UNIVERSITEIT
GENT

Faculty of Sciences

Comparison of methods for differential gene expression using RNA-seq data

Katrijn De Paepe

Master dissertation submitted to
obtain the degree of
Master of Statistical Data Analysis

Promoter: Prof. Dr. Olivier Thas
Co-promoter: Dr. Bie Verbist

Department of Mathematical Modelling, Statistics and Bioinformatics

Academic Year 2014-2015

Faculty of Sciences

Comparison of methods for differential gene expression using RNA-seq data

Katrijn De Paepe

Master dissertation submitted to
obtain the degree of
Master of Statistical Data Analysis

Promoter: Prof. Dr. Olivier Thas
Co-promoter: Dr. Bie Verbist

Department of Mathematical Modelling, Statistics and Bioinformatics

Academic Year 2014-2015

Katrijn De Paepe

Professor Olivier Thas

Foreword

This master thesis was written as my final project for the Master in Statistical Data Analysis. After having worked for four years in the consulting industry, I decided one year ago to take a leave of absence and to follow this master in order to reorient my career towards data and analytics. From that perspective, I was looking for a master thesis topic with practical relevance and preferably in cooperation with a company. Thanks to my promoter, professor Olivier Thas, and Janssen Pharmaceuticals I found such a topic. Although the beginning was a bit hard given my limited background in biology, familiarizing myself with the topic of RNA-seq and finally contributing to the research in this field turned out to be a rewarding experience. In the first place I would like to thank professor Thas for his support and advice. In addition, this master thesis would not have become what it is without the support of the Janssen team: Joke, An and in particular my co-promoter Bie. They provided me with the data and gave valuable input and feedback throughout the project. Thanks as well to Matthijs Vincke for reviewing the first part of my code and my cousin Bart for reviewing the introductory part on gene expression and RNA-sequencing.

Last but not least, I would like to thank my mentor Nicolas for encouraging me to pursue this master degree and my family and friends for their continued support throughout the past year.

Contents

List of abbreviations and symbols

List of abbreviations

AUC	area under the curve
BCV	biological coefficient of variation
cDNA	complementary DNA
cpm	counts per million
CR	Cox-Reid
DE	differentially expressed
DNA	deoxyribonucleic acid
ERCC	External RNA Control Consortium
FDR	false discovery rate
FPR	false positive rate
FWER	familywise error rate
glm	generalized linear model
GOF	goodness-of-fit
GTEx	Genotype-Tissue Expression project
HBRR	human brain reference RNA
LFC	logarithmic fold change
limmaQN	limma with quantile normalization
limmaVst	limma with variance stabilizing transformation
limmaVoomQW	limmaVoom with quality weights
mRNA	messenger RNA
MA-plot	plot of estimated log fold change (M) vs. average expression (A)
MDS	multidimensional scaling
MLE	maximum likelihood estimator
NB	negative binomial
ncRNA	non-coding RNA
non-DE	not differentially expressed
pDiff	fraction of differentially expressed genes

pOutlier	fraction of genes with outlying observations
pUp	fraction of differentially expressed genes that is upregulated
PCR	polymerase chain reaction
qCML	quantile-adjusted conditional maximum likelihood
qPCR	quantitative PCR
RMSE	root mean squared error
RNA	ribonucleic acid
ROC	receiver operating characteristic
SEQC	RNA sequencing quality control
TMM	trimmed mean of M-values
TPR	true positive rate
UHRR	universal human reference RNA

List of symbols

α_{gi}	the dispersion of the count of gene g in sample i
μ_{gi}	the expected value of the count of gene g in sample i
σ_{gi}^2	the variance of the count of gene g in sample i
d_i	sequencing depth of sample i (relative number of reads vs. other samples)
N_i	total number of reads in sample i
Y_{gi}	raw count of gene g in sample i
$Y_{gi}^{'}$	normalized count of gene g in sample i

Abstract

Background: High-throughput sequencing of cDNA (RNA-seq) is overtaking microarrays as the primary approach to transcriptome profiling. A fundamental task of RNA-seq is to identify genes that are differentially expressed between two or more conditions. Many statistical methods are available to perform this task, using different ways of normalizing the counts, modeling gene expression, testing for differential expression and dealing with outliers. However, no clear consensus exists on which of these methods perform best.

Methodology: We conduct a comparison of twelve methods for differential gene expression analysis of RNA-seq data. Our approach consists of two parts. First, we perform a concordance analysis in which we apply the methods on several real RNA-seq datasets to understand to which extent the methods come to the same results and which methods are more alike than others. Second, we conduct a simulation study to empirically assess the performance of the methods under varying conditions.

Conclusions: Whereas previous research states that no method is optimal under all circumstances, we claim that two methods have a clear advantage over the other methods: edgeR robust and limmaVoom (with and without sample quality weights). edgeR robust has the most favorable trade-off between power and real false discovery rate (FDR) in the presence of outliers, but is too liberal in the sense that the true FDR considerably exceeds the target FDR. limmaVoom also achieves a good trade-off between power and real FDR and performs best in terms of FDR control.

Introduction

Phenotypic variations are known to be largely characterized by distinct patterns of gene expression. For a long time microarrays were the primary technology to measure gene expression levels. Recently RNA-sequencing (RNA-seq) has become a competitive alternative to the microarray technology. Unlike microarrays, RNA-seq expression profiles consist of counts, reflecting the number of sequence reads mapped to each gene. Many different methods have been developed to identify differentially expressed genes from these RNA-seq data, but no clear consensus exists on which of these methods perform best. In this thesis we compare the most frequently used methods for differential gene expression analysis of RNA-seq data to ultimately come to a recommendation on which methods are best to use.

Chapter 1 of this report consists of a literature study. In Section 1.1 we summarize the biological processes behind gene expression and introduce RNA-seq as an alternative to microarrays for gene expression profiling. Subsequently, we explain the different components of differential expression analysis, pointing out the challenges in each of the steps (Section 1.2). Afterwards we discuss how the most frequently used methods concretely fill in each of these components (Section 1.3). In Section 1.4 we review the existing literature that compares and evaluates the different methods.

In Chapter 2 we explain the methodology of our own research. After defining the scope and the overall approach of our research (Section 2.1), we give a detailed description of the different datasets used (Section 2.2). It will become clear that the different datasets reflect the variety of settings in which RNA-seq is used. In Sections 2.3 and 2.4, we go into more detail on the two pillars of our research. For the concordance analysis, in which we apply the methods on several real RNA-seq datasets, we describe our choice for the parameter settings of the methods and we explain the analyses we will perform to assess similarities and dissimilarities between the methods. While the concordance analysis allows to make statements on which methods are more alike than others, it does not allow to make statements on which methods perform best. To this extent we perform a simulation study. We give a detailed explanation of the simulation setup and we define the performance metrics to be used.

Chapter 3 discusses the most insightful results of both the concordance analysis and the simulation study. Where possible we link our results to the literature we discussed in chapter 1. Based on the results we will motivate that either limmaVoom (with and without quality weights) or edgeR robust is the best methods to use, depending on whether power of FDR control is of primary interest to the researcher. We end with some suggestions for further research.

Chapter 1

Literature

1.1 Background

1.1.1 Gene expression

DNA, which stands for deoxyribonucleic acid, is a double stranded molecule with a helical structure and is present in the nucleus of all living cells. Structurally, each strand of the DNA is a polymer, which consists of monomers called nucleotides. Each nucleotide is composed of a nitrogenous base, a five-carbon sugar and a phosphate group. There are four types of nucleotides, differing only in the nitrogenous base: adenine (A), guanine (G), cytosine (C) and thymine (T). These bases are in the inside of the helix, forming the steps of the spiral staircase. Each base on one strand pairs with just one type of base on the other strand: adenine pairs to thymine in two hydrogen bonds and cytosine pairs to guanine in three hydrogen bonds. This phenomenon, called base pairing, implies that the sequence of bases on one strand uniquely determines the entire set of base pairs. The order of these bases encodes all biological information required to make proteins. The process of protein synthesis from DNA consists of two steps and is referred to as the central dogma of molecular biology. In a first step, the **transcription** step, the information contained in a section of DNA is replicated in the form of a newly assembled piece of messenger **RNA** (mRNA). In the second step, the **translation** step, mRNA finds its way to a ribosome where it is used to make **proteins**. It is exactly these proteins that regulate all biological processes of the organism.

Not all nucleotides in the DNA encode a function. The subunits of DNA that encode a function are called genes. A **gene** encodes either a **protein product** or a **functional RNA product**. Whereas protein-coding genes go through both steps as described in the central dogma of molecular biology, RNA-coding genes only go through the first step and are transcribed in non-coding RNA (ncRNA) but are not translated into proteins. The process of producing a biologically functional molecule in the form of either RNA or a protein is called **gene expression**. Measuring RNA concentration levels is a useful tool in determining how the transcriptional machinery of the cell is affected in the presence of external signals (e.g. a drug treatment), or how cells differ between a healthy state and a diseased state. Genes that show differences in expression level between two conditions, are called differentially expressed (DE) genes.

1.1.2 RNA-sequencing

For most of the past two decades, DNA **microarrays** were the predominant technology for expression profiling. The microarray technology is based on the principle of hybridization between complementary base pairs. Expression levels are assessed by measuring the intensity of the light that is emitted by the fluorescently labeled target cDNA that binds with the probes on the microarray chip. In the past few years, **RNA-sequencing** has emerged as a new revolutionary tool for transcriptomics (Wang et al., 2009). RNA-seq is based on the principle of high-throughput sequencing, a catch-all term for a set of of technologies that allow to determine the order of bases in DNA or RNA quickly and cheaply. Unlike microarrays, RNA-seq profiles consist of integer counts, reflecting the number of sequence reads mapped to each genomic feature of interest. Compared to microarrays, RNA-seq offers a number of **advantages**: it is not limited to detecting transcripts that correspond to existing genomic sequences, it has a much larger dynamic range of expression levels over which transcripts can be detected and it requires less RNA sample. While RNA-seq is used for a variety of applications, including the detection of alternative splice forms and novel transcript discovery, it is most commonly used for detecting differential expression between experimental conditions, which is also the focus of this master thesis.

Figure 1.1 illustrates the **workflow of a typical RNA-seq experiment**. The RNA from a sample is fragmented into small pieces, which are then reverse transcribed into more stable complementary DNA (cDNA) using random primers. These short pieces of cDNA are amplified by polymerase chain reaction (PCR) and sequenced by a sequencing machine, resulting in millions of short sequence read-outs, called "reads". Subsequently the reads are mapped to a reference genome (the case of de novo assembly is not illustrated here), using an algorithm that tells which region each read comes from. By counting the number of reads for a set of genomic features, we obtain a measure of the expression of these genomic features. While these genomic features could be genes, exons or junctions, we will work at gene level in the remainder of this thesis. The result of sequencing a single sample is thus a vector of feature counts. The feature counts of all samples are put together in a **count table** with the features as rows and the samples as columns. This table is the starting point for differential expression analysis, which is the topic of this master thesis.

