much larger cohort. It also would be interesting to see whether at later stages cluster formation as seen at birth persists. With the development of BPD the patterns of infants with mild BPD and infants with higher grades of BPD must become more similar and distinguish more clearly from preterm infants without BPD.

5 Summary

Bronchopulmonary dysplasia is one of the most common chronic lung diseases and contributes greatly to morbidity of preterm infants. While moderate and severe forms of BPD are the most common forms under investigation little is known about the development of mild BPD. The aim of this work is to identify mechanisms and biomarkers, which make it possible to predict at birth whether a preterm infant is prone to develop no BPD, mild BPD, or a stronger form of BPD.

Transcriptome and in particular microarray analysis plays an important role in the generation of hypotheses regarding underlying mechanisms and diagnostic tools. Microarrays are able to examine a multitude of transcripts simultaneously. In order to obtain reliable results, however, a number of data preparation steps are necessary. The statistical analysis has some peculiarities due to the high number of parameters collected and a comparatively small number of patients. In the present study, a standardized workflow for the statistical analysis of transcriptome data is developed and used to predict BPD in very preterm infants.

First, background correction and normalization steps are performed to prepare the data. This on the one hand, separates signal from noise in the gene expression, and on the other hand makes the microarrays comparable. Then informative transcripts are iteratively selected. Transcripts are reviewed for missing values, low expression levels, and extreme values and if necessary eliminated. Then remaining missing values are estimated using an imputation algorithm.

Data preparation was particularly facilitated through the implementation and automation of workflow using the programming language R. In comparison to a preparation that is based on different independent programs and tools a considerable advantage in terms of data amount that can be processed, processing time, and actuality of the algorithms can be achieved. Existing programs have been replaced by Bioconductor packages where possible to avoid data transmission errors.

The instruments for data preparation can be used for the analysis of either predefined groups (supervised) as well as without predetermined groups (un-/ semi-supervised). This way it is possible to take the nature and prerequisites of the different statistical analyses into account. The group-based (supervised) data analysis is used to work out differences between the examined groups. For the presented study two methods (Limma, PAM) were used to identify differentially regulated genes. While Limma determined individual transcripts that are differentially regulated in isolation from other transcripts, the focus of PAM is on the interplay of the transcripts to explain the different expressions of the phenotypes.

The aim of the transcriptome analysis without prior definition of groups (unsupervised) is to identify groups solely based on gene expression. Since in this case a very large number of transcripts will be taken into account, this approach is only suitable to draw conclusions about underlying diseases affecting the whole gene expression. Therefore in a semi-supervised approach the data preparation is performed without groups. However, only a selection of transcripts is used. The selection is based on clinical data associated with the phenotype. With this selection clustering techniques are then used to identify groups.

In the present case different maturities of preterm infants at time of birth caused particular difficulties while forecasting BPD groups. Frequently the gene expression patterns differ with maturity. To address this issue in particular the gestational age of preterm infants is used as a secondary variable in the selection of transcripts. In addition it is beneficiary to have only transcripts selected that show an effect in mechanical ventilation and oxygen requirement but not in GA or in addition to the effect of GA. As this cannot be achieved with the usual methods of gene selection (Limma, PLS), a multiple linear regression is performed here, which allows filtering only transcripts with additional effects.

The gene expression analysis of the present study comprising neonates born before 32 weeks of gestation shows that consideration of processes at birth significantly augments the understanding of BPD in general and its classification in different severity grades. With the help of the presented gene expression analysis tools for data preparation, data analysis and functional gene expression analysis, it is possible to predict BPD severity grades at birth and identify cytokines as biomarkers.

Our results showed that the combination of oxidative stress and inflammation at birth contributes to the severity of BPD. In light of the duration of mechanical ventilation and the duration of oxygen supply considered, it becomes evident that processes responsible for the T-cell development are associated with the development of BPD. Furthermore, the importance of tumor necrosis factor α (TNFα), interleukin 6 (IL6), interleukin 1 and interleukin 10 in the regulation of the differential gene expression in BPD becomes apparent.

6 Zusammenfassung

Bronchopulmonare Dysplasie ist eine der am meisten verbreiteten chronischen Lungenerkrankungen und trägt stark zur Morbidität von Frühgeborenen bei. Während moderate und starke Formen von BPD bevorzugt untersucht werden, ist über die milde Form von BPD nur wenig bekannt. Ziel dieser Arbeit ist es, Hinweise auf Mechanismen und Biomarker zu identifizieren, die es möglich machen bei Geburt die Entwicklung keiner BPD, einer milden BPD, oder einer stärker ausgeprägten Form von BPD vorherzusagen.

Transkriptomanalysen und insbesondere Microarray-Analysen spielen eine wichtige Rolle in der Generation von Hypothesen in Bezug auf zugrundeliegende Mechanismen und diagnostischen Hilfsmitteln. Microarrays sind in der Lage eine Vielzahl von Transkripten gleichzeitig zu untersuchen. Um jedoch belastbare Ergebnisse zu bekommen, ist eine Reihe von Datenvorbereitungsschritten notwendig. Auch die statistische Analyse birgt einige Besonderheiten aufgrund der hohen Anzahl an erhobenen Parametern bei vergleichsweise geringer Anzahl an Patienten. In der vorliegenden Arbeit wurde ein standardisierter Ablaufplan zur statistischen Analyse von Transkriptom-Daten entwickelt und zu BPD-Prognose von Frühgeborenen verwendet.

Zunächst werden die mithilfe von Microarrays gewonnen Transkriptomdaten mit den üblichen Schritten der Hintergrundkorrektur und Normalisierung aufbereitet. Dies dient zum einen dazu, die Genexpression, die durch die zu untersuchende Krankheit hervorgerufen wurde, von dem Hintergrundsignal zu trennen und zum anderen dazu, die Microarrays vergleichbar zu machen. Anschließend werden informative Transkripte iterativ ausgewählt. In diesem Abschnitt der Datenaufbereitung werden Transkripte auf fehlende Werte, niedrige Expression und Extremwerte überprüft und gegebenenfalls eliminiert. Verbleibende fehlende Werte werden mithilfe eines Imputationsverfahrens geschätzt.

Eine besondere Erleichterung der Datenvorbereitung konnte durch die Implementierung und Automatisierung des Arbeitsablaufes in der Programmiersprache R erreicht erzielt werden. Im Vergleich zu einer Vorbereitung, die auf verschiedenen unabhängigen Programmen basiert, kann ein erheblicher Vorteil in Bezug auf Datenumfang, Bearbeitungszeit und Aktualität der Algorithmen erreicht werden. Soweit möglich wurden bestehende Programme durch Bioconductor-Pakete ersetzt, die es ermöglichen Übertragungsfehler zu vermeiden.

Diese Instrumente der Datenaufbereitung können sowohl bei der Analyse von vorgegebenen Gruppen (supervised) und ohne vorgegebene Gruppen (un-/semisupervised) eingesetzt werden. Auf diese Weise wird bereits bei der Vorbereitung der Daten berücksichtigt, welche Art der statistischen Analyse durchgeführt werden wird.

Die gruppenbasierte (supervised) Datenauswertung dient dazu, Unterschiede zwischen den zu untersuchenden Gruppen herauszuarbeiten. Für die vorgestellte Studie wurden zwei Methoden (Limma, PAM) verwendet, um differentiell regulierte Gene zu identifizieren. Während Limma einzelne Transkripte ermittelt, die losgelöst von anderen Transkripten differentiell reguliert sind, liegt der Fokus von PAM auf dem Zusammenspiel der Transkripte, welches die unterschiedliche Ausprägung des Phänotyps erklären.

Ziel der Transkriptom-Analyse ohne vorherige Festlegung von Gruppen (unsupervised) ist es, rein aufgrund der Genexpression Gruppen zu identifizieren. Da in diesem Fall eine sehr große Anzahl von Transkripten berücksichtig wird, ist dieser Ansatz nur bedingt geeignet, um Rückschlüsse auf zugrundeliegende Krankheiten zu ziehen. Deshalb wird in einem semi-supervised Ansatz zwar die Datenvorbereitung ohne Gruppen durchgeführt, jedoch wird eine Auswahl an Transkripten anhand klinischer Daten getroffen, die im Zusammenhang mit dem zu untersuchenden Phänotyp stehen. Aufgrund dieser Auswahl werden dann mittels Clustering Gruppen identifiziert. Eine besondere Schwierigkeit in der Prognose von BPD-Gruppen stellt im vorliegen Fall die Berücksichtigung der Reife der Frühgeborenen dar. Häufig ist die Genexpression zum Zeitpunkt der Geburt beeinflusst durch die Reife der Frühgeborenen; deshalb sollten nur Transkripte ausgewählt werden, die in Bezug auf mechanische Ventilation und Beatmung einen zusätzlichen Effekt zeigen. Mit den bisher üblichen Methoden der Genselektion (Limma, PLS) kann dies jedoch nicht berücksichtigt werden, weshalb hier eine multiple lineare Regression durchgeführt wird, die es erlaubt nur Transkripte mit zusätzlichen Effekten zu filtern.

Die Studie der Genexpression von Neugeborenen, geboren vor der 32. Schwangerschaftswoche, zeigt, dass eine Betrachtung der Prozesse zum Zeitpunkt der Geburt deutlich zum Verständnis von BPD im Allgemeinen und der Ausprägung verschiedener Schweregrade im Speziellen beitragen kann. So ist es möglich, anhand der vorgestellten Instrumente und mit Instrumenten der funktionellen Expressionsanalyse, biologische Prozesse und Zytokine identifizieren, die dazu dienen den Schweregrad einer BPD schon bei Geburt abzuschätzen.

In der vorliegenden Studie ist zu sehen, dass bereits bei Geburt, die Kombination aus oxidativem Stress und Inflammation zur Ausprägung des BPD-Schweregrades beitragen. In der Betrachtung der Dauer der mechanischen Ventilation im Zusammenspiel mit der Dauer der Sauerstoffgabe wird deutlich, dass Prozesse der T-Zell-Entwicklung an der Entwicklung von BPD beteiligt sind. Die Betrachtung der Zytokine, die die beobachten Gen-Expression regulieren, wird die Bedeutung des Tumornekrosefaktors α (TNF-α), Interleukin 6 (IL-6), Interleukin 1 β, und Interleukin 10 für das Auftreten von BPD deutlich. Die Proteinanalyse bestätigt die Relevanz von TNF-α und IL-6 zur Differenzierung der BPD-Grade bei Geburt.

7 Literature

Aittokallio, T., 2010. Dealing with missing values in large-scale studies: microarray data imputation and beyond. Brief Bioinform 11, 253–264. doi:10.1093/bib/bbp059

Applied Microarrays, 2013. Data Analysis [WWW Document]. URL http://www.appliedmicroarrays.com/index.php?option=com_content&view=article&id=11&Itemid=17 (accessed 10.23.13).

Applied Microarrays, 2007. lod.pdf [WWW Document]. Page | 1 Improved Methodology for Assessing the Lower Limit of Detection. URL http://new.appliedmicroarrays.com/images/stories/articles/bg/lod.pdf (accessed 4.4.14).