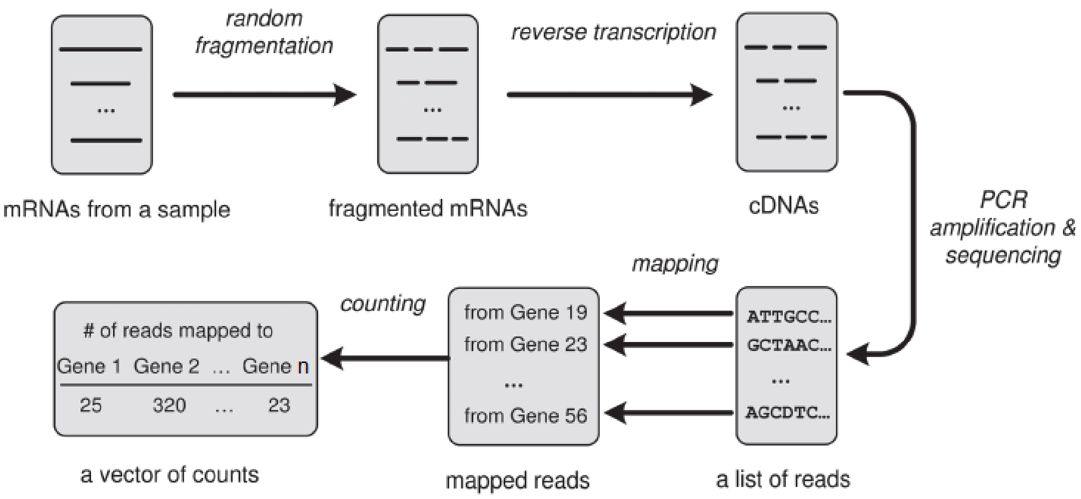


Figure 1.1: Workflow for a typical RNA-seq experiment, modified from Figure 1 of Li et al. (2012)

1.2 Components of DE methods

Differential gene expression analysis of RNA-seq data generally consists of three components: normalization of the counts, statistical modeling of gene expression and testing for differential expression (Rapaport et al., 2013). As recent publications (Zhou et al., 2014; Love et al., 2014) emphasize the influence of outliers and the need to deal with them in an appropriate way, we treat the topic of outliers in a separate subsection. It may be clear though that outlier correction is intertwined with the statistical modeling of gene expression and testing for differential expression. In Sections 1.2.1-1.2.4, we give a general discussion of the challenges in each of these steps. For each step, we will already briefly indicate the general approach used by the methods we discuss in full depth in Section 1.3: edgeR classic (Robinson et al., 2010), edgeR glm (McCarthy et al., 2012), edgeR robust (Zhou et al., 2014), DESeq (Anders and Huber, 2010), DESeq2 (Love et al., 2014), limmaQN (Rapaport et al., 2013), limmaVoom (Law et al., 2014), limmaVoom with Quality Weights (Liu et al., 2015), limmaVst (Soneson and Delorenzi, 2013), baySeq (Hardcastle and Kelly, 2010), PoissonSeq (Li et al., 2012) and SAMSeq (Li and Tibshirani, 2013).

1.2.1 Normalizing the counts

While originally it was claimed that one advantage of RNA-seq is that it does not require sophisticated normalization of the data sets (Wang et al., 2009), the current view is that normalization is an essential step in the analysis of RNA-seq data (Dillies et al., 2013). The aim of normalization is to **remove systematic technical effects** that occur in the data to make samples comparable and ensure that technical bias has minimal impact on the results.

The most obvious source of variation is the **large differences in the total number of reads** between samples. Denote Y_{gi} as the raw count of gene g ($g = 1, \cdots, n$) in sample i ($i = 1, \cdots, m$). Then the total number of reads, also called the library size, of sample i is given by $N_i = \sum_{g=1}^{n} Y_{gi}$. The sequencing depth d_i of a sample i is a measure of the relative number of reads in sample i versus the other samples. The raw counts of each sample should be normalized though division by the sample's sequencing depth. However, accurate estimation of the sequencing depth is not that trivial. The original and most straightforward approach, referred to as **total count normalization**, is to calculate the sequencing depth d_i of each sample i as the ratio of the library size of sample i versus the average library size.

$$d_i = \frac{N_i}{1/m \sum_{i=1}^{m} N_i}$$

such that the normalized counts Y'_{gi} are given by

$$Y'_{gi} = Y_{gi}/d_i$$

However, such a normalization is **generally not enough** (Robinson et al., 2010). Even if library sizes are equal, RNA-seq counts inherently represent relative abundances of the genes. If some genes have extremely high expression in one sample, or a large number of genes are only expressed in one sample, these genes may repress the counts for all other genes. The latter group of genes may, perhaps incorrectly, seem to have lower expression compared to a sample where the counts are distributed more evenly and this may lead to a lot of false positives. To account for this, **more complex normalization schemes** have been proposed: TMM normalization (edgeR and limmaVoom with and without quality weights), median-of-ratios normalization (DESeq, DESeq2, limmaVst), upper-quartile normalization (baySeq), quantile normalization (limmaQN), normalization by total count of least differential genes (PoissonSeq) and Poisson sampling normalization (SAMSeq). Often these normalization schemes start from the assumption that most genes are not differentially expressed and that the set of DE genes is less or more equally divided between upregulated and downregulated genes. Each of these normalizations is discussed in more detail in Section 1.3 when we discuss the respective DE-methods they are used for. All of them try to achieve comparability between samples. Other approaches exist that also aim to facilitate comparison of expression levels between genes within a sample. These approaches rescale gene counts to also correct for gene length and GC-content. We will not further discuss these latter approaches.

1.2.2 Statistical modeling of gene expression

Whereas a number of early RNA-seq publications applied statistical methods developed for microarrays to analyze RNA-seq data (Cloonan et al., 2008; Perkins et al., 2009), later publications argued that these microarray methods are not applicable to RNA-seq data, as these data are **discrete counts** rather than continuous measurements.

A natural representation of gene read counts would be the **Poisson distribution**. The PoissonSeq method (Li et al., 2012) amongst others uses the Poisson distribution to model the counts. An important property of the Poisson distribution is that the variance equals the mean. Marioni et al. (2008) reported that count data from technical replicates are indeed well characterized by a Poisson distribution. Technical replicates are samples that share the same underlying RNA-sample, but that have been sequenced in different runs. Biological replicates on the other hand share the same condition but originate from different RNA-samples taken from different cell lines, organisms etc. It turns out that for biological replicates the variance is larger than the mean and the **negative binomial (NB) distribution**, also called the overdispersed Poission model, is more appropriate (Robinson and Smyth, 2008). The simulation studies of Lu et al. (2005) show that the NB assumption can be reliable in non-NB sampling situations as well and as such provides a more flexible framework for real data. DE-methods like edgeR (Robinson et al., 2010), DESeq (Anders and Huber, 2010), DESeq2 (Love et al., 2014) and baySeq (Hardcastle and Kelly, 2010) model the counts by means of a NB distribution. The negative binomial distribution is uniquely determined by the mean μ and the variance σ^2. For the count Y_{gi} of gene g in sample i, we thus have that $Y_{gi} \sim \text{NB}(\mu_{gi}, \sigma^2_{gi})$. The relationship between σ^2_{gi} and μ_{gi} is defined as $\sigma^2_{gi} = \mu_{gi} + \alpha_{gi}\mu^2_{gi}$, where α_{gi} is called the dispersion. Most methods assume that within a condition, a gene's dispersion is the same across replicates, such that α_{gi} can be replaced with α_g in the aforementioned formula. Even then, **estimation of the dispersions** remains a challenge, as in most cases there is only a **limited number of samples** (experimental designs with two to three replicates per condition are common) resulting in highly variable dispersion estimates for each gene. Using these noisy estimates directly would compromise the accuracy of DE testing. Therefore, DE methods use some sharing of information across genes to come to more reliable dispersion estimates. The way how this is done varies from method to method as further discussed in Section 1.3.

The limma-based methods use a completely different approach: Law et al. (2014) revisits the idea of applying **normal-based statistical methods** developed for microarrays on RNA-seq data. They start from the observation that statistical methods developed specifically for RNA-seq counts rely on approximations of various kinds. They suggest that it might be more important to correctly model the mean-variance relationship than to specify the exact probabilistic distribution of the counts. Yet other methods, of which SAMSeq (Li and Tibshirani, 2013) is the most well-known, get around the difficulty of modelling counts by using a **non-parametric approach**.

1.2.3 Testing for differential expression

Once the statistical model of gene expression has been defined, **a test for differential expression** is performed to determine which genes show evidence for differences in expression level between experimental groups, taking into account technical and biological variation in these expression levels. While edgeR classic and DESeq use an exact test, edgeR glm and edgeR robust use a likelihood ratio test, DESeq2 a Wald test and the limma-based methods a moderated t-test. PoissonSeq performs a score test for the significance of the term that models differential expression in the log-linear model it estimated. SAM-seq calculates a kind of Wilcoxon rank-sum test of which the p-value is calculated by means of a permutation approach. baySeq estimates the posterior probabilities of two models, one reflecting differential expression and the other no differential expression. It does not generate p-values but the posterior likelihoods can be used for hypothesis testing.

Since thousands of tests are performed (one corresponding to each gene), a **correction for multiple testing** needs to be included to avoid an uncontrolled inflation of type I-errors. All methods that generate p-values correct for multiple testing by applying the Benjamini-Hochberg FDR procedure (Benjamini and Hochberg, 1995), except for PoissonSeq and SAM-Seq that propose a different approach based on permutation plug-in (Tusher et al., 2001). In Bayesian statistics no correction for multiple testing is required. As such, the posterior probability of differential expression as returned by baySeq can be used directly for testing.

1.2.4 Correcting for outliers

We say that a (normalized) count Y'_{gi} of gene g in replicate i is an outlier if its value is markedly outside the range of (normalized) counts of gene g in the other replicates. For example, if gene g has single-digit counts for all samples belonging to a specific condition, except for one sample where the count of gene g is in the thousands, we say this latter count value is an outlier or that gene g is an outlier gene. Li and Tibshirani show that **parametric methods are very sensitive to outliers**: the log fold change (LFC) is overly influenced by individual outliers, leading to a too high true false discovery rate (FDR). There are several possible reasons for outliers. A gene may be highly expressed in one individual, but not in others. In this case expression is related to the individual, not to the condition. Mapping errors as well can produce outliers.

While older parametric DE-methods did not pay a lot of attention to outliers, newer DE-methods specifically address this issue. DESeq2 removes genes with outlying observations or imputes a trimmed value for them, edgeR robust downweights outlying genes. Non-parametric methods like SAMSeq don't make distributional assumptions and are by nature less sensitive to outliers.