Auten, R.L., Davis, J.M., 2009. Oxygen toxicity and reactive oxygen species: the devil is in the details. Pediatric research 66, 121–127. doi:10.1203/PDR.0b013e3181a9eafb

Bair, E., Tibshirani, R., 2004. Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. PLoS Biology 2, e108. doi:10.1371/journal.pbio.0020108

Baldi, P., Long, A.D., 2001. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics 17, 509–519. doi:10.1093/bioinformatics/17.6.509

Ballabh, P., Simm, M., Kumari, J., Krauss, A.N., Jain, A., Auld, P.A.M., Cunningham-Rundles, S., 2003. Lymphocyte subpopulations in bronchopulmonary dysplasia. Am J Perinatol 20, 465–475. doi:10.1055/s-2003-45387

Benjamini, Y., 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, 289–300. doi:10.2307/2346101

Bhandari, V., Bizzarro, M.J., Shetty, A., Zhong, X., Page, G.P., Zhang, H., Ment, L.R., Gruen, J.R., 2006. Familial and Genetic Susceptibility to Major Neonatal Morbidities in Preterm Twins. PEDIATRICS 117, 1901–1906. doi:10.1542/peds.2005-1414

Bhattacharya, S., Go, D., Krenitsky, D.L., Huyck, H.L., Solleti, S.K., Lunger, V.A., Metlay, L., Srisuma, S., Wert, S.E., Mariani, T.J., Pryhuber, G.S., 2012. Genome-Wide Transcriptional Profiling Reveals Connective Tissue Mast Cell Accumulation in Bronchopulmonary Dysplasia. Am J Respir Crit Care Med 186, 349–358. doi:10.1164/rccm.201203-0406OC

Blencowe, H., Cousens, S., Chou, D., Oestergaard, M., Say, L., Moller, A.-B., Kinney, M., Lawn, J., the Born Too Soon Preterm Birth Action Group (see acknowledgement for full list), 2013. Born Too Soon: The global epidemiology of 15 million preterm births. Reproductive Health 10, S2. doi:10.1186/1742-4755-10-S1-S2

Bolstad, B.M., Irizarry, R.A., Åstrand, M., Speed, T.P., 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19, 185–193. doi:10.1093/bioinformatics/19.2.185

Bolstad, B.M., Irizarry, R.A., Gautier, L., Wu, Z., 2005. Preprocessing High-density Oligonucleotide Arrays, in: Gentleman, R., Carey, V.J., Huber, W., Irizarry, R.A., Dudoit, S. (Eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Statistics for Biology and Health. Springer New York, pp. 13–32.

Bose, C.L., Laughon, M.M., Allred, E.N., O'Shea, T.M., Van Marter, L.J., Ehrenkranz, R.A., Fichorova, R.N., Leviton, A., 2013. Systemic inflammation associated with mechanical ventilation among extremely preterm infants. Cytokine 61, 315–322. doi:10.1016/j.cyto.2012.10.014

Bø, T.H., Dysvik, B., Jonassen, I., 2004. LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucl. Acids Res. 32, e34–e34. doi:10.1093/nar/gnh026

Boulesteix, A.-L., Lambert-Lacroix, S., Peyre, J., Strimmer, K., 2012. plsgenomics: PLS analyses for genomics.

Boulesteix, A.-L., Porzelius, C., Daumer, M., 2008. Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. Bioinformatics 24, 1698–1706. doi:10.1093/bioinformatics/btn262

Boulesteix, A.-L., Strimmer, K., 2007. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinform 8, 32–44. doi:10.1093/bib/bbl016

Boutros, P.C., Okey, A.B., 2005. Unsupervised pattern recognition: An introduction to the whys and wherefores of clustering microarray data. Brief Bioinform 6, 331–343. doi:10.1093/bib/6.4.331

Breitling, R., Armengaud, P., Amtmann, A., Herzyk, P., 2004. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Letters 573, 83–92. doi:10.1016/j.febslet.2004.07.055

Brevern, A.G. de, Hazout, S., Malpertuy, A., 2004. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. BMC Bioinformatics 5, 114. doi:10.1186/1471-2105-5-114

Brock, G.N., Shaffer, J.R., Blakesley, R.E., Lotz, M.J., Tseng, G.C., 2008. Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. BMC Bioinformatics 9, 12. doi:10.1186/1471-2105-9-12

Celton, M., Malpertuy, A., Lelandais, G., Brevern, A.G. de, 2010. Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. BMC Genomics 11, 15. doi:10.1186/1471-2164-11-15

Chipman, H., Hastie, T.J., Tibshirani, R., 2003. Clustering microarray data, in: Speed, T. (Ed.), Statistical Analysis of Gene Expression Data. CRC Press, pp. 159–200.

Chu, G., Li, J., Narasimhan, B., Tibshirani, R., Tusher, V.G., 2014. SAM "Significance Analysis of Microarrays" Users guide and technical document [WWW Document]. URL http://statweb.stanford.edu/~tibs/SAM/sam.pdf (accessed 5.6.14).

Cleveland, W.S., Devlin, S.J., 1988. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. Journal of the American Statistical Association 83, 596–610. doi:10.1080/01621459.1988.10478639

Cohen, J., Marter, L.J.V., Sun, Y., Allred, E., Leviton, A., Kohane, I.S., 2007. Perturbation of gene expression of the chromatin remodeling pathway in premature newborns at risk for bronchopulmonary dysplasia. Genome Biology 8, R210. doi:10.1186/gb-2007-8-10-r210

Cousineau, D., Chartier, S., 2010. Outliers detection and treatment: a review. International Journal of Psychological Research 3, 58–67.

Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., Lempicki, R.A., 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 4, R60. doi:10.1186/gb-2003-4-9-r60

DeRisi, J.L., 1997. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. Science 278, 680–686. doi:10.1126/science.278.5338.680

D'haeseleer, P., 2005. How does gene expression clustering work? Nature biotechnology 23, 1499–1502. doi:10.1038/nbt1205-1499

Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J.C., Hernandez-Boussard, T., Rees, C.A., Cherry, J.M., Botstein, D., Brown, P.O., Alizadeh, A.A., 2003. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucl. Acids Res. 31, 219–223. doi:10.1093/nar/gkg014

Diez, D., 2013. Codelink_Introduction.pdf [WWW Document]. Introduction to the Codelink package. URL http://bioconductor.org/packages/devel/bioc/vignettes/codelink/inst/doc/Codelink_Introduction.pdf (accessed 10.27.13).

Diez, D., Alvarez, R., Dopazo, A., 2007. Codelink: an R package for analysis of GE healthcare gene expression bioarrays. Bioinformatics 23, 1168–1169. doi:10.1093/bioinformatics/btm072

Dudoit, S., Fridlyand, J., 2003. Classification in microarray experiments, in: Speed, T. (Ed.), Statistical Analysis of Gene Expression Data. CRC Press, pp. 93–158.

Eady, J.J., Wortley, G.M., Wormstone, Y.M., Hughes, J.C., Astley, S.B., Foxall, R.J., Doleman, J.F., Elliott, R.M., 2005. Variation in gene expression profiles of peripheral blood mononuclear cells from healthy volunteers. Physiol. Genomics 22, 402–411. doi:10.1152/physiolgenomics.00080.2005

Edwards, D., 2003. Non-linear normalization and background correction in one-channel cDNA microarray studies. Bioinformatics 19, 825–833. doi:10.1093/bioinformatics/btg083

Ernst, C., Bureau, A., Turecki, G., 2008. Application of microarray outlier detection methodology to psychiatric research. BMC Psychiatry 8, 29. doi:10.1186/1471-244X-8-29

Fox, R.J., Dimmic, M.W., 2006. A two-sample Bayesian t-test for microarray data. BMC Bioinformatics 7, 126. doi:10.1186/1471-2105-7-126

Freyhult, E., Landfors, M., Onskog, J., Hvidsten, T.R., Ryden, P., 2010. Challenges in microarray class discovery: a comprehensive examination of normalization, gene selection and clustering. BMC Bioinformatics 11, 503. doi:10.1186/1471-2105-11-503

Gevaert, O., Smet, F.D., Timmerman, D., Moreau, Y., Moor, B.D., 2006. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics 22, e184–e190. doi:10.1093/bioinformatics/btl230

Gruvberger-Saal, S.K., Edén, P., Ringnér, M., Baldetorp, B., Chebil, G., Borg, Å., Fernö, M., Peterson, C., Meltzer, P.S., 2004. Predicting continuous values of prognostic markers in breast cancer from microarray gene expression profiles. Mol Cancer Ther 3, 161–168.

Hastie, T., Tibshirani, R., Narasimhan, B., Chu, G., 2013a. impute: Imputation for microarray data [WWW Document]. URL http://www.bioconductor.org/packages/release/bioc/html/impute.html (accessed 7.11.13).

Hastie, T., Tibshirani, R., Narasimhan, B., Chu, G., 2013b. pamr: Pam: prediction analysis for microarrays [WWW Document]. URL http://CRAN.R-project.org/package=pamr (accessed 4.28.14).

Hatfield, G.W., Hung, S., Baldi, P., 2003. Differential analysis of DNA microarray gene expression data. Molecular Microbiology 47, 871–877. doi:10.1046/j.1365-2958.2003.03298.x

Hong, F., Breitling, R., 2008. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinformatics 24, 374–382. doi:10.1093/bioinformatics/btm620

Hong, F., Breitling, R., McEntee, C.W., Wittner, B.S., Nemhauser, J.L., Chory, J., 2006. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics 22, 2825–2827. doi:10.1093/bioinformatics/btl476

Hua, D., Lai, Y., 2007. An ensemble approach to microarray data-based gene prioritization after missing value imputation. Bioinformatics 23, 747–754. doi:10.1093/bioinformatics/btm010

Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4, 44–57. doi:10.1038/nprot.2008.211

Huber, W., Irizarry, R.A., Gentleman, R., 2005. Preprocessing Overview, in: Gentleman, R., Carey, V.J., Huber, W., Irizarry, R.A., Dudoit, S. (Eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Statistics for Biology and Health. Springer New York, pp. 3–12.

Ingenuity®Systems, 2014. Ingenuity IPA - Integrate and understand complex ’omics data [WWW Document]. Ingenuity. URL http://www.ingenuity.com/products/ipa (accessed 2.10.14).

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostat 4, 249–264. doi:10.1093/biostatistics/4.2.249

Jeanmougin, M., de Reynies, A., Marisa, L., Paccard, C., Nuel, G., Guedj, M., 2010. Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies. PLoS ONE 5, e12336. doi:10.1371/journal.pone.0012336

Jobe, A.H., 2006. The New BPD. Neoreviews 7, e531–e545. doi:10.1542/neo.7-10-e531

Jobe, A.H., Bancalari, E., 2001. Bronchopulmonary Dysplasia. American Journal of Respiratory and Critical Care Medicine 163, 1723–1729. doi:10.1164/ajrccm.163.7.2011060

Johansson, D., Lindgren, P., Berglund, A., 2003. A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. Bioinformatics 19, 467–473. doi:10.1093/bioinformatics/btg017

Johansson, P., Häkkinen, J., 2006. Improving missing value imputation of microarray data by using spot quality weights. BMC Bioinformatics 7, 306. doi:10.1186/1471-2105-7-306

Kauffmann, A., Huber, W., 2010. Microarray data quality control improves the detection of differentially expressed genes. Genomics 95, 138–142. doi:10.1016/j.ygeno.2010.01.003

Kelder, T., Conklin, B.R., Evelo, C.T., Pico, A.R., 2010. Finding the Right Questions: Exploratory Pathway Analysis to Enhance Biological Discovery in Large Datasets. PLoS Biology 8, e1000472. doi:10.1371/journal.pbio.1000472

Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine 7, 673–679. doi:10.1038/89044

Kim, H., Golub, G.H., Park, H., 2006. Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics 22, 1410–1411. doi:10.1093/bioinformatics/btk053

Kim, K.-Y., Kim, B.-J., Yi, G.-S., 2004. Reuse of imputed data in microarray analysis increases imputation efficiency. BMC Bioinformatics 5, 160. doi:10.1186/1471-2105-5-160

Kim, K.-Y., Yi, G.-S., lab, Csb., Information, University, C., 2008. SeqKnn: Sequential KNN imputation method [WWW Document]. URL http://CRAN.R-project.org/package=SeqKnn (accessed 7.11.13).