1.3 Overview selected methods

Many different methods are available for DE analysis of RNA-seq data. Xiong et al. (2014) states that edgeR, DESeq, limma-based methods, baySeq, PoissonSeq and Cuffdiff are among the most widely used tools for DE analysis. Besides two exceptions, we limit the scope of this thesis to these methods, both for the detailed discussion of DE methods in this section as for the set of methods that we include in our own research. A first exception is that Cuffdiff is not included as it is not available in R. A second exception is that we add SAMSeq as we also want to include a non-parametric method. An overview of the overall approach of the selected methods is shown in Table 1.1, a more detailed description is given below.

1.3.1 edgeR classic, edgeR glm, edgeR robust

The edgeR package (Robinson et al., 2010) uses **TMM normalization**, which stands for the trimmed mean of M-values normalization, as the default normalization method as proposed by Robinson and Oshlack (2010). TMM normalization is based on the hypothesis that most genes are not differentially expressed. One sample r is picked as the reference sample, the other samples are the test samples. For each test sample i a TMM factor is computed based on the genes' log expression ratios M_{gi}^{r} (the M-values):

$$M_{gi}^{r} = \log_2 \frac{Y_{gi}/N_i}{Y_{gr}/N_r}$$

with

M_{gi}^{r} the log expression ratio of test sample i versus reference sample r for gene g

Y_{gi} and Y_{gr} the raw counts of gene g in test sample i and reference sample r respectively

N_i and N_r the library sizes of test sample i and reference sample r respectively

The TMM for test sample i is then the weighted average of these M-values, after excluding the genes with the most extreme M-values and the genes with the highest absolute expression levels. The weights are determined as the inverse of the approximate asymptotic variance and account for the fact that log fold changes from genes with larger read counts have lower variance on the logarithmic scale. Assuming that most genes are not differentially expressed, the TMM should be close to 1. If it is not, its value provides an estimate of the correction factor that must be applied to the library sizes (not the raw counts) in the statistical analysis. The correction factor for the reference sample is 1.

edgeR models the counts by means of a **NB model**. It estimates a common or trended dispersion for all tags and then applies an **empirical Bayes** strategy for squeezing the tagwise dispersions toward the common or trended dispersion (Robinson and Smyth, 2007). The

Method	Publication	Normalization	Statistical modelling gene expression	Test for differential expression	Correction for multiple testing	Outlier Correction
edgeR classic	Robinson et al. (2010)	TMM	Assume NB distribution; Use of empirical Bayes for squeezing the tagwise dispersions toward common or trended dispersion	Exact test	Benjamini Hochberg	None
edgeR glm	McCarthy et al. (2012)			Likelihood ratio test		None
edgeR robust	Zhou et al. (2014)			Likelihood ratio test		Downweight observations with high Pearson residual
DESeq	Anders and Huber (2010)	Median-of-ratios	Assume NB distribution; Dispersion of each gene is maximum of gene-wise dispersion estimate and fitted value (that has been estimated with a parametric fit)	Exact test	Benjamini Hochberg	None
DESeq2	Love et al. (2014)		Assume NB distribution; Use empirical Bayes to shrink gene-wise dispersion estimates towards fitted values (that have been estimated with a parametric fit)	Empirical Bayes to shrink LFC + Wald Test		Detect with Cook's Distance; Exclude from further analysis or impute values
limmaQN	Rapaport et al. (2013)	Quantile Normalization on log transformed counts	Assume Gaussian distribution for normalized log-transformed counts	Fit linear model (with weights if provided); Apply moderated t-test	Benjamini Hochberg	Robust empirical Bayes options of limma package could be used
limmaVoom	Law et al. (2014)	TMM	Assume Gaussian distribution after log-transforming normalized counts; Calculate observation weights based on remaining mean-variance relationship			
limmaVoom QW	Liu et al. (2015)		Assume Gaussian distribution after log-transforming normalized counts; Calculate observation weights based on remaining mean-variance relationship and combine with sample weights			
limmaVst	Soneson and Delorenzi (2013)	Median-of-ratios	Assume Gaussian distribution after applying the DESeq variance stabilizing transformation on the normalized counts			
baySeq	Hardcastle and Kelly (2010)	Upper Quartile *(TMM*[1]*)*	Assume NB distribution; Estimate prior distribution of NB parameters by means of bootstrapping from the data and calculate posterior probabilities	Use posterior distribution of DE model	None	None
PoissonSeq	Li et al. (2012)	Total count of least differental genes (assessed by GOF)	Assume Poisson distribution; Take power transformation of the data first if needed)	Score statistic	Permutation plug-in	None
SAMSeq	Li and Tibshirani (2013)	Poisson Sampling	None	Wilcoxon rank-sum statistic	Permutation plug-in	None

Table 1.1: Comparison selected DE-methods

[1] while Upper Quartile normalization is the default normalization procedure for baySeq, we will use TMM as in Soneson and Delorenzi (2013)

amount of shrinkage is determined by the prior weight given to the common or trended dispersion and the precision of the tagwise estimates. As such edgeR allows the estimation of gene-specific biological variation, even for experiments with minimal levels of biological replication. edgeR classic estimates the dispersions using quantile-adjusted conditional maximum likelihood (qCML), conditioning on the total count of the particular gene (Robinson and Smyth, 2008). edgeR classic can only be used for designs with a single factor. edgeR glm can also be used for multifactor designs and fits generalized linear models given a count matrix and a design matrix (McCarthy et al., 2012). edgeR glm uses the Cox-Reid profile-adjusted likelihood (CR) to estimate the tagwise dispersions. Robust edgeR modifies the edgeR glm approach by attaching weights to each observation (Zhou et al., 2014). Observations with a high Pearson residual in the current fit are given lower weight in the next fit. As such the influence of outliers is dampened.

To test for DE genes, edgeR **classic** uses an **exact test** based on quantile-adjusted pseudocounts that are generated by the qCML approach. Let Z_{gA} and Z_{gB} denote the sum of pseudocounts for condition A and condition B respectively. Under the null hypothesis, both Z_{gA} and Z_{gB} follow a NB distribution. An exact test similar to the Fisher's exact test can then be constructed. Conditioning on $Z_{gA} + Z_{gB}$, also an NB random variable, the probability of observing class totals at least as extreme as the ones observed can be calculated. The 2-sided p-value is defined as the sum of probabilities of condition totals that are not more likely than those observed. edgeR **glm** tests for DE genes using the GLM **likelihood ratio test**. This test is based on the idea of fitting negative binomial GLMs with the CR dispersion estimates. edgeR **robust** follows the same approach but again incorporates the **weights** when fitting the GLM model and performing the **likelihood ratio test**. For all three edgeR variants, the p-values generated are then corrected for multiple testing by means of the **Benjamini-Hochberg** procedure.

1.3.2 DESeq and DESeq2

DESeq (Anders and Huber, 2010) and DESeq2 (Love et al., 2014) use a **'median-of-ratios'** approach to normalize counts. Similar to TMM, this approach assumes that most genes are not DE. For each sample a scaling factor is calculated as the median of the ratios of each gene's read count in the particular sample over its geometric mean across all samples. The underlying idea is that non-DE genes should have similar read counts across samples leading to a ratio of 1. If most genes are non-DE, the median of this ratio for each sample is the estimated correction factor to be applied to all counts of this sample.

The counts Y_{gi} are modeled by a **NB distribution**. The relationship between the variance and the mean is modeled by of means of a gene-specific dispersion parameter. The procedure to estimate the dispersions consists of three steps. First, a dispersion value is estimated for each gene using maximum likelihood. Second, a curve is fitted through the estimates. While the original version of DESeq used a local regression, the current default for DESeq and DESeq2 is to use a parametric fit. In a third and last step a dispersion value is assigned to each gene, choosing a value between the gene-wise estimate and the fitted value. **DESeq** adopts a rather conservative approach using the **maximum of the fitted value and the gene-wise estimate**. **DESeq2** shrinks the gene-wise dispersion estimates towards the fitted values to obtain the final dispersion values. To this extent it uses an **empirical Bayes** approach, which lets the strength of shrinkage depend on an estimate of how close the true dispersion values tend to be to the fit and on the degrees of freedom.

The **DESeq** approach to test for differential expression is very similar to the one of edgeR classic, using an **exact test** for differences between two negative binomial variables. **DESeq2** on the other hand adopts a different approach. It first **shrinks** the MLE-estimates for the **LFC** towards zero in a way such that shrinkage is stronger when the available information for a gene is low. Therefore it employs an empirical Bayes procedure. These shrunken LFCs together with their standard errors are then used in a **Wald test** for differential expression. The p-values of both the DESeq and the DESeq2 approach are then adjusted for multiple testing using the procedure of **Benjamini and Hochberg**.

To avoid the gene-wise LFC estimates of being overly influenced by individual outliers, DESeq2 adopts an approach to detect **outliers** and reduce their impact. Outliers are detected using the Cook's distance. By default, outliers in conditions with six or less replicates cause the whole gene to be removed from subsequent analysis. For conditions that contain seven or more replicates, DESeq2 replaces the outlier counts with an imputed value. One final remark on DESeq2 is that it applies an automatic filtering. By default, DESeq2 chooses an average expression cutoff that maximizes the number of genes found at user-specified target FDR.

1.3.3 limma-based methods

Limma-based methods **transform the count** data before entering them into the **limma pipeline**, a toolkit with statistical methods to perform differential expression analysis on microarray data. We will first describe the approach behind limmaVoom (Law et al., 2014) and limmaVoom with quality weights (Liu et al., 2015). At the end of the section we briefly outline two alternative limma-based methods: limmaQN (Rapaport et al., 2013) and limmaVst (Soneson and Delorenzi, 2013).

limmaVoom (with or without quality weights) starts with a normalization of the count data. While Law et al. proposed a straightforward counts per million normalization, the current standard is to use **TMM**. An obstacle to use these normalized counts in normal-based statistical methods is that they have unequal variabilities: larger counts have larger variance than smaller counts. Taking a $\log_2$-transformation of the normalized counts counteracts this, but it even overdoes the adjustment a bit: while the variance is roughly stable for larger log-transformed normalized counts, it shows a smoothly decreasing trend for the small to medium log-transformed normalized counts.

As a consequence, the **mean-variance relationship** needs to be estimated before feeding the log-transformed normalized counts to the limma pipeline. limmaVoom models the mean-variance trend of the log-transformed normalized counts **at the individual observation level**, rather than applying a gene-level variability estimate to all observations from the same gene. The trend is estimated in a non-parametric way. A difficulty here is that there is no replication at observational level from which variances could be estimated. To work around this, the mean-variance trend is estimated at gene level and this trend is then interpolated to predict the variances of individual observations. The inverse variances are used as weights in the rest of the procedure to eliminate the mean-variance relationship in the log-transformed counts. limmaVoom with **quality weights** augments this procedure by combining these observation weights with sample weights that account for variations in sample quality.