Köksal, N., Kayik, B., Çetinkaya, M., Özkan, H., Budak, F., Kiliç, Ş., Canitez, Y., Oral, B., 2012. Value of serum and bronchoalveolar fluid lavage pro- and anti-inflammatory cytokine levels for predicting bronchopulmonary dysplasia in premature infants. Eur. Cytokine Netw. 23, 29–35. doi:10.1684/ecn.2012.0304

Kompass, K.S., Deslee, G., Moore, C., McCurnin, D., Pierce, R.A., 2010. Highly conserved transcriptional responses to mechanical ventilation of the lung. Physiological Genomics 42, 384–396. doi:10.1152/physiolgenomics.00117.2009

Kramer, A., Green, J., Pollard, J., Tugendreich, S., 2014. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics 30, 523–530. doi:10.1093/bioinformatics/btt703

Lavoie, P.M., Pham, C., Jang, K.L., 2008. Heritability of Bronchopulmonary Dysplasia, Defined According to the Consensus Statement of the National Institutes of Health. PEDIATRICS 122, 479–485. doi:10.1542/peds.2007-2313

Leys, C., Ley, C., Klein, O., Bernard, P., Licata, L., 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology 49, 764–766. doi:10.1016/j.jesp.2013.03.013

Li, J., Tibshirani, R., 2013. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res 22, 519–536. doi:10.1177/0962280211428386

Liu, L., Hawkins, D.M., Ghosh, S., Young, S.S., 2003. Robust singular value decomposition analysis of microarray data. PNAS 100, 13167–13172. doi:10.1073/pnas.1733249100

Lorkowski, S., Cullen, P.M. (Eds.), 2003. Analysing Gene Expression, A Handbook of Methods: Possibilities and Pitfalls. Wiley.

Melville, J.M., Moss, T.J.M., 2013. The immune consequences of preterm birth. Front Neurosci 7. doi:10.3389/fnins.2013.00079

Modlich, O., Munnes, M., 2007. Statistical Framework for Gene Expression Data Analysis, in: Korenberg, M.J. (Ed.), Microarray Data Analysis, Methods in Molecular Biology. Humana Press, Totowa, NJ, pp. 111–130.

Murphy, D., 2002. Gene Expression Studies Using Microarrays: Principles, Problems, and Prospects. Advan in Physiol Edu 26, 256–270. doi:10.1152/advan.00043.2002

Mutch, D.M., Berger, A., Mansourian, R., Rytz, A., Roberts, M.-A., 2002. The limit fold change model: A practical approach for selecting differentially expressed genes from microarray data. BMC Bioinformatics 3, 17. doi:10.1186/1471-2105-3-17

Nguyen, D.V., Wang, N., Carrol, R.J., 2004. Evaluation of Missing Value Estimation for Microarray Data. Journal of Data Science 2, 347–370.

Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K., Ishii, S., 2003. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 19, 2088–2096. doi:10.1093/bioinformatics/btg287

Oh, S., Kang, D.D., Brock, G.N., Tseng, G.C., 2011. Biological impact of missing-value imputation on downstream analyses of gene expression profiles. Bioinformatics 27, 78–86. doi:10.1093/bioinformatics/btq613

Pearson, R.K., Gonye, G.E., Schwaber, J.S., 2003. Outliers in Microarray Data Analysis, in: Johnson, K.F., Lin, S.M. (Eds.), Methods of Microarray Data Analysis III. Springer US, pp. 41–55.

Pérez-Enciso, M., Tenenhaus, M., 2003. Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. Hum Genet 112, 581–592. doi:10.1007/s00439-003-0921-9

Perrone, S., Tataranno, M., Buonocore, G., 2012. Oxidative stress and bronchopulmonary dysplasia. Journal of Clinical Neonatology 1, 109–114. doi:10.4103/2249-4847.101683

Pietrzyk, J.J., Kwinta, P., Wollen, E.J., Bik-Multanowski, M., Madetko-Talowska, A., Günther, C.-C., Jagła, M., Tomasik, T., Saugstad, O.D., 2013. Gene Expression Profiling in Preterm Infants: New Aspects of Bronchopulmonary Dysplasia Development. PLoS ONE 8, e78585. doi:10.1371/journal.pone.0078585

Quackenbush, J., 2002. Microarray data normalization and transformation. Nat Genet 32, 496–501. doi:10.1038/ng1032

Quackenbush, J., 2001. Computational genetics: Computational analysis of microarray data. Nature Reviews Genetics 2, 418–427. doi:10.1038/35076576

Raychaudhuri, S., Stuart, J.M., Altman, R.B., 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 455–466.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Reimers, M., 2010. Making Informed Choices about Microarray Data Analysis. PLoS Comput Biol 6, e1000786. doi:10.1371/journal.pcbi.1000786

Reiner, A., Yekutieli, D., Benjamini, Y., 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19, 368–375. doi:10.1093/bioinformatics/btf877

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. doi:10.1093/nar/gkv007

Ritchie, M.E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., Smyth, G.K., 2007. A comparison of background correction methods for two-colour microarrays. Bioinformatics 23, 2700–2707. doi:10.1093/bioinformatics/btm412

Rousseeuw, P.J., Croux, C., 1993. Alternatives to the Median Absolute Deviation. Journal of the American Statistical Association 88, 1273–1283. doi:10.1080/01621459.1993.10476408

Sartor, M.A., Tomlinson, C.R., Wesselkamper, S.C., Sivaganesan, S., Leikauf, G.D., Medvedovic, M., 2006. Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. BMC Bioinformatics 7, 538. doi:10.1186/1471-2105-7-538

Saugstad, O.D., 2010. Oxygen and oxidative stress in bronchopulmonary dysplasia. Journal of Perinatal Medicine 38, 571–578. doi:10.1515/jpm.2010.108

Scheel, I., 2007. linImp Package [WWW Document]. URL http://folk.uio.no/idasch/imp/ (accessed 8.19.14).

Scheel, I., Aldrin, M., Glad, I.K., Sørum, R., Lyng, H., Frigessi, A., 2005. The influence of missing value imputation on detection of differentially expressed genes from microarray data. Bioinformatics 21, 4272–4279. doi:10.1093/bioinformatics/bti708

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., Herzel, H., 2000. Normalization strategies for cDNA microarrays. Nucl. Acids Res. 28, e47. doi:10.1093/nar/28.10.e47

Shaw, G.M., O’Brodovich, H.M., 2013. Progress in understanding the genetics of Bronchopulmonary Dysplasia. Semin Perinatol 37, 85–93. doi:10.1053/j.semperi.2013.01.004

Sherman, B.T., Huang, D.W., Tan, Q., Guo, Y., Bour, S., Liu, D., Stephens, R., Baseler, M.W., Lane, H.C., Lempicki, R.A., 2007. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics 8, 426. doi:10.1186/1471-2105-8-426

Shieh, A.D., Hung, Y.S., 2009. Detecting Outlier Samples in Microarray Data. Statistical Applications in Genetics and Molecular Biology 8, 1–24. doi:10.2202/1544-6115.1426

Silver, J.D., Ritchie, M.E., Smyth, G.K., 2009. Microarray background correction: maximum likelihood estimation for the normal–exponential convolution. Biostat 10, 352–363. doi:10.1093/biostatistics/kxn042

Smyth, G.K., 2005. Limma: linear models for microarray data, in: Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., Huber, W. (Eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, New York, pp. 397–420.

Smyth, G.K., 2004. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical Applications in Genetics and Molecular Biology 3. doi:10.2202/1544-6115.1027

Smyth, G.K., Speed, T., 2003. Normalization of cDNA microarray data. Methods 31, 265–273. doi:10.1016/S1046-2023(03)00155-5

Smyth, M.G., biocViews Microarray, O., TwoChannel, D., QualityControl, P., MultipleComparisons, T., 2013. Package "limma."

Speer, C.P., 2006. Inflammation and bronchopulmonary dysplasia: A continuing story. Seminars in Fetal and Neonatal Medicine 11, 354–362. doi:10.1016/j.siny.2006.03.004

Speer, C.P., 2003. Inflammation and bronchopulmonary dysplasia. Seminars in Neonatology 8, 29–38. doi:10.1016/S1084-2756(02)00190-2

Stacklies, W., Redestig, H., Scholz, M., Walther, D., Selbig, J., 2007. pcaMethods—a bioconductor package providing PCA methods for incomplete data. Bioinformatics 23, 1164–1167. doi:10.1093/bioinformatics/btm069

Storey, J.D., 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, 479–498. doi:10.1111/1467-9868.00346

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., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102, 15545–15550. doi:10.1073/pnas.0506580102

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T.R., 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. PNAS 96, 2907–2912. doi:10.1073/pnas.96.6.2907

Thompson, A., Bhandari, V., 2008. Pulmonary Biomarkers of Bronchopulmonary Dysplasia. Biomark Insights 3, 361–373.

Tibshirani, R., Chu, G., Narasimhan, B., Li, J., 2011. samr: SAM: Significance Analysis of Microarrays.

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS 99, 6567–6572. doi:10.1073/pnas.082099299

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17, 520–525. doi:10.1093/bioinformatics/17.6.520

Tuikkala, J., Elo, L.L., Nevalainen, O.S., Aittokallio, T., 2008. Missing value imputation improves clustering and interpretation of gene expression microarray data. BMC Bioinformatics 9, 202. doi:10.1186/1471-2105-9-202

Tuikkala, J., Elo, L., Nevalainen, O.S., Aittokallio, T., 2006. Improving missing value estimation in microarray data with gene ontology. Bioinformatics 22, 566–572. doi:10.1093/bioinformatics/btk019

Tusher, V.G., Tibshirani, R., Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. PNAS 98, 5116–5121. doi:10.1073/pnas.091062498

Tu, Y., Stolovitzky, G., Klein, U., 2002. Quantitative noise analysis for gene expression microarray experiments. PNAS 99, 14031–14036. doi:10.1073/pnas.222164199

Walsh, M.C., Szefler, S., Davis, J., Allen, M., Van Marter, L., Abman, S., Blackmon, L., Jobe, A., 2006. Summary proceedings from the bronchopulmonary dysplasia group. Pediatrics 117, S52–56. doi:10.1542/peds.2005-0620I

Wang, H., St. Julien, K.R., Stevenson, D.K., Hoffmann, T.J., Witte, J.S., Lazzeroni, L.C., Krasnow, M.A., Quaintance, C.C., Oehlert, J.W., Jelliffe-Pawlowski, L.L., Gould, J.B., Shaw, G.M., O’Brodovich, H.M., 2013. A Genome-Wide Association Study (GWAS) for Bronchopulmonary Dysplasia. PEDIATRICS 132, 290–297. doi:10.1542/peds.2013-0533

Ward, J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58, 236–244. doi:10.1080/01621459.1963.10500845

Workman, C., Jensen, L.J., Jarmer, H., Berka, R., Gautier, L., Nielser, H.B., Saxild, H.-H., Nielsen, C., Brunak, S., Knudsen, S., 2002. A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biology 3, research0048. doi:10.1186/gb-2002-3-9-research0048

Wu, W., Dave, N., Tseng, G.C., Richards, T., Xing, E.P., Kaminski, N., 2005. Comparison of normalization methods for CodeLink Bioarray data. BMC Bioinformatics 6, 309. doi:10.1186/1471-2105-6-309

Yang, S., Guo, X., Yang, Y.-C., Papcunik, D., Heckman, C., Hooke, J., Shriver, C.D., Liebman, M.N., Hu, H., 2007. Detecting Outlier Microarray Arrays by Correlation and Percentage of Outliers Spots. Cancer Inform 2, 351–360.

Yeoh, E.-J., Ross, M.E., Shurtleff, S.A., Williams, W.K., Patel, D., Mahfouz, R., Behm, F.G., Raimondi, S.C., Relling, M.V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C.-H., Evans, W.E., Naeve, C., Wong, L., Downing, J.R., 2002. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 1, 133–143. doi:10.1016/S1535-6108(02)00032-6

Yu, L., Gulati, P., Fernandez, S., Pennell, M., Kirschner, L., Jarjoura, D., 2011. Fully Moderated T-statistic for Small Sample Size Gene Expression Arrays. Stat Appl Genet Mol Biol 10. doi:10.2202/1544-6115.1701

8 Appendix

8.1 Supplemental material and tables

Table 8-1 Differentially regulated transcripts (39) in mild BPD compared to no BPD and moderate/severe BPD

False discovery rate (FDR) and fold change (FC) were calculated with LIMMA. ACCN: Genbank Accession number, Gene ID from Entrez Gene ID database, and Symbols are Hugo gene symbols obtained with SOURCE.

ACCN	Name	Symbol	Gene ID	FDR mild – no BPD	FDR mod./ severe – no BPD	FDR mod./ severe – mild BPD	FC mild – no BPD	FC mod./ severe – no BPD	FC mod./ severe – mild BPD
NM_001311	Cysteine-rich protein 1 (intestinal)	CRIP1	1396	0.000	0.764	0.000	-8.53	1.44	12.30
NM_005129				0.001	0.992	0.007	-5.64	1.03	5.81
NM_001607	Acetyl-CoA acyltransferase 1	ACAA1	30	0.004	0.724	0.008	-3.12	1.51	4.72
NM_004545	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 1, 7kDa	NDUFB1	4707	0.025	0.732	0.031	-3.06	1.62	4.95
NM_005608	Protein tyrosine phosphatase, receptor type, C-associated protein	PTPRCAP	5790	0.001	0.925	0.017	-2.58	1.18	3.05
NM_018335	Zinc finger protein 839	ZNF839	55778	0.001	0.978	0.007	-2.54	1.06	2.71
AB011126	Formin binding protein 1	FNBP1	23048	0.008	0.916	0.033	-2.33	1.20	2.79
NM_004891	Mitochondrial ribosomal protein L33	MRPL33	9553	0.010	0.780	0.022	-2.33	1.36	3.17
NM_004427	Polyhomeotic homolog 2 (Drosophila)	PHC2	1912	0.012	0.587	0.012	-2.29	1.60	3.67
BC010420	Arginyl-tRNA synthetase 2, mitochondrial	RARS2	57038	0.001	0.860	0.008	-2.26	1.20	2.70
NM_032031				0.022	0.855	0.041	-2.20	1.29	2.83
NM_015414	Ribosomal protein L36	RPL36	25873	0.034	0.694	0.033	-2.04	1.43	2.91
M77140	Galanin prepropeptide	GAL	51083	0.000	0.975	0.000	15.49	1.14	-13.59
NM_020689	Solute carrier family 24 (sodium/potassium/calcium exchanger), member 3	SLC24A3	57419	0.000	0.997	0.006	4.10	-1.01	-4.16
NM_001257	Cadherin 13, H-cadherin (heart)	CDH13	1012	0.000	0.985	0.008	4.06	1.05	-3.87
NM_001585				0.001	0.993	0.034	3.70	1.03	-3.58
AL049332				0.000	0.910	0.001	2.92	-1.15	-3.34
NM_002387	Mutated in colorectal cancers	MCC	4163	0.003	0.993	0.045	2.65	1.03	-2.58
AB028949				0.000	0.981	0.003	2.63	-1.04	-2.75
NM_003672	CDC14 cell division cycle 14 homolog A (S. cerevisiae)	CDC14A	8556	0.000	0.862	0.003	2.52	-1.18	-2.97
NM_031885	Bardet-Biedl syndrome 2	BBS2	583	0.001	0.986	0.026	2.48	-1.04	-2.58
NM_002570	Proprotein convertase subtilisin/kexin type 6	PCSK6	5046	0.010	0.913	0.040	2.46	-1.25	-3.06
NM_014240	LIM domains containing 1	LIMD1	8994	0.001	1.000	0.024	2.45	1.00	-2.45

ACCN	Name	Symbol	Gene ID	FDR mild – no BPD	FDR mod./ severe – no BPD	FDR mod./ severe – mild BPD	FC mild – no BPD	FC mod./ severe – no BPD	FC mod./ severe – mild BPD
AL122083				0.006	0.982	0.048	2.42	-1.06	-2.56
NM_016735				0.003	0.972	0.036	2.38	-1.10	-2.62
NM_005267	Gap junction protein, alpha 8, 50kDa	GJA8	2703	0.002	0.981	0.038	2.33	1.05	-2.21
NM_003146	Structure specific recognition protein 1	SSRP1	6749	0.001	0.985	0.018	2.31	-1.03	-2.38
NM_018050	MANSC domain containing 1	MANSC1	54682	0.008	0.853	0.024	2.25	-1.26	-2.83
NM_017803	Dihydrouridine synthase 2-like, SMM1 homolog (S. cerevisiae)	DUS2L	54920	0.008	0.951	0.038	2.22	-1.14	-2.53
AL096732	Dynein, axonemal, heavy chain 3	DNAH3	55567	0.006	0.808	0.022	2.14	-1.31	-2.80
NM_005201	Chemokine (C-C motif) receptor 8	CCR8	1237	0.006	0.971	0.038	2.13	-1.09	-2.32
AL080111	NIMA (never in mitosis gene a)-related kinase 7	NEK7	140609	0.011	0.836	0.029	2.09	-1.26	-2.63
AF063936	Immunoglobulin superfamily, DCC subclass, member 3	IGDCC3	9543	0.003	0.961	0.029	2.07	-1.10	-2.29
NM_133631	Roundabout, axon guidance receptor, homolog 1 (Drosophila)	ROBO1	6091	0.006	0.971	0.038	2.06	-1.09	-2.25
NM_002939	Ribonuclease/angiogenin inhibitor 1	RNH1	6050	0.021	0.528	0.013	2.05	-1.63	-3.33
NM_001466	Frizzled homolog 2 (Drosophila)	FZD2	2535	0.010	0.208	0.003	2.04	-1.86	-3.79
NM_013305	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 5	ST8SIA5	29906	0.017	0.716	0.026	2.03	-1.39	-2.83
NM_001065	Tumor necrosis factor receptor superfamily, member 1A	TNFRSF1A	7132	0.010	0.944	0.040	2.02	-1.14	-2.30
NM_007147	Zinc finger protein 175	ZNF175	7728	0.011	0.266	0.003	2.01	-1.80	-3.61

Table 8-2 Differentially regulated transcripts (27) in mild BPD compared to no BPD and moderate/severe BPD

False discovery rate (FDR) and fold change (FC) were calculated with LIMMA. ACCN: Genbank Accession number, Gene ID from Entrez Gene ID database, and Symbols are Hugo gene symbols obtained with SOURCE.

ACCN	Name	Symbol	Gene ID	FDR mild – no BPD	FDR mod./ severe – no BPD	FDR mod./ severe – mild BPD	FC mild – no BPD	FC mod./ severe – no BPD	FC mod./ severe – mild BPD
NM_020995	Haptoglobin-related protein	HPR	3250	0.974	0.001	0.001	-1.03	19.73	20.31
NM_003618	Mitogen-activated protein kinase 3	MAP4K3	8491	0.220	0.001	0.022	1.68	6.67	3.96
NM_005091	Peptidoglycan recognition protein 1	PGLYRP1	8993	0.553	0.014	0.005	-1.37	5.76	7.87
AF453583	Secretogranin III	SCG3	29106	0.068	0.000	0.000	1.53	5.44	3.55
NM_021958	H2.0-like homeobox	HLX	3142	0.845	0.005	0.005	-1.10	4.55	5.01
NM_001785	Cytidine deaminase	CDA	978	0.384	0.015	0.004	-1.40	4.45	6.25
NM_005849	Immunoglobulin superfamily, member 6	IGSF6	10261	0.616	0.045	0.018	-1.31	4.06	5.33
NM_000348	Steroid-5-alpha-reductase, alpha polypeptide 2 (3-oxo-5 alpha-steroid delta 4-dehydrogenase alpha 2)	SRD5A2	6716	0.523	0.003	0.017	1.24	3.87	3.11
NM_014870	Zinc finger and BTB domain containing 40	ZBTB40	9923	0.906	0.003	0.006	1.05	3.59	3.41
NM_006145	DnaJ (Hsp40) homolog, subfamily B, member 1	DNAJB1	3337	0.527	0.033	0.010	-1.24	3.12	3.86
NM_016364	Dual specificity phosphatase 13	DUSP13	51207	0.780	0.015	0.013	-1.13	3.04	3.43
NM_007219	Ring finger protein 24	RNF24	11237	0.843	0.047	0.033	-1.11	3.00	3.31
NM_014213	Homeobox D9	HOXD9	3235	0.815	0.026	0.040	1.10	2.84	2.58
NM_001039	Sodium channel, nonvoltage-gated 1, gamma	SCNN1G	6340	0.323	0.041	0.007	-1.35	2.74	3.70
NM_022154	Solute carrier family 39 (zinc transporter), member 8	SLC39A8	64116	0.529	0.042	0.013	-1.25	2.69	3.36
NM_024327	Zinc finger protein 576	ZNF576	79177	0.785	0.003	0.003	-1.09	2.68	2.91
NM_006399	Basic leucine zipper transcription factor, ATF-like	BATF	10538	0.809	0.043	0.029	-1.09	2.45	2.68
NM_007222	Zinc fingers and homeoboxes 1	ZHX1	11244	0.844	0.032	0.046	1.07	2.33	2.18
NM_000892	Kallikrein B, plasma (Fletcher factor) 1	KLKB1	3818	0.690	0.005	0.003	-1.09	2.03	2.20
NM_017946	FK506 binding protein 14, 22 kDa	FKBP14	55033	0.698	0.006	0.004	1.13	-3.04	-3.43
BC009709	Guanine nucleotide binding protein (G protein), gamma 11	GNG11	2791	0.237	0.001	0.029	-1.35	-2.98	-2.21
NM_018638	Ethanolamine kinase 1	ETNK1	55500	0.994	0.037	0.036	1.00	-2.73	-2.74
NM_032010				0.933	0.036	0.040	-1.04	-2.60	-2.50
NM_000053	ATPase, Cu++ transporting, beta polypeptide	ATP7B	540	0.731	0.014	0.007	1.09	-2.37	-2.58
NM_002878	RAD51-like 3 (S. cerevisiae)	RAD51L3	5892	0.849	0.043	0.029	1.07	-2.28	-2.43
NM_003941	Wiskott-Aldrich syndrome-like	WASL	8976	0.989	0.042	0.040	1.01	-2.26	-2.28
NM_017588	WD repeat domain 5	WDR5	11091	0.719	0.014	0.031	-1.09	-2.26	-2.07

Table 8-3 Differentially regulated transcripts (4) in mild BPD and moderate/severe BPD compared to no BPD

False discovery rate (FDR) and fold change (FC) were calculated with LIMMA. ACCN: Genbank Accession number, Gene ID from Entrez Gene ID database, and Symbols are Hugo gene symbols obtained with SOURCE.

ACCN	Name	Symbol	Gene ID	FDR mild – no BPD	FDR mod./ severe – no BPD	FDR mod./ severe – mild BPD	FC mild – no BPD	FC mod./ severe – no BPD	FC mod./ severe – mild BPD
NM_030807	Solute carrier family 2 (facilitated glucose transporter), member 11	SLC2A11	66035	0.000	0.014	0.888	6.93	6.15	-1.13
NM_001103	Actinin, alpha 2	ACTN2	88	0.036	0.033	0.455	2.16	3.28	1.52
NM_018104				0.005	0.029	0.840	2.11	2.29	1.09
NM_003832				0.000	0.013	0.870	-16.74	-13.81	1.21

Table 8-4 Transcripts (71) able to discriminate between BPD groups, PAM threshold = 2.2

False discovery rate (FDR) and fold change (FC) were calculated with LIMMA. ACCN: Genbank Accession number, Gene ID from Entrez Gene ID database, and Symbols are Hugo gene symbols obtained with SOURCE.