In a final stage the log-transformed normalized counts and their associated weights are passed to the usual limma pipeline for differential expression. A linear model is fit and an empirical Bayes procedure is applied to test for differential expression by means of a an **empirical Bayes moderated t-test** in which both the standard error and the degrees of freedom are modified. Empirical Bayes smoothing is applied to the standard errors, borrowing information from all genes. The degrees of freedom are also adjusted by a term that represents the a priori number of degrees of freedom for the model. The p-values produced are corrected for multiple testing by applying the **Benjamini and Hochberg procedure**.

Two alternative limma-based methods that have been proposed are limmaQN and limmaVst. **limma QN** performs a **quantile normalization** on the log-transformed counts. Quantile normalization ensures that the counts across all samples have the same empirical distribution by sorting the counts from each sample and setting the values to be equal to the quantile mean from all samples. **limmaVst** applies the **variance stabilizing transformation** provided by the DESeq package. Both methods feed the transformed counts to the limma-pipeline without passing any quality weights.

Limma-based methods are said to inherit the robustness properties from the normal-based procedures in limma and can be made even more robust using the robust empirical Bayes options of the limma package.

1.3.4 baySeq

The default normalization procedure in baySeq (Hardcastle and Kelly, 2010) is an **upper-quartile normalization** as proposed by Bullard et al. (2010). Upper-quartile normalization implies that for each sample the non-zero gene counts are summed up to the upper 25% quantile. Subsequently the counts of each sample are divided by the upper-quartile value of the sample and multiplied by the average upper-quartile value across samples. We will however use a TMM-normalization, similar to Soneson and Delorenzi.

baySeq uses a completely different inferential approach compared to the methods discussed above. baySeq requires the user to define a **set of models**. Each model divides the samples into groups, where samples in the same group are assumed to share the same parameters of the underlying distribution. Imagine the situation where we have two experimental conditions A and B and that for each experimental condition we have 2 samples, respectively A_1, A_2 and B_1, B_2. A first model of no differential expression is then defined by the set of samples $\{A_1, A_2, B_1, B_2\}$ which all share the same parameters for the underlying distribution . A second model of differential expression between the two conditions, divides the samples in two groups, namely $\{A_1, A_2\}$ and $\{B_1, B_2\}$ where each group has its own set of parameters. Subsequently baySeq uses an empirical Bayes framework to **estimate the posterior probability of each model** for each gene. To this extent baySeq assumes that the counts follow an **NB distribution**. The prior distribution of the parameters of the NB model is estimated by bootstrapping from the data, taking individual counts and finding the quasi-likelihood parameters.

1.3.5 PoissonSeq

PoissonSeq uses a log-linear model with a different approach to normalization and a novel procedure for estimating the false discovery rate (Li et al., 2012). The underlying assumption of this approach is that the counts follow a Poisson distribution. Potential overdispersion in the data is handled by taking a power transformation.

The **log-linear model** is estimated in two steps. The first step fits a model under the null hypothesis that no gene is associated with the outcome. In this step the **normalization** factors are estimated by an iterative procedure. A goodness-of-fit (GOF) statistic is used to estimate which set of **genes** is **least differential** between two (or more) conditions. The normalization factor for a sample i is then calculated by comparing the average total count for this subset of genes across all samples versus the total count for this subset of genes in sample i.

In the second step, an additional term is added to the model to accommodate differential expression. The main interest here lies in determining whether the parameter estimate related to this additional term is different from zero. Therefore a **score statistic** is used. The authors showed that when the Poisson log-linear model holds exactly and under the null hypothesis of no differential expression, the empirical sampling distribution of this score statistic closely follows the chi-squared law. Instead of using the FDR-procedure of Benjamini Hochberg, the authors use a modified version of the **permutation plug-in estimate for FDR**. The modification exists in that only the genes whose observed score is small are used in the pooled permutation distribution that is used **to estimate the FDR**.

1.3.6 SAMSeq

SAMSeq (Li and Tibshirani, 2013) is the only nonparametric method we discuss in more detail. The authors claim that, by not relying on underlying distributional assumptions, their method is less sensitive to outliers and allows to better detect consistent patterns in differential expression.

Normalization of the counts is done by using a resampling strategy. First the sequencing depths of the samples $d_1, d_2, ..., d_m$ are estimated. One could think of several methods to do this, but Li and Tibshirani use the approach of PoissonSeq. Instead of just scaling each count by the sequencing depth of the sample to which it belongs, a **Poisson sampling** strategy is applied. The geometric mean $\bar{d}$ of the sequencing depths is determined. For each sample i, a normalized count $Y_{gi}^{'}$ is resampled using

$$Y_{gi}^{'} \sim \text{Poisson}(\frac{\bar{d}}{d_i} N_{gi})$$

In order to avoid ties between the normalized counts, a small random number is added to each count.

Once the counts are on a comparable scale, a **Wilcoxon rank-sum statistic** is calculated for each gene to test for a difference in ranks between two conditions. As some randomness is introduced by the resampling procedure, the resampling procedure is repeated a number of times and for each gene the average is taken of the corresponding Wilcoxon statistics. Since the distribution of the averaged Wilcoxon statistic is unknown, a **permutation plug-in estimate** is used to generate the null distribution and to estimate the FDR.

1.3.7 Other methods

Many more DE-methods are available. DEGSeq (Wang et al., 2010), TSPM (Auer and Doerge, 2011), NBPSeq (Di et al., 2011), NOISeq (Tarazona et al., 2011), EBSeq (Leng et al., 2013), DSS (Wu et al., 2013) and ShrinkBayes (Van De Wiel et al., 2013) are only a few of them. Cuffdiff (Trapnell et al., 2013) is yet another method, but this one is not available in R. We refer the interested reader to the respective research papers for more information.

1.4 Performance comparison of DE methods

Many research has been conducted to compare performance between DE methods. Most of this research was done in the context of a publication of a new method, where the developer of the method tries to demonstrate that their method outperforms the other methods for at least one metric. We will only discuss the most recent research that was done in this context: Zhou et al. (2014), Love et al. (2014), Law et al. (2014) and Liu et al. (2015) that include benchmarks to demonstrate good performance of their newly developed methods, respectively edgeR robust, DESeq2, limmaVoom and limmaVoom with quality weights. The drawback of this type of benchmark is that the authors are less neutral and might select parameter settings that are most favorable for their method. Therefore, we also discuss the publications of Rapaport et al. (2013) and Soneson and Delorenzi (2013), which are more neutral. The drawback of this research is that it starts already to be outdated in the sense that it does not include all the latest methods or the methods' most recent version. An overview of the methods compared in these publications is given in Table 1.2. In the remainder of this section, we summarize the findings of each of these publications, focusing on those findings that relate to the methods that we discussed above and that we will use in our own analyses (grey part of Table 1.2).

	edgeR_classic	edgeR_glm	edgeR_rob	DESeq	DESeq2	LimmaQN	LimmaVoom	LimmaVoom_QW	LimmaVst	BaySeq	PoissonSeq	SAMSeq	Cuffdiff	DEGSeq	TSPM	NBPSeq	NOISeq	EBSeq	DSS	ShrinkBayes
Rapaport et al. (2013)	x			x		x	x			x	x									
Soneson and Delorenzi (2013)	x			x			x		x	x		x			x	x	x	x		x
Law et al. (2014)	x	x		x			x			x	x			x					x	
Love et al. (2014)		x	x	x	x		x					x	x					x	x	
Zhou et al. (2014)		x	x	x	x		x			x		x						x		x
Liu et al. (2015)							x	x												
Our research	x	x	x	x	x	x	x	x	x	x	x									

Table 1.2: Methods compared by publication - methods in grey are the ones used in our own research

Instead of using simulations, **Rapaport et al.** performs a benchmarking exercise that is based on a real dataset. They mainly use the Sequencing Quality Control (SEQC) dataset which consists out of 2x5 technical replicates and includes 92 spike-in controls as well as a set of about one thousand genes that were validated by TaqMan qPCR. The methods compared are edgeR classic, DESeq, limmaQN, limmaVoom, baySeq, PoissonSeq and Cuffdiff. Based on hierarchical clustering of samples after normalization and on the correlation of the estimated LFCs as reported by each method and the qPCR-values, they conclude that all methods considered perform well in terms of normalization. The ability of the methods to detect DE genes is assessed in terms of the AUC for the qPRC-validated genes, using an LFC threshold of 0.5 to classify a gene as DE. The AUC values indicate comparable performance among the methods with a slight advantage for edgeR and DESeq. The control of type I errors (at a nominal FDR of 0.05) is evaluated by checking the distribution of the p-values and looking at the number of false positives when performing an intra-condition comparison of the SEQC samples. limmaVoom turned out to be the only method without any false positives, followed by baySeq, edgeR classic and PoissonSeq. DESeq and limmaQN have considerably more false positives, the latter one mainly for genes with low average expression levels. A last aspect that the authors pay specific attention to is the methods' ability to detect genes expressed in only one condition. This aspect is evaluated on a different dataset, the ENCODE dataset, by investigating the relationship between the adjusted p-values of this set of genes with the signal-to-noise ratio (the ratio of mean over standard deviation) in the expressed condition. It appears that the limmaQN, limmaVoom and baySeq are the only methods that exhibit the desired monotonic behavior in this relation, indicating they are better able to detect genes expressed in only one condition. As an overall observation, the authors state that none of the methods emerged as favorable in all comparisons and that array-based methods adapted to RNA-seq perform comparably to methods designed for RNA-seq.

Soneson and Delorenzi examine performance of eleven methods, amongst which edgeR classic, DESeq, limmaVoom, limmaVst, baySeq and SAMSeq, by means of simulations. They assess the impact of the percentage of DE genes, the direction of differential expression, sample size and the presence of outlier genes amongst others. In terms of the ability to discriminate between DE and non-DE genes (expressed by the AUC), the authors find that in case of more samples (5 to 10 samples per condition) and symmetric differential expression, all methods perform similarly. In case of less samples and symmetric expression (2 samples per condition), edgeR classic, DESeq and the limma-based methods generally produce the best results. Asymmetry in differential expression only negatively influences the AUC for larger percentages of DE genes. SAMSeq is least affected by the asymmetry. Outliers reduce the AUC slightly for all methods, but less for the limma-based methods and SAMSeq. Looking at the type I error rate in the absence of truly DE genes and in case there are no outliers, all methods are found to control the type I error quite well, with DESeq at the conservative side. In the presence of outliers, the limma-based methods best control the type I error. Subsequently the authors look at the methods' ability to control the false discovery rate (FDR) when there are DE genes. For the lowest number of samples (2 per condition), the FDR is always poor: either the method does not detect any DE genes (the limma-based methods and SAMSeq) or the real FDR is way higher than the nominal FDR (the other methods). For a larger numbers of samples and in case of 10% DE genes (both symmetric and asymmetric differential expression), FDR control improves, but DESeq and baySeq are at the conservative side while edgeR classic remains too liberal. A higher percentage of DE genes improves FDR control in the case of symmetric differential expression and impairs FDR control in the case of asymmetric differential expression. The FDR of the methods that are based on the

NB distribution increased when outliers were introduced, while the FDR of the limma-based methods and SAMSeq was largely unaffected. The performance in terms of power is related to the performance in terms of FDR: DESeq and baySeq which have the lowest real FDR, also have the lowest power. edgeR classic has a higher power, but at the cost of a higher real FDR. SAMSeq appears to have high power and good FDR control for large numbers of samples, but does not detect any DE genes at all in case of a low number of samples.
Besides a simulation study, Soneson and Delorenzi also analyze real RNA-seq data from two mouse strains (Bottomly et al., 2011) and compare the results from the different methods. It turns out that baySeq and DESeq classify less genes as DE compared to the other methods. SAMSeq (after ShrinkSeq) has the highest number of DE genes. Overlap in the set of DE genes is large: almost all genes in the DE-set of a method that classifies less genes as DE are also included in the DE-set of methods that classify more genes as DE. Comparing the gene rankings, edgeR classic, DESeq, limmaVoom, limmaVst and SAMSeq give similar rankings, while baySeq gives a considerably different ranking.

Zhou et al. compare performance of edgeR glm, edgeR robust, DESeq, DESeq2, limmaVoom, EBSeq and ShrinkBayes. They do their simulation based on the Pickrell dataset (Pickrell et al., 2010) with 5+5 samples, 10% symmetric DE genes, a fold difference 3 for the DE genes, with and without 10% outliers generated by the 'simple' outlier generating mechanism. They find that the introduction of outliers results in more false positives and lower power at the same nominal FDR. In the absence of outliers, edgeR glm, edgeR robust and DESeq2 have a small advantage in power at the 5% FDR. In the presence of outliers, edgeR robust performs better than the other methods and DESeq experiences the strongest drop in power. Looking at the DE genes with outliers separately, they find that robust edgeR clearly outperforms the other methods in terms of power and in particular DESeq2 which seems to suffer form its hard threshold. Considering FDR control, the authors must admit that edgeR glm and edgeR robust do not meet the target FDR, while limma-voom controls the FDR well. They argue however that edgeR glm, edgeR robust, DESeq2 and limma-voom achieve similar power-to-achieved-FDR tradeoffs across sample sizes with a slight advantage for edgeR robust if outliers are present and with an advantage for DESeq2 for smaller fold changes.

Love et al. perform a simulation study that is similar to the one of Zhou et al.. It is also based on the Pickrell dataset but uses different parameter settings (varying number of samples, 20% symmetric DE genes, fold changes of 2, 3 and 4 and a nominal FDR of 10%). The authors emphasize that DESeq2 often has the highest power of the algorithms that control the FDR in the sense that the actual FDR is at or below the nominal FDR. In addition they state that DESeq2 estimates fold changes more precisely than edgeR [2] in that it con-

[2] It is not fully clear if Love et al. refers to edgeR glm or edgeR robust here

sistently has a lower low root-mean-square error. Besides benchmarking through simulation, the authors also performed benchmarking on real datasets. Power and FDR were assessed by splitting a large dataset into an evaluation set and a larger verification set and compare the calls from the evaluation set with the calls from the verification set which were taken as the truth. We won't further discuss these results, as we think that calls from the verification are not necesseraliy a good approximation of the true differential state.

Law et al. find in their nullsimulation that their limmaVoom method most accurately controls the type I error rate. In their simulation with DE genes, they use a rather low fraction of DE genes (2%) and a moderate fold change of 2. They find that limmaVoom has the best power of the methods that control the false discovery rate, both in the case of equal and unequal library sizes. In terms of gene ranking, limmaVoom achieves the lowest FDR at any cutoff, followed by edgeR classic and edgeR glm in case of unequal library sizes and by PoissonSeq in case of equal library sizes. Liu et al. compare performance between limmaVoom with and without quality weights in case of sample level variability. In the absence of variability at sample level, limmaVoom and limmaVoom with quality weights have similar performance. However, in case of substantial sample variability, limmaVoom with quality weights successfully down-weighs samples with higher variability and achieves a higher power and a lower FDR than limmaVoom without quality weights.

Chapter 2

Data & Methodology

2.1 Overall approach, goals and scope

The research of this master thesis was conducted in cooperation with Janssen Pharmaceuticals (referred to as Janssen in the remainder of this work). Key purpose is to compare and evaluate differential gene expression analysis methods for RNA-seq data, with a focus on methods that are available in R. Our research consists of **two pillars** each of which have their own goals.

The first part consists of a **concordance analysis** that should help to **understand similarities and dissimilarities** between DE methods. In this part of the research the selected methods are applied on both publicly available and in-house Janssen data. The different datasets reflect a variety of settings in which DE analysis is applied. The outputs of the different methods are then compared in order to evaluate which methods are more alike than others in terms of the set of DE genes, the ranking of the genes and the estimated fold changes. In addition, we analyze the top 100 of most significant genes to see if some methods are more likely than others to classify specific types of genes as DE.

The second part consists of a **simulation study** that aims to empirically **assess the performance** of the different methods. While the concordance analysis allows to make statements on how methods differ from each other, these analyses do not allow to make statements on which method is better than the others. Therefore we resort to a simulation study. We generate a number of count datasets according to a model that reflects a real dataset as well as possible. As such we can control which genes are differentially expressed and which are not. By comparing the output of each DE method with the underlying truth of the model, we are then able to assess how well each method classifies the genes as DE or non-DE. We evaluate each method according to multiple criteria and we investigate how performance depends on a number of relevant parameter settings, namely the fraction of DE genes, symmetric or asymmetric expression, the number of samples, the fraction of outliers, the size of the fold change and the average count.

In both parts of our research, we limit the scope to a setting with two conditions. All analyses are run on a Windows 7 64-bit computer with 8GB RAM and an Intel(R) Core(TM) i7-4600U CPU @ 2.10Ghz, in R version 3.1.2 (2014-10-31) 'Pumpkin Helmet'.

2.2 Datasets

The **concordance analysis** has been run on a number of datasets of which one in-house Janssen dataset and four publicly available datasets. A summary of the datasets is given in Table 2.1 and some more context is given below:

- The **CRC AZA dataset** is an in-house Janssen dataset. It includes data on 2x3 biological replicates of colorectal cancer cell lines (HCT-116): three controls and three which have been treated with Azacytidine. As Azacytidine is an inhibitor of methyltransferase, it impacts transcription regulation. Azacytidine is an aspecific drug such that a global response can be expected.
- **Bottomly et al. (2011)** uses both RNA-seq data and microarray data to detect differential gene expression between the C57BL/6J (B6) and DBA/2J (D2) mouse strains, two of the most commonly used inbred mouse strains in neuroscience research. The study evaluates concordance of the RNA-Seq results with the result of two microarray platforms. The RNA-seq count dataset includes 10 biological replicates of the B6 strain and 11 biological replicates of the D2 strain.
- **Hammer et al. (2010)** performs RNA-seq on the L4 dorsal root ganglion (DRG) of rats with chronic neuropathic pain induced by spinal nerve ligation (SNL) of the neighboring (L5) spinal nerve to demonstrate its potential for in vivo transcriptomics in the nervous system. They use count data of 2x2 biological replicates: two controls and two with L5-SNL induced chronic neuropathic pain.
- **The GTEx project**, which stands for the Genotype-Tissue Expression project, provides a large dataset with gene expression data of both RNA-seq and microarrays. More than 3000 RNA-seq samples of in total 54 different tissues are included. For our research, we randomly selected 2x10 biological replicates of the Hippocampus and Hypothalamus.
- **Rapaport et al. (2013)** uses samples from two sources that are part of the SEQC (RNA sequencing quality control) study, each generated from a mixture of biological sources and a set of synthetic RNAs from the External RNA Control Consortium (ERCC) at known concentrations. The first group contains the Strategene Universal Human Reference RNA (UHRR), the second group contains Ambion's Human Brain Reference RNA (HBRR). The samples in both groups are mixed with 2% by volume of

respectively ERCC mix 1 and ERCC mix 2. The ERCC mixtures in both groups contain concentrations of four subgroups of in total 92 synthetic genes. The log expression change of these spike-in genes is predefined, such that they can be used to benchmark DE performance. For each group, there are five technical replicates (of which the first four were prepared by a single technician and the last one by Illumina).

Dataset	Type	Cond A	Cond B	Replicates (A+B)	Replicate type
CRC AZA	Janssen in-house	Control	Treated with AZA	3+3	biological
Bottomly	Public	B6 mouse strain	D2 mouse strain	10+11	biological
Hammer	Public	Control	L5 SNL	2+2	biological
GTEx	Public	Hippocampus	Hypothalamus	10+10	biological
Rapaport	Public	Universal Human Reference RNA	Human Brain Reference RNA	5+5	technical

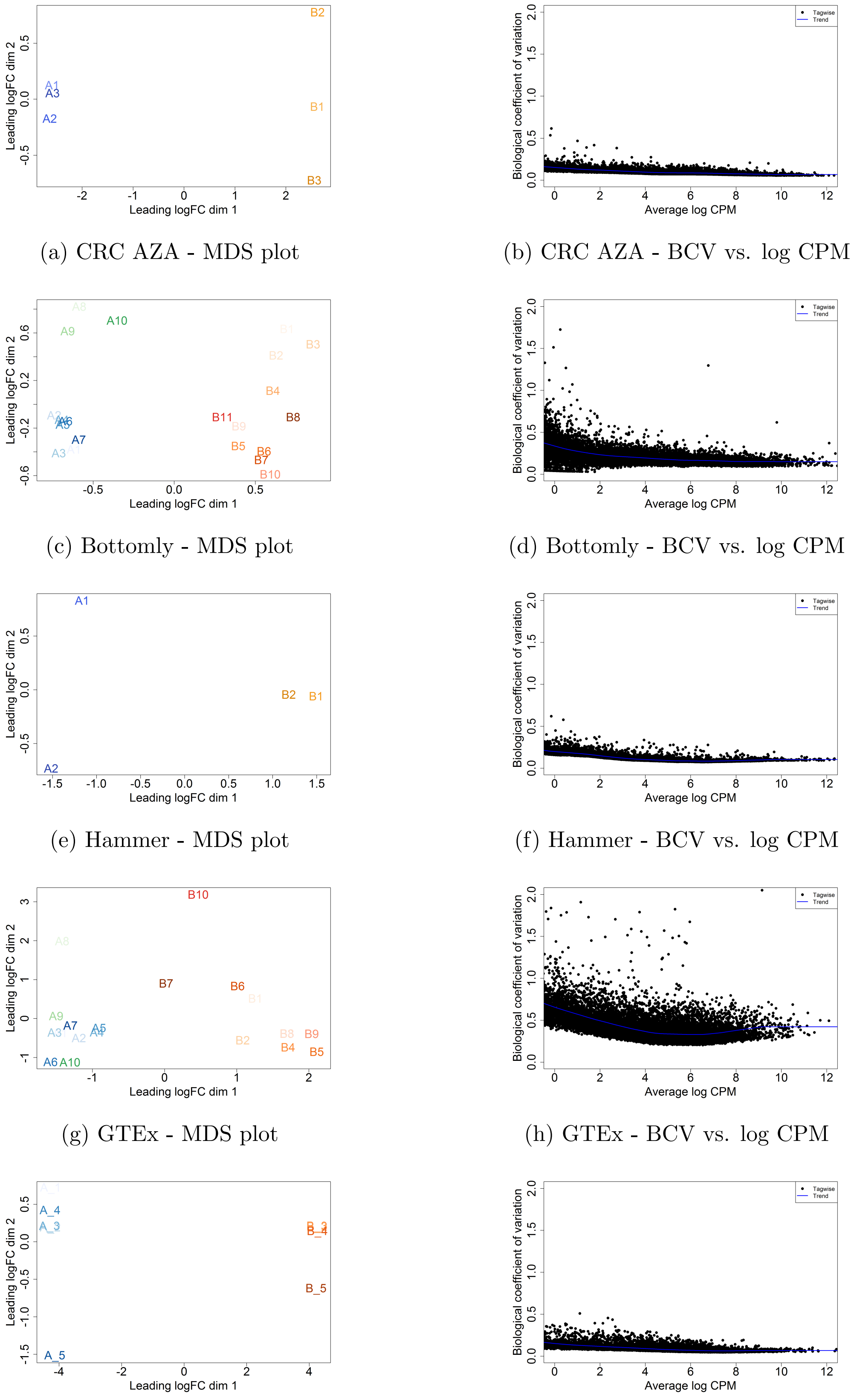
Table 2.1: Overview datasets concordance analysis