ACCN	Name	Symbol	Gene ID	FDR mild – no BPD	FDR mod./ severe – no BPD	FDR mod./ severe – mild BPD	FC mild – no BPD	FC mod./ severe – no BPD	FC mod./ severe – mild BPD
AA318707	S100 calcium binding protein A9	S100A9	6280	0.944	0.351	0.192	-1.2	21.3	25.3
AB028949				0.000	0.981	0.003	2.6	-1.0	-2.7
AF095735	Sarcosine dehydrogenase	SARDH	1757	0.034	0.943	0.093	-2.8	1.3	3.6
AJ223280	Linker for activation of T cells	LAT	27040	0.010	0.975	0.135	-3.2	-1.1	2.8
AK001143				0.011	0.973	0.162	3.6	1.2	-3.1
AK001814	Hypothetical LOC100505876	LOC100505876	100505876	0.100	0.115	0.520	2.2	3.6	1.6
AK002039	Murine retrovirus integration site 1 homolog	MRVI1	10335	0.001	0.584	0.154	2.1	1.3	-1.6
AK024496	Kelch domain containing 4	KLHDC4	54758	0.011	0.855	0.256	-2.8	-1.4	2.0
AL049274	Mannosidase, alpha, class 1C, member 1	MAN1C1	57134	0.003	0.981	0.057	-2.8	-1.1	2.6
AL049332	BTG family, member 3	BTG3	10950	0.000	0.910	0.001	2.9	-1.1	-3.3
AL110274	aldehyde dehydrogenase 1 family, member A2	ALDH1A2	8854	0.006	0.941	0.126	2.7	1.2	-2.3
AL162053	F-box protein 3	FBXO3	26273	0.018	0.124	0.883	-2.1	-2.2	-1.1
AL539691				0.000	0.548	0.101	-4.8	-1.8	2.6

ACCN	Name	Symbol	Gene ID	FDR mild – no BPD	FDR mod./ severe – no BPD	FDR mod./ severe – mild BPD	FC mild – no BPD	FC mod./ severe – no BPD	FC mod./ severe – mild BPD
BM741997	Solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 3	SLC25A3	5250	0.001	0.191	0.413	-4.8	-2.8	1.7
M77140	Galanin prepropeptide	GAL	51083	0.000	0.975	0.000	15.5	1.1	-13.6
NM_000275	Oculocutaneous albinism II	OCA2	4948	0.006	0.999	0.067	2.5	1.0	-2.4
NM_000732	CD3d molecule, delta (CD3-TCR complex)	CD3D	915	0.039	0.971	0.278	-2.9	-1.2	2.4
NM_000985	Ribosomal protein L17	RPL17	6139	0.012	0.962	0.053	-3.9	1.2	4.8
NM_001257	Cadherin 13, H-cadherin (heart)	CDH13	1012	0.000	0.985	0.008	4.1	1.0	-3.9
NM_001311	Cysteine-rich protein 1 (intestinal)	CRIP1	1396	0.000	0.764	0.000	-8.5	1.4	12.3
NM_001585				0.001	0.993	0.034	3.7	1.0	-3.6
NM_001607	Acetyl-CoA acyltransferase 1	ACAA1	30	0.004	0.724	0.008	-3.1	1.5	4.7
NM_001730	Kruppel-like factor 5 (intestinal)	KLF5	688	0.023	0.389	0.008	-2.0	1.7	3.3
NM_001752	Catalase	CAT	847	0.007	0.389	0.532	6.7	3.4	-2.0
NM_001803	CD52 molecule	CD52	1043	0.027	0.992	0.147	-2.7	1.0	2.9
NM_002108	Histidine ammonia-lyase	HAL	3034	0.020	0.971	0.207	2.6	1.2	-2.3
NM_002416	Chemokine (C-X-C motif) ligand 9	CXCL9	4283	0.006	0.983	0.086	2.9	1.1	-2.7
NM_002570	Proprotein convertase subtilisin/kexin type 6	PCSK6	5046	0.010	0.913	0.040	2.5	-1.2	-3.1
NM_002777	Proteinase 3	PRTN3	5657	0.321	0.015	0.068	1.4	3.9	2.8
NM_003578	Sterol O-acyltransferase 2	SOAT2	8435	0.015	0.780	0.354	2.7	1.5	-1.8
NM_003726	Src kinase associated phosphoprotein 1	SKAP1	8631	0.032	0.993	0.183	-2.7	-1.0	2.6
NM_003832				0.000	0.013	0.870	-16.7	-13.8	1.2
NM_003841	Tumor necrosis factor receptor superfamily, member 10c, decoy without an intracellular domain	TNFRSF10C	8794	0.006	0.640	0.342	2.8	1.6	-1.7
NM_004049	BCL2-related protein A1	BCL2A1	597	0.363	0.312	0.492	2.8	8.7	3.1
NM_004221	Interleukin 32	IL32	9235	0.073	0.955	0.142	-2.5	1.2	3.1
NM_004305	Bridging integrator 1	BIN1	274	0.039	0.994	0.202	-2.3	-1.0	2.2
NM_004418	Dual specificity phosphatase 2	DUSP2	1844	0.025	0.735	0.465	-6.8	-2.5	2.7
NM_004427	Polyhomeotic homolog 2 (Drosophila)	PHC2	1912	0.012	0.587	0.012	-2.3	1.6	3.7
NM_004545	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 1, 7kDa	NDUFB1	4707	0.025	0.732	0.031	-3.1	1.6	5.0
NM_005091	Peptidoglycan recognition protein 1	PGLYRP1	8993	0.553	0.014	0.005	-1.4	5.8	7.9
NM_005129				0.001	0.992	0.007	-5.6	1.0	5.8
NM_005389	Protein-L-isoaspartate (D-aspartate) O-methyltransferase	PCMT1	5110	0.022	0.887	0.056	-2.2	1.3	2.7
NM_005608	Protein tyrosine phosphatase, receptor	PTPRCAP	5790	0.001	0.925	0.017	-2.6	1.2	3.1

ACCN	Name	Symbol	Gene ID	FDR mild – no BPD	FDR mod./ severe – no BPD	FDR mod./ severe – mild BPD	FC mild – no BPD	FC mod./ severe – no BPD	FC mod./ severe – mild BPD
	type, C-associated protein								
NM_005849	Immunoglobulin superfamily, member 6	IGSF6	10261	0.616	0.045	0.018	-1.3	4.1	5.3
NM_006308	Heat shock 27kDa protein 3	HSPB3	8988	0.001	0.862	0.056	2.6	1.2	-2.2
NM_006746	Sex comb on midleg-like 1 (Drosophila)	SCML1	6322	0.006	0.256	0.822	-2.4	-2.1	1.1
NM_007147	Zinc finger protein 175	ZNF175	7728	0.011	0.266	0.003	2.0	-1.8	-3.6
NM_014299	Bromodomain containing 4	BRD4	23476	0.060	0.978	0.176	2.5	-1.1	-2.9
NM_014381	MutL homolog 3 (E. coli)	MLH3	27030	0.025	0.960	0.086	-2.6	1.2	3.2
NM_014801	Pecanex-like 2 (Drosophila)	PCNXL2	80003	0.003	0.961	0.076	-2.9	-1.2	2.5
NM_015987	Heme binding protein 1	HEBP1	50865	0.039	0.554	0.766	12.9	6.8	-1.9
NM_016735				0.003	0.972	0.036	2.4	-1.1	-2.6
NM_017947	Molybdenum cofactor sulfurase	MOCOS	55034	0.865	0.059	0.076	1.1	5.3	4.6
NM_018050	MANSC domain containing 1	MANSC1	54682	0.008	0.853	0.024	2.2	-1.3	-2.8
NM_018104				0.005	0.029	0.840	2.1	2.3	1.1
NM_018356	Chromosome 5 open reading frame 22	C5orf22	55322	0.053	0.999	0.225	2.5	1.0	-2.5
NM_018427	RRN3 RNA polymerase I transcription factor homolog (S. cerevisiae)	RRN3	54700	0.001	0.928	0.063	-4.0	-1.2	3.2
NM_018457	Proline rich 13	PRR13	54458	0.012	0.971	0.056	-3.4	1.2	3.9
NM_020689	Solute carrier family 24 (sodium/potassium/calcium exchanger), member 3	SLC24A3	57419	0.000	0.997	0.006	4.1	-1.0	-4.2
NM_020995	Haptoglobin-related protein	HPR	3250	0.974	0.001	0.001	-1.0	19.7	20.3
NM_021958	H2.0-like homeobox	HLX	3142	0.845	0.005	0.005	-1.1	4.5	5.0
NM_022154	Solute carrier family 39 (zinc transporter), member 8	SLC39A8	64116	0.529	0.042	0.013	-1.2	2.7	3.4
NM_025084				0.013	0.997	0.106	-3.0	1.0	3.1
NM_030751	Zinc finger E-box binding homeobox 1	ZEB1	6935	0.064	0.043	0.404	-2.3	-4.2	-1.8
NM_030807	Solute carrier family 2 (facilitated glucose transporter), member 11	SLC2A11	66035	0.000	0.014	0.888	6.9	6.2	-1.1
NM_031296	RAB33B, member RAS oncogene family	RAB33B	83452	0.007	0.925	0.172	2.3	1.2	-1.9
NM_031885	Bardet-Biedl syndrome 2	BBS2	583	0.001	0.986	0.026	2.5	-1.0	-2.6
NM_032031				0.022	0.855	0.041	-2.2	1.3	2.8
NM_032621	Brain expressed X-linked 2	BEX2	84707	0.004	0.997	0.057	-2.1	1.0	2.2
NM_052972	Leucine-rich alpha-2-glycoprotein 1	LRG1	116844	0.521	0.037	0.109	1.4	4.2	3.0
X00437	Interleukin 23, alpha subunit p19	IL23A	51561	0.034	0.992	0.195	-2.9	-1.1	2.8

Table 8-5 DAVID functional annotation clustering for 58 transcripts able to differentiate between BPD groups and are differentially regulated between at least two BPD groups

False discovery rate (FDR) and fold change (FC) were calculated with LIMMA. ACCN: Genbank Accession number, Gene ID from Entrez Gene ID database, and Symbols are Hugo gene symbols obtained with SOURCE.

ACCN	Name	Symbol	Gene ID	FDR mild – no BPD	FDR moderate/ severe – no BPD	FDR moderate/ severe – mild BPD	FC mild – no BPD	FC moderate/ severe – no BPD	FC moderate/ severe – mild BPD
regulation of leukocyte activation (ES: 1.65)									
M77140	Galanin prepropeptide	GAL	51083	0.000	0.975	0.000	15.5	1.1	-13.6
NM_021958	H2.0-like homeobox	HLX	3142	0.845	0.005	0.005	-1.1	4.5	5.0
NM_030751	Zinc finger E-box binding homeobox 1	ZEB1	6935	0.064	0.043	0.404	-2.3	-4.2	-1.8
AJ223280	Linker for activation of T cells	LAT	27040	0.010	0.975	0.135	-3.2	-1.1	2.8
regulation of cell proliferation/ embryonic organ development (ES: 1.19)									
M77140	Galanin prepropeptide	GAL	51083	0.000	0.975	0.000	15.5	1.1	-13.6
NM_001257	Cadherin 13, H-cadherin (heart)	CDH13	1012	0.000	0.985	0.008	4.1	1.0	-3.9
AL049332	BTG family, member 3	BTG3	10950	0.000	0.910	0.001	2.9	-1.1	-3.3
AL110274	aldehyde dehydrogenase 1 family, member A2	ALDH1A2	8854	0.006	0.941	0.126	2.7	1.2	-2.3
NM_002777	Proteinase 3	PRTN3	5657	0.321	0.015	0.068	1.4	3.9	2.8
NM_021958	H2.0-like homeobox	HLX	3142	0.845	0.005	0.005	-1.1	4.5	5.0
NM_001730	Kruppel-like factor 5 (intestinal)	KLF5	688	0.023	0.389	0.008	-2.0	1.7	3.3
NM_030751	Zinc finger E-box binding homeobox 1	ZEB1	6935	0.064	0.043	0.404	-2.3	-4.2	-1.8

Table 8-6 **Transcripts (210) obtained from regression model explaining gene expression by the need of assisted ventilation, oxygen support, or the interaction of ventilation and oxygen support under consideration of maturity.**

ACCN	Symbol	Gene ID	p-value for ventilation	p-value for oxygen	p-value for interaction of ventilation and oxygen	FC ventilation	FC oxygen	FC ventilation: oxygen	FC cluster 1 vs. 2	FC cluster 1 vs. 3	FC cluster 2 vs. 3
NM_016471			0.176	0.000	0.051	-0.98	-1.00	0.50	-1.31	1.18	1.55
NM_012417	PITPNC1	26207	0.698	0.000	0.686	1.07	1.00	-0.50	-1.01	1.25	1.26
NM_007233			0.020	0.000	0.021	1.02	-1.00	0.50	-1.13	1.16	1.31
NM_003107	SOX4	6659	0.680	0.000	0.143	1.25	-1.00	-2.00	1.47	3.52	2.39
NM_006963	ZNF22	7570	0.326	0.001	0.870	0.91	-1.00	-2.00	-1.07	1.31	1.41
NM_001558	IL10RA	3587	0.276	0.001	0.968	1.14	-1.00	-2.00	-1.26	1.67	2.10
NM_000734	CD247	919	0.469	0.001	0.489	1.08	-1.00	-2.00	1.10	1.71	1.56
AL137521	TMEM104	54868	0.333	0.001	0.140	-1.00	1.00	-0.50	1.19	-1.14	-1.37
NM_004585	RARRES3	5920	0.161	0.001	0.286	1.11	-1.00	0.50	-1.09	1.40	1.53
NM_001383	DPH1	1801	0.033	0.001	0.303	-0.92	1.00	2.00	-1.39	-1.22	1.14
NM_013416	NCF4	4689	0.065	0.001	0.805	1.23	-1.00	-2.00	-1.32	3.06	4.03
NM_004305	BIN1	274	0.084	0.001	0.734	1.15	-1.00	-2.00	1.09	1.66	1.52
NM_017415	KLHL3	26249	0.091	0.001	0.461	1.15	-1.00	0.50	-1.18	1.55	1.82
AK057700	ATP6V0E2	155066	0.138	0.001	0.717	1.06	-1.00	0.50	1.00	1.71	1.71
AJ223280	LAT	27040	0.548	0.002	0.641	1.32	1.00	2.00	-1.07	2.37	2.55
NM_004716	PCSK7	9159	0.378	0.002	0.403	1.09	-1.00	0.50	-1.13	1.33	1.51
NM_006746	SCML1	6322	0.117	0.002	0.182	1.01	-1.00	0.50	-1.27	1.20	1.52
NM_004310	RHOH	399	0.388	0.002	0.069	0.96	-1.00	-2.00	1.35	1.39	1.03
X00437	IL23A	51561	0.269	0.002	0.833	1.22	-1.00	0.50	-1.11	1.88	2.08
NM_030807	SLC2A11	66035	0.097	0.002	0.093	0.97	1.00	-0.50	1.63	-1.11	-1.81
NM_014875	KIF14	9928	0.202	0.002	0.360	-1.03	1.00	-0.50	1.19	-1.20	-1.43
AL162053	FBXO3	26273	0.145	0.002	0.242	-0.98	-1.00	0.50	-1.35	1.37	1.85
NM_000449	RFX5	5993	0.233	0.002	0.111	-0.94	-1.00	0.50	-1.24	1.14	1.41
NM_002309			0.765	0.002	0.031	-0.97	-1.00	0.50	-1.27	1.12	1.42
NM_002827	PTPN1	5770	0.222	0.002	0.450	-1.08	1.00	-0.50	1.05	-1.35	-1.42
NM_033544	RCCD1	91433	0.998	0.002	0.032	-0.97	-1.00	0.50	-1.84	1.24	2.29
NM_022761	C11orf1	64776	0.505	0.002	0.001	-0.95	-1.00	0.50	-1.86	-1.03	1.81
NM_030911	CDADC1	81602	0.978	0.003	0.952	-1.10	1.00	2.00	-1.05	-1.56	-1.49
NM_031209	QTRT1	81890	0.186	0.003	0.807	1.13	-1.00	-2.00	-1.02	1.63	1.67
NM_014911	AAK1	22848	0.210	0.003	0.937	-1.03	1.00	2.00	1.03	-1.16	-1.19
NM_004580	RAB27A	5873	0.303	0.003	0.213	-0.77	-1.00	0.50	-1.64	2.79	4.57
NM_001894	CSNK1E	1454	0.514	0.003	0.710	1.10	-1.00	-2.00	1.05	1.85	1.76
NM_004418	DUSP2	1844	0.801	0.003	0.679	1.00	-1.00	0.50	-1.30	1.86	2.43
NM_000073	CD3G	917	0.081	0.003	0.560	1.04	-1.00	-2.00	1.16	1.41	1.21

ACCN	Symbol	Gene ID	p-value for ventilation	p-value for oxygen	p-value for interaction of ventilation and oxygen	FC ventilation	FC oxygen	FC ventilation: oxygen	FC cluster 1 vs. 2	FC cluster 1 vs. 3	FC cluster 2 vs. 3
NM_004753	DHRS3	9249	0.759	0.004	0.604	1.12	-1.00	0.50	-1.09	1.57	1.71
AA601902			0.453	0.004	0.943	1.15	-1.00	-2.00	1.20	2.38	1.98
AF244129	LY9	4063	0.383	0.004	0.418	1.12	-1.00	0.50	-1.18	1.42	1.68
NM_030819	GFOD2	81577	0.257	0.004	0.049	0.95	-1.00	0.50	-1.16	1.14	1.32
NM_005682	GPR56	9289	0.453	0.004	0.586	0.99	-1.00	-2.00	1.23	1.52	1.24
NM_000732	CD3D	915	0.521	0.004	0.341	1.25	-1.00	0.50	-1.37	2.32	3.17
U43368	VEGFB	7423	0.693	0.004	0.794	1.13	-1.00	-2.00	-1.02	1.83	1.87
NM_003726	SKAP1	8631	0.096	0.004	0.972	1.17	-1.00	-2.00	-1.03	1.64	1.68
NM_003890	FCGBP	8857	0.455	0.004	0.905	1.08	-1.00	0.50	-1.09	1.54	1.68
BE378990	OGDH	4967	0.116	0.004	0.372	-0.93	-1.00	0.50	-1.44	1.24	1.78
NM_003463	PTP4A1	7803	0.490	0.004	0.057	1.05	1.00	-0.50	1.37	1.04	-1.31
NM_007011	ABHD2	11057	0.182	0.004	0.100	-1.16	-1.00	0.50	-1.31	-1.95	-1.49
NM_003330	TXNRD1	7296	0.788	0.004	0.075	-1.07	1.00	2.00	-1.20	-1.51	-1.26
NM_018075	ANO10	55129	0.443	0.004	0.580	-1.08	1.00	2.00	-1.29	-1.53	-1.18
NM_002833	PTPN9	5780	0.489	0.005	0.956	-1.13	1.00	2.00	1.00	-1.77	-1.77
NM_018346	RSAD1	55316	0.905	0.005	0.757	1.08	-1.00	-2.00	1.00	1.63	1.62
NM_021064	HIST1H2AG	8969	0.774	0.005	0.299	-1.01	1.00	-0.50	1.34	-1.16	-1.55
NM_002539	ODC1	4953	0.772	0.005	0.258	1.04	-1.00	0.50	-1.00	1.35	1.35
NM_018381	C19orf66	55337	0.815	0.005	0.937	1.12	-1.00	0.50	-1.06	1.63	1.73
NM_018074	CCDC94	55702	0.404	0.005	0.184	-0.95	-1.00	0.50	-1.26	1.25	1.58
AB029010	SLC8A2	6543	0.505	0.005	0.651	-1.08	1.00	2.00	1.02	-1.69	-1.73
NM_030984	TBXAS1	6916	0.180	0.005	0.121	0.96	1.00	-0.50	1.35	-1.30	-1.75
NM_002106	H2AFZ	3015	0.460	0.006	0.752	1.10	-1.00	0.50	-1.07	1.51	1.62
AL137416	C15orf55	256646	0.003	0.006	0.766	-0.95	1.00	-0.50	-1.12	1.03	1.16
NM_006357	UBE2E3	10477	0.442	0.006	0.542	0.89	-1.00	-2.00	1.21	-1.39	-1.68
NM_007240	DUSP12	11266	0.700	0.006	0.186	0.96	-1.00	0.50	-1.49	1.08	1.61
NM_019884	GSK3A	2931	0.402	0.006	0.094	0.96	1.00	-0.50	1.44	-1.22	-1.76
NM_005858	AKAP8	10270	0.556	0.007	0.825	0.95	-1.00	-2.00	1.08	1.36	1.26
NM_004328	BCS1L	617	0.668	0.007	0.434	1.02	-1.00	-2.00	1.10	1.41	1.28
AB020671	MPRIP	23164	0.506	0.007	0.259	-0.93	-1.00	0.50	-1.25	1.33	1.67
NM_021874			0.566	0.007	0.128	1.20	1.00	-0.50	1.27	2.13	1.67
NM_006453	TBL3	10607	0.763	0.007	0.769	1.07	-1.00	0.50	1.00	1.38	1.38
AL539691			0.642	0.007	0.716	0.98	-1.00	-2.00	-1.11	1.23	1.37
NM_001436	FBL	2091	0.143	0.007	0.803	1.05	-1.00	0.50	-1.06	1.36	1.44
Y12395	IFRD2	7866	0.937	0.007	0.678	0.97	-1.00	0.50	-1.01	1.33	1.34
CAA94614			0.092	0.008	0.262	-0.91	1.00	-0.50	1.21	-1.02	-1.23
NM_005923	MAP3K5	4217	0.561	0.008	0.682	-1.03	1.00	-0.50	-1.03	-1.40	-1.36

ACCN	Symbol	Gene ID	p-value for ventilation	p-value for oxygen	p-value for interaction of ventilation and oxygen	FC ventilation	FC oxygen	FC ventilation: oxygen	FC cluster 1 vs. 2	FC cluster 1 vs. 3	FC cluster 2 vs. 3
NM_002386	MC1R	4157	0.463	0.008	0.297	-1.06	1.00	-0.50	1.11	-1.35	-1.50
NM_013279	C11orf9	745	0.665	0.008	0.102	-1.04	-1.00	0.50	-1.13	1.09	1.24
AI984373			0.209	0.008	0.322	-0.92	1.00	-0.50	1.04	-1.05	-1.09
NM_014289	CAPN6	827	0.828	0.008	0.838	-1.06	1.00	-0.50	1.00	-1.27	-1.27
NM_018133	MSL2	55167	0.470	0.009	0.137	1.00	1.00	-0.50	1.07	-1.17	-1.25
NM_016543	SIGLEC7	27036	0.734	0.009	0.863	-1.02	1.00	2.00	-1.03	-1.18	-1.15
AF038440	PLD2	5338	0.808	0.009	0.145	0.98	1.00	-0.50	1.16	-1.25	-1.45
AB032967	ZNF473	25888	0.114	0.009	0.533	-0.96	1.00	2.00	-1.00	-1.31	-1.31
NM_018104			0.341	0.009	0.752	-1.00	1.00	-0.50	1.02	-1.36	-1.39
AB029012	SMG5	23381	0.366	0.010	0.462	-1.03	1.00	-0.50	1.15	-1.15	-1.32
NM_005451	PDLIM7	9260	0.832	0.010	0.942	-0.97	1.00	2.00	1.01	-1.40	-1.41
NM_004907	IER2	9592	0.848	0.010	0.172	-0.96	-1.00	0.50	-1.59	1.58	2.51
L29376	HCG26	352961	0.036	0.010	0.303	1.02	-1.00	0.50	-1.14	1.13	1.28
AF327354	WDR20	91833	0.377	0.010	0.067	0.99	1.00	-0.50	1.18	-1.06	-1.25
AL137489	C9orf123	90871	0.000	0.710	0.151	-0.93	1.00	-0.50	-1.30	1.36	1.78
NM_013308	GPR171	29909	0.001	0.265	0.623	-0.93	1.00	-0.50	-1.35	1.36	1.83
NM_019025			0.002	0.614	0.029	1.01	1.00	-0.50	-1.03	3.74	3.85
NM_004538	NAP1L3	4675	0.003	0.067	0.040	-1.03	1.00	2.00	-1.55	-1.31	1.19
NM_017980	LIMS2	55679	0.003	0.591	0.033	1.10	1.00	-0.50	1.44	1.19	-1.21
AB028960	WDTC1	23038	0.004	0.185	0.383	-0.98	-1.00	0.50	-1.09	1.05	1.15
NM_016940	RWDD2B	10069	0.004	0.356	0.703	-0.95	1.00	2.00	-1.10	1.14	1.25
NM_006145	DNAJB1	3337	0.004	0.532	0.213	-1.10	-1.00	0.50	-2.01	-1.02	1.96
AF007155	LPCAT4	254531	0.004	0.039	0.404	-1.10	-1.00	0.50	-1.13	-1.21	-1.06
BE866015			0.005	0.319	0.613	0.99	-1.00	-2.00	-1.01	1.05	1.06
NM_003984	SLC13A2	9058	0.006	0.751	0.939	-0.93	1.00	-0.50	-1.13	1.21	1.37
NM_017586	CACFD1	11094	0.006	0.737	0.967	-0.93	1.00	-0.50	-1.11	1.16	1.29
U79265	B3GNTL1	146712	0.006	0.277	0.057	1.01	1.00	-0.50	1.09	1.18	1.08
AK001228	UHRF1BP1	54887	0.006	0.018	0.720	-0.99	1.00	-0.50	-1.08	-1.06	1.02
NM_002574	PRDX1	5052	0.008	0.824	0.880	-0.81	1.00	-0.50	-1.10	3.61	3.96
AL050381	DNAJB12	54788	0.010	0.251	0.543	-0.93	1.00	-0.50	-1.03	1.23	1.26
NM_018089	ANKZF1	55139	0.010	0.401	0.179	-0.95	1.00	2.00	-1.20	1.05	1.27
NM_016211	SEC31A	22872	0.821	0.406	0.000	1.02	1.00	-0.50	2.29	1.63	-1.40
NM_015367	BCL2L13	23786	0.108	0.351	0.000	0.95	1.00	-0.50	4.01	1.38	-2.91
NM_006336	ZER1	10444	0.793	0.844	0.000	0.96	1.00	-0.50	2.26	1.51	-1.49
NM_001418	EIF4G2	1982	0.498	0.860	0.000	1.11	1.00	-0.50	2.28	1.81	-1.26
AL133623	XRN1	54464	0.769	0.159	0.000	-1.04	-1.00	0.50	-1.79	-1.46	1.23
NM_003217	TMBIM6	7009	0.268	0.277	0.000	0.96	1.00	-0.50	2.21	1.37	-1.61