These datasets reflect the **variety of settings** in which RNA-Seq is used in terms of:

- The type of conditions:
    - Comparison of a single tissue in two different populations: non-treated vs. treated (CRC AZA), non-diseased vs. diseased (Hammer) or different strains (Bottomly)
    - Comparison of different tissues in one population: GTEx and Rapaport
- The type of replicates: biological vs. technical replicates (see also Section 1.2.2)

The setting is related to both between-condition and within-condition variability of the samples. To get an insight in the between-condition and within-condition variability, two types of plots are constructed: a multidimensional scaling (MDS) plot of the samples and a plot of the biological coefficient of variation (BCV) against gene abundance (in $\log_2$ counts per million). The MDS-plots, which are constructed by means of a built-in procedure in edgeR, visualize for each dataset the distances between the gene expression profiles in a two-dimensional space. For its construction, counts are converted to log counts per million (log cpm) and the Euclidean distances between the samples are calculated based on the 500 genes that are most distinguishing between the samples. From the MDS-plots (Figure 2.1 a-c-e-g-i) we can see that the replicates of the different conditions are well separated for each dataset. The plots showing the biological coefficient of variation versus the log average expression are based on the dispersion estimates as produced by the edgeR robust procedure. These BCV-plots (Figure 2.1 b-d-f-h-j) indicate that the Bottomly and the GTEx dataset are characterized by higher biological variability relative to the other datasets.

The **simulation analysis** is based on simulated datasets that try to **mimic the CRC AZA dataset**. A detailed explanation of how the simulation was set up is given in Section 2.4.



(a) CRC AZA - MDS plot

(b) CRC AZA - BCV vs. log CPM

(c) Bottomly - MDS plot

(d) Bottomly - BCV vs. log CPM

(e) Hammer - MDS plot

(f) Hammer - BCV vs. log CPM

(g) GTEx - MDS plot

(h) GTEx - BCV vs. log CPM

(i) Rapaport - MDS plot

(j) Rapaport - BCV vs. log CPM

Figure 2.1: MDS-plot and BCV-plot by dataset. In the MDS-plots A and B indicate two different conditions. The MDS-plots show that the replicates of the different conditions are well separated for each dataset. The BCV-plots indicate higher biological variability for the Bottomly and GTEx datasets relative to the other datasets.

Two final remarks need to be made on the way we treated these datasets. First, while most count datasets were originally at the level of the ensembl gene ID, we preferred to do the DE analyses **at the level of the external gene ID** as this allows an easier biological interpretation of the results afterwards. A small minority of external gene IDs matched multiple ensembl gene IDs. For these external gene IDs the counts of the multiple corresponding ensembl gene IDs were added up. Second, a **pre-filtering** of the genes was done. Only protein coding genes were kept as the RNA-seq datasets were prepared using polyA capture and the primary focus of this method is mRNA. In the remainder of this document, when we talk about RNA-seq we actually refer to mRNA-seq. In addition, of these protein coding genes only these genes with an average count larger than 1 were retained for the final dataset on which DE analysis was applied.

2.3 Concordance Analysis

2.3.1 Methods used

There are many different methods for RNA-seq analysis. As discussed in Section 1.3, we limit ourselves to the most frequently used ones: edgeR, DESeq(2), limma, baySeq and PoissonSeq. Cuffdiff which is another frequently used method is not included because it is not available in R. On the other hand, we include SAMSeq to also have a non-parametric method in our set of methods tested. A discussion of the theory behind each method is given in Section 1.3. Here we only list the version of the software packages and the parameter settings used:

- **edgeR (v.3.8.6)**: we use three variants, each time applying the default parameter settings. First, we use the original variant that uses an exact test to test for DE (edgeR classic). Second, we apply the variant developed to deal with multifactor designs and that uses a likelihood ratio test (edgeR glm). Third, we use the extension of edgeR glm that was developed to deal with outliers (robust edgeR).
- **DESeq (v.1.18.0)**: the dispersion estimate call `estimateDispersions` is used with its default parameter values `sharingMode="maximum"`, `method="pooled"` and `fitType="parametric"`. These default values are different from the default values in earlier DESeq versions and also differ from the parameter settings used by Rapaport et al. and Soneson and Delorenzi.
- **DESeq2 (v.1.6.3)**: the default parameters are used which implies a parametric fit to estimate dispersions. This more recent version of DESeq was not available yet at the publication of Rapaport et al. and Soneson and Delorenzi.
- **limma (v.3.22.7)**: four limma-based variants are used, which differ in the way how they transform the count data. limmaQN uses the `normalizeBetweenArrays` function with `method="quantile"` to perform a quantile normalization on the $\log_2$ transformed gene counts. limmaVoom and limmaVoom with quality weights respectively use

the `voom` and `voomWithQualityWeights` functions. While the original limmaVoom approach used a total count normalization, the current default is to use TMM. limmaVst employs the variance stabilizing transformation (with parameters `fittype="local"` and `method="blind"`) provided by the DESeq package.

- **baySeq (v.2.0.50)**: contrary to Rapaport et al., but in line with Soneson and Delorenzi baySeq is applied without sequence length correction and with TMM normalization. NB parameter estimation is done assuming equal dispersions and using quasi-likelihood estimation. It needs to be mentioned that baySeq does not report p-values. Instead we use the posterior probability of the DE model to classify genes as DE or non-DE. If in the remainder of this document we compare adjusted p-values between methods, we use these posterior probabilities for baySeq.
- **PoissonSeq (v.1.1.2)**: no minimum for the total number of reads across samples is applied and the number of permutations is set to 500.
- **SAMSeq (v.2.0)**: we use the default parameter values. In contrast to Love et al. (2014) who performs a Benjamini Hochberg correction of the p-values, but in line with Zhou et al. (2014), we classify the genes based on their q-value which can be considered as the adjusted p-value.

2.3.2 Analysis

Once the different methods are run on the dataset of interest, we compare the results to understand similarities and dissimilarities between the methods.
First, we look into the **number of genes** that have been **classified as DE** by each of the methods at an FDR of 5% and we check which methods classify most or least genes as DE across the datasets.

Subsequently, we compare **MA-plots and volcano plots** to see if we can observe any meaningful differences between the methods. MA-plots plot the estimated LFC (M) between two conditions vs. the average expression level (A), with a color code indicating the method's classification of the gene (DE vs. non-DE). Volcanoplots plot the adjusted p-value vs. the estimated LFC, again with a color code for the method's classification.

Afterwards, the **overlap in the set of DE genes** is compared by means of a pairwise overlap table. This table is also visualized by means of an MDS-plot. Here, the MDS-plot is constructed based on the chi-square distances of the pairwise overlap table. In addition, we plot the adjusted p-values (or the posterior probabilities for baySeq) of the different methods versus each other to assess similarity of the methods in terms of gene ranking by significance. For each combination of methods, the degree of similarity in the ranking of the genes is expressed as a Spearman rank correlation coefficient. As we are primarily interested in the most significant genes, we zoom in on these genes by making a scattermatrix of the -$\log_{10}$ transformed adjusted p-values of all genes. A pairwise overlap table of the top 100 most significant genes completes our view on the degree of correspondence between the methods for the most significant genes. A methodological side note here is that if ties occur in the smallest p-values, the genes are taken in the order they are returned by the method.

Besides assessing to which extent the methods return the same set of (top) DE genes, we also want to understand similarities in the **estimated LFC-values**. Again we construct a scatter matrix and calculate a correlation coefficient between the estimated LFCs for each combination of methods. This time we use the Pearson correlation coefficient as we do expect to see a linear relationship between the estimated LFCs.

Finally, we perform an **explorative analysis** to see whether some **methods classify particular genes** more easily as DE compared to other methods. To that extent we plot the normalized counts per condition and per sample of the top 3 most significant genes for each of the methods. To ensure consistency in the graphs, we use a total count normalization for all methods (rather than using the normalized counts as calculated by the method's default normalization function). Besides these graphs, we also calculate some key metrics for the top 100 most significant genes:

- the number of top 100 genes with low counts in one of the conditions: to check if some methods are more likely than others to classify genes with very low expression in one condition as DE. Low count in a condition has been defined as an average normalized count in the condition of less than 5.
- the number of top 100 genes with a dispersion factor that is in the upper quartile of the dispersions of all genes: to check if some methods are more likely than others to classify genes with high dispersions as DE.
- the number of the top 10 most expressed genes included in the top 100 most significant genes: to check if some methods are more likely than others to classify genes with the highest average counts as DE.

A watch out is again that if there are ties in the smallest p-values, we just take the genes in the order they are returned by the respective method. To have a feel of the size of the issue, we report for each dataset and for each method the number of unique p-values in the top 100 most significant genes.

2.4 Simulations

2.4.1 Settings

For the simulations we start from **the flexible framework offered by Zhou et al. (2014)**, which we then further adapt to facilitate our own needs. For a pre-specified number of genes and samples, counts are simulated according to a negative binomial distribution. In order to mimic real data as well as possible, a real count dataset serves as the basis to generate both the library sizes and the parameters of negative binomial distribution. The library sizes are generated following a uniform distribution over the interval between 70% and 130% of the median library size of the real dataset used. Then for each gene the parameters of the NB are set by sampling from the joint distribution of the estimated log-cpm and dispersion of the real count dataset (of which the 10% genes with most extreme dispersion have been removed though). Subsequently, a pre-specified percentage of these genes is randomly sampled to serve as DE genes. For these genes, the mean parameter of the NB distribution is adapted in both conditions, to ensure that the LFC between the two conditions takes a pre-specified value, while the average count of the gene remains unchanged. Then for each gene and each sample a count is generated according to the NB distribution. Finally, a pre-specified percentage of outlier genes is introduced in the dataset that was simulated as such.