ACCN	Symbol	Gene ID	p-value for ventilation	p-value for oxygen	p-value for interaction of ventilation and oxygen	FC ventilation	FC oxygen	FC ventilation: oxygen	FC cluster 1 vs. 2	FC cluster 1 vs. 3	FC cluster 2 vs. 3
NM_006304	SHFM1	7979	0.937	0.327	0.000	-1.00	-1.00	0.50	-2.04	-1.54	1.32
NM_000175	GPI	2821	0.710	0.994	0.000	0.98	1.00	-0.50	2.31	1.75	-1.32
NM_005778	RBM5	10181	0.169	0.971	0.000	0.96	1.00	-0.50	1.91	1.27	-1.50
NM_014390	SND1	27044	0.930	0.084	0.000	0.92	1.00	-0.50	2.09	1.18	-1.77
AB014530	HIPK1	204851	0.808	0.078	0.000	0.96	1.00	-0.50	2.04	1.10	-1.87
NM_006170	NOP2	4839	0.046	0.206	0.000	1.06	1.00	-0.50	1.71	1.50	-1.14
NM_014402	UQCRQ	27089	0.100	0.673	0.001	-1.01	-1.00	0.50	-1.69	-1.58	1.07
NM_001191	BCL2L1	598	0.664	0.695	0.001	0.98	1.00	-0.50	2.13	1.45	-1.47
NM_032204	ASCC2	84164	0.506	0.601	0.001	0.97	1.00	-0.50	3.21	1.46	-2.20
NM_007176	C14orf1	11161	0.360	0.250	0.001	-1.03	-1.00	0.50	-1.70	-1.36	1.26
NM_003746	DYNLL1	8655	0.578	0.888	0.001	-1.07	-1.00	0.50	-1.24	-1.23	1.01
AL046016	FAM46C	54855	0.762	0.481	0.001	0.95	1.00	-0.50	2.00	1.24	-1.61
NM_013291	CPSF1	29894	0.343	0.432	0.001	1.10	1.00	-0.50	2.19	1.50	-1.46
AB037788	CPSF2	53981	0.954	0.118	0.001	-1.01	1.00	2.00	-1.43	-1.40	1.02
NM_000274	OAT	4942	0.649	0.140	0.001	1.03	1.00	-0.50	3.11	1.37	-2.26
NM_004238	TRIP12	9320	0.737	0.251	0.002	1.03	1.00	-0.50	2.04	1.30	-1.57
NM_001923	DDB1	1642	0.731	0.176	0.002	0.98	1.00	-0.50	2.18	1.57	-1.39
NM_013236	ATXN10	25814	0.825	0.836	0.002	1.09	1.00	-0.50	2.18	1.64	-1.33
NM_012394	PFDN2	5202	0.548	0.760	0.002	-1.06	-1.00	0.50	-1.50	-1.40	1.07
NM_004879	EI24	9538	0.537	0.836	0.002	0.93	-1.00	-2.00	2.16	1.48	-1.46
NM_003639	IKBKG	8517	0.098	0.442	0.002	-1.05	-1.00	0.50	-1.40	-1.31	1.07
NM_032179			0.909	0.878	0.002	0.83	-1.00	-2.00	2.09	1.10	-1.90
NM_021078	KAT2A	2648	0.487	0.345	0.002	0.95	1.00	-0.50	1.71	1.16	-1.48
NM_033103	RHPN2	85415	0.269	0.205	0.003	1.00	1.00	-0.50	1.83	1.06	-1.73
NM_014761	IST1	9798	0.136	0.900	0.003	0.94	-1.00	-2.00	1.45	1.17	-1.23
NM_032305	POLR3GL	84265	0.596	0.046	0.003	-1.01	-1.00	0.50	-1.77	-1.13	1.57
NM_005001	NDUFA7	4701	0.398	0.633	0.003	-1.05	-1.00	0.50	-1.46	-1.39	1.05
AL137257	UHMK1	127933	0.354	0.335	0.003	0.99	1.00	-0.50	1.77	1.10	-1.60
NM_004120	GBP2	2634	0.380	0.482	0.003	1.09	1.00	-0.50	2.07	1.64	-1.26
NM_031902	MRPS5	64969	0.466	0.053	0.003	-1.02	-1.00	0.50	-1.51	-1.09	1.39
NM_001358	DHX15	1665	0.403	0.349	0.003	0.96	1.00	-0.50	1.92	1.30	-1.48
NM_004541	NDUFA1	4694	0.635	0.619	0.003	-1.03	-1.00	0.50	-1.66	-1.67	-1.00
NM_006402	HBXIP	10542	0.948	0.651	0.003	-1.09	-1.00	0.50	-1.39	-1.35	1.03
D50683			0.373	0.549	0.003	0.94	-1.00	-2.00	1.39	1.12	-1.24
AK026880	BRD3	8019	0.677	0.090	0.003	-0.98	1.00	2.00	-1.54	-1.35	1.14
NM_006711	RNPS1	10921	0.142	0.231	0.003	0.98	1.00	-0.50	1.64	1.15	-1.43
NM_002080	GOT2	2806	0.449	0.619	0.004	0.99	1.00	-0.50	1.90	1.41	-1.35

ACCN	Symbol	Gene ID	p-value for ventilation	p-value for oxygen	p-value for interaction of ventilation and oxygen	FC ventilation	FC oxygen	FC ventilation: oxygen	FC cluster 1 vs. 2	FC cluster 1 vs. 3	FC cluster 2 vs. 3
NM_018206	VPS35	55737	0.253	0.156	0.004	1.04	1.00	-0.50	1.62	1.14	-1.42
NM_014453	CHMP2A	27243	0.634	0.688	0.004	-1.03	-1.00	0.50	-1.52	-1.30	1.17
NM_014637	MTFR1	9650	0.446	0.485	0.004	-1.03	-1.00	0.50	-1.61	-1.51	1.07
NM_015710	GLTSCR2	29997	0.072	0.940	0.004	1.06	1.00	-0.50	1.66	1.43	-1.16
NM_004127	GPS1	2873	0.769	0.304	0.004	1.01	1.00	-0.50	2.28	1.54	-1.48
NM_017742	ZCCHC2	54877	0.932	0.753	0.005	-1.05	-1.00	0.50	-2.08	-1.61	1.30
NM_024552	CERS4	79603	0.334	0.082	0.005	-1.02	1.00	2.00	-1.71	-1.71	-1.00
NM_019082	DDX56	54606	0.075	0.449	0.005	1.04	1.00	-0.50	1.21	1.25	1.03
NM_004604	STX4	6810	0.810	0.111	0.005	-1.04	-1.00	0.50	-1.57	-1.13	1.40
U66042	CXorf40B	541578	0.177	0.199	0.005	-1.12	-1.00	0.50	-1.35	-1.40	-1.04
NM_025211	GKAP1	80318	0.088	0.854	0.005	-1.17	-1.00	0.50	-1.78	-1.69	1.05
NM_031844	HNRNPU	3192	0.816	0.985	0.005	0.95	-1.00	-2.00	2.05	1.53	-1.34
NM_001549	IFIT3	3437	0.475	0.336	0.005	-1.10	-1.00	0.50	-1.43	-1.25	1.15
NM_018196	TMLHE	55217	0.910	0.411	0.005	-0.93	1.00	2.00	-1.72	-1.19	1.45
NM_004046	ATP5A1	498	0.469	0.109	0.006	1.16	1.00	-0.50	2.04	2.50	1.23
BE560878	MAP2K3	5606	0.867	0.027	0.006	-1.08	-1.00	0.50	-1.43	1.01	1.44
BC004988	FEM1A	55527	0.522	0.345	0.006	0.98	1.00	-0.50	2.01	1.29	-1.56
AB029032	KIAA1109	84162	0.748	0.034	0.006	0.98	1.00	-0.50	2.40	1.19	-2.02
NM_014943	ZHX2	22882	0.562	0.066	0.006	1.02	1.00	-0.50	1.60	1.18	-1.35
NM_021943	ZFAND3	60685	0.211	0.177	0.006	0.98	1.00	-0.50	1.95	1.14	-1.71
NM_002873	RAD17	5884	0.814	0.104	0.006	1.09	1.00	-0.50	1.69	1.18	-1.43
NM_005341	ZBTB48	3104	0.212	0.409	0.006	1.02	1.00	-0.50	1.15	1.16	1.01
NM_003339	UBE2D2	7322	0.357	0.852	0.006	-1.04	-1.00	0.50	-1.42	-1.23	1.16
AF334405	SPRYD7	57213	0.867	0.967	0.006	-1.02	-1.00	0.50	-1.67	-1.36	1.24
U90878	PDLIM1	9124	0.175	0.796	0.006	1.06	1.00	-0.50	1.32	1.30	-1.02
NM_004515	ILF2	3608	0.105	0.973	0.006	1.03	1.00	-0.50	1.27	1.12	-1.14
NM_018307	RHOT1	55288	0.619	0.094	0.006	0.92	1.00	-0.50	1.71	-1.04	-1.78
NM_000359	TGM1	7051	0.179	0.720	0.007	-1.02	-1.00	0.50	-1.50	-1.21	1.24
NM_000527	LDLR	3949	0.827	0.384	0.007	-1.06	-1.00	0.50	-1.73	-1.70	1.02
NM_024009	GJB3	2707	0.770	0.120	0.007	0.97	1.00	-0.50	1.98	1.14	-1.75
NM_004794	RAB33A	9363	0.231	0.346	0.007	-1.04	-1.00	0.50	-1.60	-1.33	1.20
AL050277	ATP5L	10632	0.684	0.331	0.007	-1.04	-1.00	0.50	-1.45	-1.26	1.15
NM_019598	KLK12	43849	0.062	0.181	0.007	1.01	1.00	-0.50	1.79	1.47	-1.22
AA536084	MCM7	4176	0.727	0.201	0.007	0.98	1.00	-0.50	1.80	1.21	-1.49
NM_022831	AIDA	64853	0.182	0.313	0.007	0.97	1.00	-0.50	1.83	-1.02	-1.86
D85730	HSPA1L	3305	0.089	0.669	0.008	-1.14	-1.00	0.50	-1.90	-1.71	1.11
NM_003003	SEC14L1	6397	0.062	0.838	0.008	-1.01	-1.00	0.50	-1.73	-1.08	1.61

ACCN	Symbol	Gene ID	p-value for ventilation	p-value for oxygen	p-value for interaction of ventilation and oxygen	FC ventilation	FC oxygen	FC ventilation: oxygen	FC cluster 1 vs. 2	FC cluster 1 vs. 3	FC cluster 2 vs. 3
NM_024894	NOL10	79954	0.535	0.465	0.008	-0.90	1.00	2.00	-1.67	-1.46	1.15
NM_005291	GPR17	2840	0.495	0.025	0.008	1.00	1.00	-0.50	2.07	1.10	-1.89
NM_032102	SRSF8	10929	0.338	0.079	0.008	0.95	1.00	-0.50	1.87	-1.08	-2.01
NM_018320	RNF121	55298	0.754	0.027	0.008	-0.97	-1.00	0.50	-1.49	-1.03	1.45
NM_006405	TM9SF1	10548	0.343	0.022	0.008	-1.13	-1.00	0.50	-1.34	-1.18	1.13
NM_002947	RPA3	6119	0.850	0.121	0.008	-0.95	-1.00	0.50	-1.66	-1.02	1.63
AF014403	PPAP2A	8611	0.574	0.456	0.008	-0.93	-1.00	0.50	-1.85	-1.30	1.42
AB037861	INTS1	26173	0.898	0.027	0.008	1.01	1.00	-0.50	1.41	-1.01	-1.43
NM_001164	APBB1	322	0.377	0.621	0.008	1.11	1.00	-0.50	1.96	1.71	-1.15
NM_006022	TSC22D1	8848	0.706	0.706	0.008	-1.08	-1.00	0.50	-1.68	-1.57	1.07
AL133052	TMEM183A	92703	0.421	0.173	0.008	1.06	1.00	-0.50	2.09	1.24	-1.69
NM_005412	SHMT2	6472	0.035	0.084	0.009	1.03	-1.00	-2.00	1.64	1.46	-1.12
AB020861			0.503	0.429	0.009	-1.10	-1.00	0.50	-1.32	-1.40	-1.06
NM_016565	CHCHD8	51287	0.836	0.103	0.009	-1.03	-1.00	0.50	-1.51	-1.12	1.34
NM_014168	METTL5	29081	0.543	0.123	0.009	-1.09	-1.00	0.50	-1.72	-1.22	1.41
AK024426	EMR2	30817	0.123	0.173	0.009	-1.10	-1.00	0.50	-1.46	-1.64	-1.12
NM_020150	SAR1A	56681	0.772	0.734	0.009	-1.05	-1.00	0.50	-1.80	-1.48	1.21
NM_002686	PNMT	5409	0.393	0.228	0.009	1.03	1.00	-0.50	1.56	1.33	-1.18
NM_004712	HGS	9146	0.273	0.161	0.009	-1.00	1.00	2.00	-1.41	-1.24	1.14
NM_022898	BCL11B	64919	0.358	0.186	0.009	1.01	1.00	-0.50	1.32	1.16	-1.14
AK057343	ZNF131	7690	0.346	0.614	0.009	0.93	-1.00	-2.00	1.38	1.02	-1.36
NM_005006	NDUFS1	4719	0.793	0.837	0.009	-1.02	-1.00	0.50	-1.53	-1.36	1.12
NM_003139	SRPR	6734	0.176	0.097	0.009	0.97	1.00	-0.50	1.53	1.13	-1.36
NM_001896	CSNK2A2	1459	0.805	0.132	0.009	-1.06	-1.00	0.50	-1.45	-1.15	1.27
NM_014403			0.897	0.732	0.009	0.97	1.00	-0.50	2.23	1.45	-1.54
AB029016	TNRC6B	23112	0.845	0.627	0.009	0.94	-1.00	-2.00	1.63	1.17	-1.40
BC008861	ATP6V0D1	9114	0.712	0.629	0.010	1.05	1.00	-0.50	1.89	1.49	-1.26
NM_003953	MPZL1	9019	0.491	0.835	0.010	-1.04	-1.00	0.50	-1.75	-1.56	1.12

8.2 List of abbreviations

AIS	Amniotic infection syndrome
ANOVA	Analysis of variance
BLCI	Biomarker list concordance index
BPCA	Bayesian Principal Component Analysis
BPD	Bronchopulmonary dysplasia
cDNA	Complementary deoxyribonucleic acid
CPAP	Continuous positive airway pressure
CPP	Conserved pair proportions
CRP	C-reactive protein
DAVID	Database for Annotation, Visualization, and Integrated Discovery
EM	Expectation maximization
ES	Enrichment score
FC	Fold Change
FDR	False discovery rate
GA	Gestational age
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
IPA	Ingenuity Pathway Analysis
IQR	Interquartiles range
Iset	Invariant-set-normalization
IUGR	Intrauterine growth restriction
KEGG	Kyoto Encyclopedia for Genes and Genomes
KNN	K-nearest neighbor
LIMMA	Linear Models for Microarray analysis
LinImp	Linear model-based imputation
LLSI	Local least squares imputation
Loess	locally weighted regression
LS	Least squares
LSI	Least squares imputation
MAD	Mean absolute deviation

mRNA	Messenger ribonucleic acid
MV	Mechanical ventilation
NIPPV	Nasal intermittent positive pressure ventilation
OMIM	Online Mendelian Inheritance in Men
OPLSDA	Orthogonal partial least squares discriminant analysis
PAM	Predictive analysis of microarrays
PCA	Principal component analysis
PLS	Partial least squares
Qspline	Quantile-spline-normalization
RMA	Robust Multiarray Average
RMSE	Root mean squared error
ROS	Reactive oxygen species
RP	Rank Products
SAM	Significance analysis of microarrays
SD	Standard deviation
SE	Standard error
SeqKNN	sequential K-nearest neighbor
SNR	Signal-to-noise ratio
SOM	Self-organizing map
SSD	Sum of squared deviations
SVD	Singular value decomposition

8.3 List of figures

8.4 List of tables

8.5 List of equations

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und ohne unzulässige Hilfe oder Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder nichtveröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht. Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ niedergelegt sind, eingehalten sowie ethische, datenschutzrechtliche und tierschutzrechtliche Grundsätze befolgt. Ich versichere, dass Dritte von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen, oder habe diese nachstehend spezifiziert. Die vorgelegte Arbeit wurde weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde zum Zweck einer Promotion oder eines anderen Prüfungsverfahrens vorgelegt. Alles aus anderen Quellen und von anderen Personen übernommene Material, das in der Arbeit verwendet wurde oder auf das direkt Bezug genommen wird, wurde als solches kenntlich gemacht. Insbesondere wurden alle Personen genannt, die direkt und indirekt an der Entstehung der vorliegenden Arbeit beteiligt waren. Mit der Überprüfung meiner Arbeit durch eine Plagiatserkennungssoftware bzw. ein internetbasiertes Softwareprogramm erkläre ich mich einverstanden.

Ort, Datum

Unterschrift

Publikationsverzeichnis

Begutachtete Veröffentlichungen:

Xiao, Y., Karnati, S., Qian, G., Nenicu, A., Fan, W., Tchatalbachev, S., **Höland, A.**, Hossain, H., Guillou, F., Lüers, G.H., Baumgart-Vogt, E., 2012. Cre-Mediated Stress Affects Sirtuin Expression Levels, Peroxisome Biogenesis and Metabolism, Antioxidant and Proinflammatory Signaling Pathways. PLoS ONE 7, e41097. doi:10.1371/journal.pone.0041097

In Vorbereitung:

Anita Hoeland, Maren Schwarz, Tina Pritzke, Ludwig Gortner, Trinad Chakraborty, Hamid Hossain, Anne Hilgendorff. Discrimination of preterm infants at risk for the development of bronchopulmonary dysplasia at birth - differential expression pattern in transcriptome analysis. In preparation.

Anita Höland, Hamid Hossain, Anne Hilgendorff. Oxidative Stress and Inflammatory processes as modulators for BPD? In preparation.

Anne Hilgendorff, **Anita Höland**, Manuel Klein, Svetlin Tchatalbachev, Christine Windemuth-Kieselbach, Joachim Kreuder, Matthias Heckmann, Anna Gkatzofli, Harald Ehrhardt, Josef Mysliwietz, Micheal Maier, Benjamin Izar, Andre Billion, Ludwig Gortner, Trinad Chakraborty. Gene expression profiling from cord blood of preterm infants with early onset infection. In preparation.

Konferenzbeiträge:

Höland, A.C., Hossain, H., Chakraborty, T., 2010. Biostatistical analyses linking causative genotypes and transcriptional profiling in trauma-induced sepsis. Presented at the 3rd GGL Annual Conference 2010. URL: http://www.uni-giessen.de/cms/fbz/zentren/ggl/events/conference2010/dateienconf2010/poster3.

Best Poster award, 3rd place.

Höland A.C., Hossain H., Greene B., Lahl H., Daniel H., Schäfer H., Chakraborty, T., 2011. Biostatistical analysis linking causative genotypes and pathways involved in trauma-induced sepsis: Genome-Wide Association Study Pathway Analysis. Presented at the 4th GGL Annual Conference 2011.

Höland A.C., Hossain H., Chakraborty, T., 2012. Developing a biostatistical microarray analysis pipeline for reclassification using continuous phenotypes. Presented at the 5th GGL Annual Conference 2012.

Hoeland A, Gimm T, Hossain H, Ehrhardt H, Gortner L, Scholz M, Schwarz M, Reicherzer T, Hauck S, Hilgendorff A., 2012. Identification of biomarkers for early diagnosis and monitoring of bronchopulmonary dysplasia (BPD). Presented at the Munich Lung Conference; October 2012.

Danksagung

Hiermit danke ich Prof. Dr. Trinad Chakraborty und Dr. Hamid Hossain auf das herzlichste für die Bereitstellung eines Dissertationsthemas, eines freundlichen und produktiven Arbeitsumfeldes und die umfassende Betreuung während der Anfertigung der Dissertation.

Herzlich danken möchte ich auch allen Kollegen aus dem Institut für medizinische Mikrobiologie, die mich unterstützt haben und für eine rundum freundliche und offene Atmosphäre gesorgt haben. Besonders erwähnen möchte ich dabei Dr. Melanie Markmann, mit welcher die Zusammenarbeit in verschiedenen Projekten stets produktiv und lehrreich und voll fruchtbarer Diskussionen war.

Besonders zu erwähnen sind auch PD Dr. Anne Hilgendorff und Dr. Tina Pritzke aus dem Comprehensive Pneumology Center des Helmholtz-Zentrums in München, die den vorliegenden Datensatz zur Verfügung gestellt haben und einen nicht unerheblichen Teil bei der Betreuung der Arbeit geleistet haben. Die Zusammenarbeit war immer produktiv, effizient und ideenreich.

Ein besonderer Dank gilt Dr. Gabriel Schachtel, der mir meinen ersten Job als studentische Hilfskraft an der Uni gegeben hat und mich ermutigt hat, eine Promotion im Bereich Biostatistik zu wagen. Er stand mir im Laufe der Promotion immer mit einem offenen Ohr, Ratschlägen, Motivation und Schokolode zur Seite. Nicht vergessen möchte ich auch die Unterstützung von Dr. Birgit Samans und Dr. Jörn Pons-Kühnemann, die stets bereit waren meine Fragen zu verschiedensten Themen rund um Microarrays und R zu beantworten.

Eine herausragende Rolle haben Freunde und Familie gespielt, die mir stets mit Ablenkung, Motivation und Liebe zur Seite standen. Allen Voran stand David, ohne dessen Fürsorge, Liebe und Geduld hätte es nicht funktioniert.