A large number of **parameters** can be adapted to ensure performance of the methods can be tested across a variety of conditions. These parameters can be grouped in **3 categories**:

- Parameters related to general characteristics of the simulation: the number of simulated datasets, the number of features, the number of experimental conditions, the number of samples per experimental conditions and the underlying dataset which is used for determining the mean-dispersion relationship
- Parameters related to differential expression: the proportion of DE features, the direction of differential expression and the relative expression level of truly DE features
- Parameters related to outliers: the proportion of outliers, the magnitude of the outlying observations and the outlier mechanism

We only **manipulate** a subset of these **parameters**:

- The proportion of DE genes (pDiff) takes values of 1%, 5%, 20% and 70%
- The proportion of DE genes that is upregulated (pUp) is set to 50% and 70%
- The proportion of outliers (pOutlier) takes values of 0%, 5% and 10%
- The number of samples per condition is set to 3, 5 and 7

We will always vary one parameter at a time and compare to the baseline (parameter settings: 5+5 samples, pDiff=5%, pUp=50%, pOutlier=0%).

The other **parameters** we keep **fixed** over the different simulations:

- For each set of parameter values, we generate 20 simulated datasets
- The number of features is set to 10,000
- The number of conditions is kept at two
- The CRC AZA dataset is used as the underlying dataset for the simulations

- For the absolute LFCs of the truly DE genes, we use a rescaled beta-distribution. This approach is discussed in more detail below
- For the outlier mechanism we always use the 'simple' method (S). In this method a gene is chosen at some probability to have a single outlier randomly added. Other outlying generating mechanisms are described in Zhou et al.
- Counts are multiplied by a random factor between 1.5 and 10 for outlier genes

The most important difference of our approach versus the Zhou et al. approach is the way the **relative expression levels of truly DE genes** are modeled. Zhou et al. randomly samples a certain percentage of genes that will serve as DE genes in the simulated dataset. For these genes, the mean parameter of the NB distribution in both conditions is modified so as to impose a certain LFC which is the same for all DE genes. We do the same, but instead of using the same LFC for all DE genes, we sample the LFC for each DE gene from a distribution as we believe this better approaches reality. To this extent, we use a rescaled beta-distribution with shape parameters such that the absolute LFC of the truly DE genes shows an exponential-like form between a lower bound $LFC_{min} = \log_2(1.5)$ and an upper bound $LFC_{max} = 4$ with mean $LFC_{mean} = 1$ and 95% quantile $LFC_{95p} = 2$. This is illustrated in Figure 2.2. Note that the user can easily choose alternative values for LFC_{min}, LFC_{max}, LFC_{mean} and LFC_{95p}.

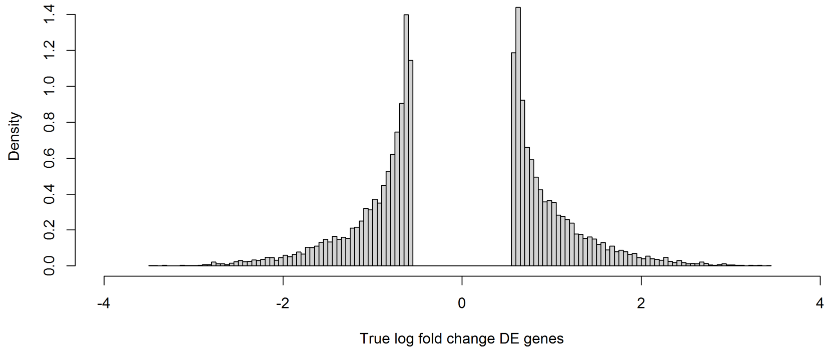
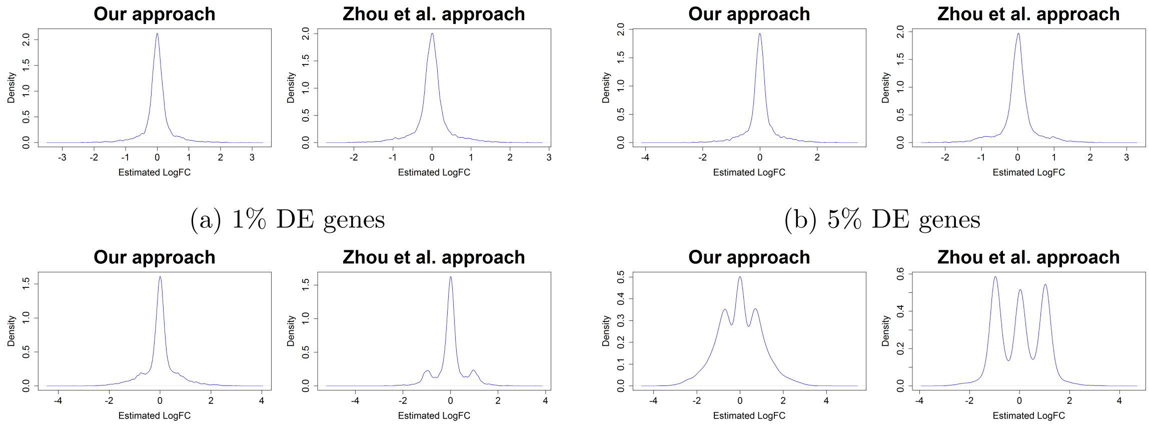


Figure 2.2: Distribution true LFC of DE genes for a dataset with an equal percentage of upregulated and downregulated genes - A certain percentage of genes is randomly sampled to serve as DE genes. The LFC of these DE genes is sampled from the rescaled beta-distribution depicted above.

For real datasets the distribution of the estimated LFC is usually a unimodal curve with the maximum very close to zero. When using a single value for the absolute LFC of the truly DE genes, the distribution of the estimated LFCs quickly becomes multimodal. With our approach the distribution of the estimated LFC only starts to deviate from the expected form for extreme values of pDiff. This is illustrated in Figure 2.3. Note that with our approach the effect of the LFC can be assessed within a single simulation, rather than doing different simulations for different values of the fold change.



(a) 1% DE genes

(b) 5% DE genes

(c) 20% DE genes

(d) 70% DE genes

Figure 2.3: LimmaVoom estimated LFC distribution for different fractions of DE genes (5-5 samples, pUp=0.5, no outliers) in case a single value is used for the true LFC of the DE genes (Zhou et al. approach) and in case the LFC of the DE genes is drawn from the distribution in Figure 2.2 (our approach). When using a single value for the LFC of the DE genes, we already see a multimodal distribution of the estimated LFC for relatively small fractions of DE genes.

A second difference with the Zhou framework is that we compare a **different set of methods**. The Zhou framework can easily be extended with new methods or variations of existing ones by writing simple wrapper functions with the correct inputs and outputs. We wrote wrapper functions with exactly the same settings for the methods described in Section 2.3.1. Note that we dropped baySeq from the simulations because of the method's high computation time, as illustrated by Figure 2.4 which shows the computation time by method for the CRC AZA dataset.

2.4.2 Performance analysis

The **performance** of each DE method under varying conditions was assessed using a number of standard **metrics**:

- The False Discovery Rate **(FDR)** expresses the proportion of incorrect null hypothesis rejections ('false discoveries') versus the total number of null hypothesis rejections. FDR-control provides less stringent control of false positives compared to the familywise error rate (FWER). In the ideal scenario, a method has a true FDR that coincides with the nominal FDR, which we set at 5%.
- The False Positive Rate **(FPR)** expresses the proportion of non-DE genes for which the null hypothesis is incorrectly rejected at a certain predefined nominal FDR. It provides

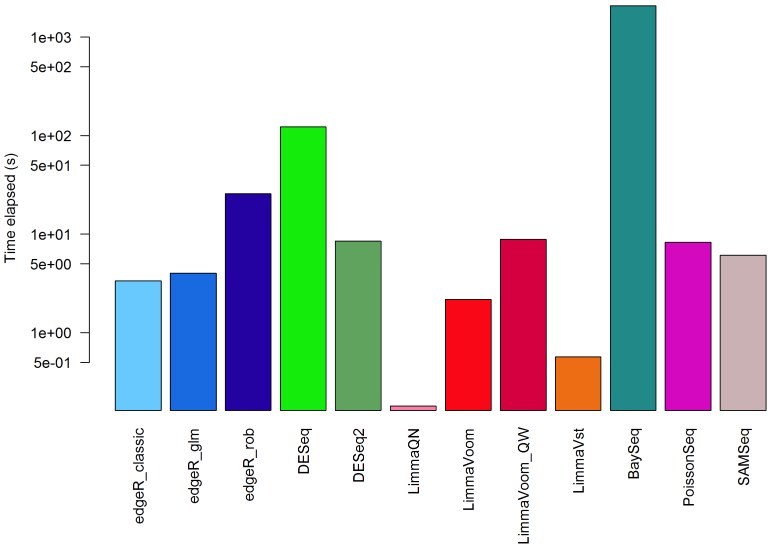


Figure 2.4: Elapsed time by method for analyzing the CRC AZA dataset; Running baySeq on the CRC AZA dataset takes 34 minutes, which is almost 17 times longer than DESeq, the method with the second longest run time.

an alternative view on the number of false positives. While the DE methods are designed to control the FDR, they do not directly control the FPR.

- The **Power** is empirically estimated by means of the true positive rate (TPR) which expresses the proportion of truly DE genes that are detected as such at a certain predefined nominal FDR.
- The area under a Receiver Operating Characteristic (ROC) curve **(AUC)** expresses the methods' ability to rank truly DE genes ahead of non-DE genes.

As some genes are excluded from analysis by some of the methods, we assigned these genes an adjusted p-value of 1 so as to be able to calculate the above metrics for exactly the same set of genes for all methods. Per metric, per set of conditions and per method, a boxplot was created showing the values of the metric for the simulated datasets. We will see that all methods involve a trade-off between power and FDR. As not all methods control the FDR equally well, assessing the power of the methods for a given nominal FDR is not always a fair comparison. From this perspective, we will also plot for each method the trade-off between the power and the true FDR. An alternative visualization of the methods' ability to rank truly DE genes ahead of non-DE genes has been included under the form of a FDR-curve depicting the number of false detections encountered when going through the list of genes ranked according to their adjusted p-value.

Whereas the above metrics help assess the methods' ability to correctly classify genes as DE or non-DE, they don't capture how well the methods **estimate the LFC**. To assess this dimension of the methods' performance, we calculate for each method the average bias which is the average difference between the estimated and the true LFC, and the root mean squared error (RMSE) which expresses the sample standard deviation of the prediction errors. In addition, the relation between the estimated and the true LFC for each method and per level of the average count has been visualized by means of a scatterplot. The analysis is done once for all genes and once for outlier genes only.

Chapter 3

Results & Discussion

The full set of outputs of both the concordance analysis and the simulations is included in the separate document ‘Supplementary figures and tables’. In this chapter we will only show the most insightful figures and tables. For the concordance analysis, we will mainly focus on the outputs of the CRC AZA dataset. For the simulations we focus on the simulation run with 5 samples in each condition, 5% DE genes, a symmetric DE pattern, both with and without 5% outlier genes. Where needed we will refer to the figures and tables in the ‘Supplementary figures and tables’ document, indicating these outputs with a prefix ‘S’.

3.1 Concordance analysis

By comparing the outputs of the different DE methods for several real RNA-seq datasets, we try to get an understanding which DE methods are more alike than others in terms of the set of DE genes, the ranking of the genes by adjusted p-value and the estimated fold changes. Looking at the **number of genes that are classified as DE** by each of the methods at a nominal FDR of 5%, we see that DESeq and baySeq and to a lesser extent limmaQN return less DE genes compared to each of the other methods. This is shown for the CRC AZA dataset in Figure 3.1. The conservative behaviour of DESeq and baySeq, which has also been observed by Soneson and Delorenzi, is even more outspoken for the Bottomly dataset (Figure S16) and for the GTEx dataset (Figure S40), two datasets that are characterized by higher biological variability and a higher number of replicates compared to the other datasets. While limmaVoom with quality weights classifies way more genes as DE compared to the other methods for the CRC AZA dataset, this is not the case for the other datasets. Table 3.1 shows that for the Bottomly, Hammer and Rapaport datasets, SAMSeq classifies most genes as DE relative to the other methods and for the GTEx Dataset DESeq2 is most liberal. Probably the most important key takeaway from this table is the high spread in the number of genes classified as DE by the different methods. For the Rapaport and GTEx dataset, the ratio of the number of DE classified genes between the most liberal and most conservative method amounts to values larger than 3.

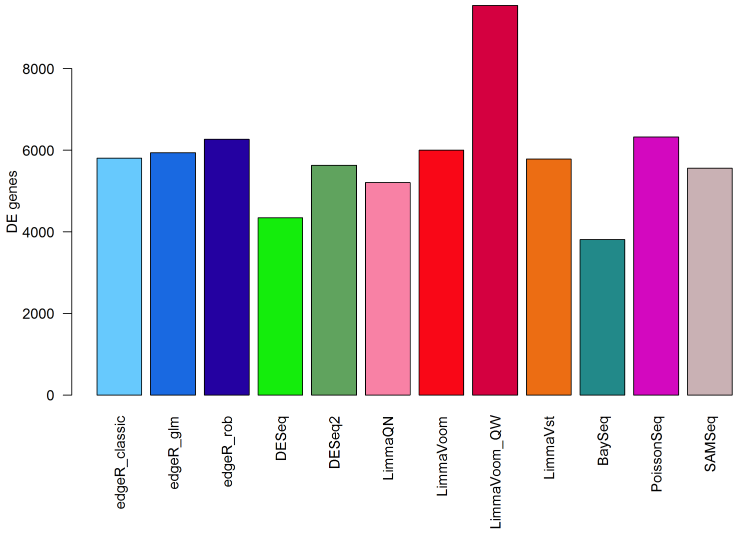
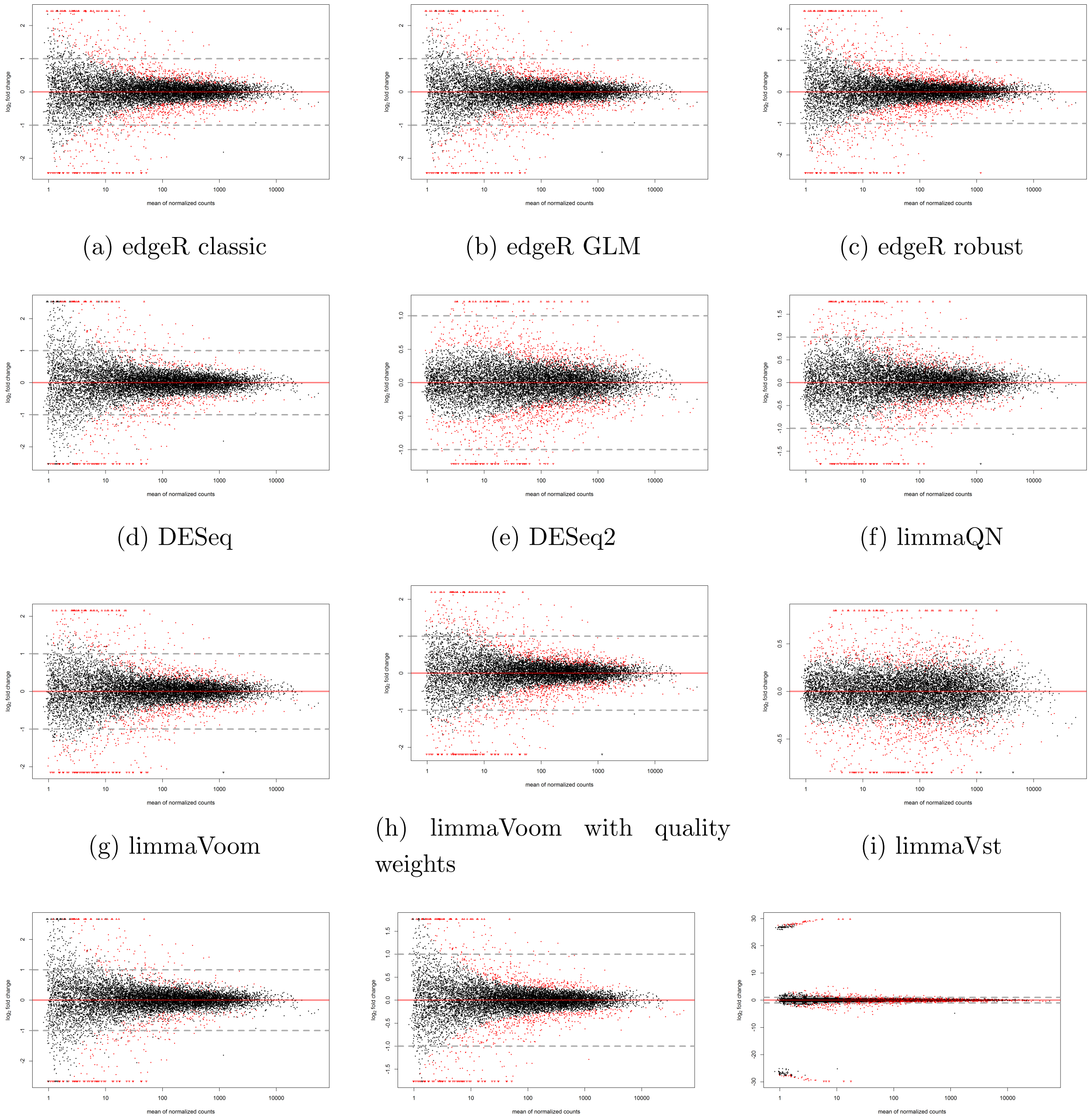


Figure 3.1: Number of genes detected as DE by method at a nominal FDR of 5% for the CRC AZA dataset - DESeq and baySeq classify less genes as DE compared to the other methods

Dataset	Least DE		Most DE		Ratio Most/Least
	Method	Number DE	Method	Number DE	
CRC AZA	baySeq	3809	limmaVoom QW	5988	1.57
Bottomly	baySeq	506	SAMSeq	1894	3.74
Hammer	limmaQN	4416	SAMSeq	10392	2.35
GTEx	DESeq	788	DESeq2	2461	3.12
Rapaport	DESeq	14644	SAMSeq	16554	1.13

Table 3.1: Number of DE genes by dataset for the most conservative and the most liberal method at a nominal FDR of 5% - the number of genes classified as DE can vary considerably between methods

A few things can be learned from comparing the **MA-plots** of the different methods. Here, we only include the MA-plot (Figure 3.2) of the Bottomly dataset, for the other datasets we refer to Figures S5 (CRC AZA), S29 (Hammer), S41 (GTEx) and S53 (Rapaport). For most methods we see the familiar trumpet-shape form in the MA-plots. A first exception is SAMSeq which consistently displays two 'tails' with extreme estimated LFCs for a number of low count genes. A second exception is the MA-plot of the DESeq2 and limmaVst methods. For these methods, the strongest LFCs are no longer exhibited by genes with the lowest expression. For DESeq2, the reason is that the MLE-estimates for the LFC are shrunken towards zero by means of an empirical Bayes procedure. The variance stabilizing transformation of limmaVst seems to have a similar effect. Other often used plots are volcano plots. We don't discuss them in detail here, but they are included as Figures S6, S18, S30, S42 and S54 in the supplementary materials.



(a) edgeR classic

(b) edgeR GLM

(c) edgeR robust

(d) DESeq

(e) DESeq2

(f) limmaQN

(g) limmaVoom

(h) limmaVoom with quality weights

(i) limmaVst

(j) baySeq

(k) PoissonSeq

(l) SAMSeq

Figure 3.2: MA-plots by method for the Bottomly dataset; Red points indicate genes with an adjusted p-value<0.05, black points indicate genes with an adjusted p-value≥0.05; Small triangles at the side of the plotting window indicate points that would fall outside the plotting window. The dashed grey lines indicate the values +1 and -1 for the LFC.

Next we look into the **overlap in the set of DE genes** as returned by the different methods. Here we see two types of patterns occurring. We illustrate them on the Hammer and Bottomly datasets as these datasets most clearly contrast the two patterns observed and also have the highest percentage of variance explained for their MDS-plots. Tables 3.2 and 3.3 show the pairwise overlaps in the DE genes for each combination of methods for these datasets. Figure 3.3 and 3.4 visualize these overlap tables by means of an MDS-plot which is based on a chi-square distance measure. The overlap tables and the MDS-plots of the CRC AZA dataset and the Rapaport dataset are similar to those of the Hammer dataset and are given by Tables S-1, S-13 and Figures S-7, S-55 respectively. The picture for the GTEx dataset is similar to the one of the Bottomly dataset and is given in Table S-10 and Figure S-43.

First and most importantly, we see there is good overlap between the methods: the vast majority of genes that are included in the DE-set of one method are also included in the DE-set of any other method. Second, SAMseq appears to be most different from the other methods. Part of this is driven by the fact that SAMSeq in general classifies more genes as DE compared to the other methods. However, in the CRC AZA dataset the number of genes classified as DE by SAMSeq is in line with the other methods, still it appears in isolation on the MDS-plot because it has a lower overlap with the other methods. The relative position of the other methods shows two different patterns according to the dataset. For the Hammer, CRC AZA and Rapaport datasets, limmaQN is positioned further away from the other methods, indicating it shows somewhat less overlap with the other methods. The other methods are grouped together with some subgrouping by family of methods: edgeR classic, glm and robust give very similar results, the same holds for DESeq and DESeq2 and for LimmaVoom and LimmaVoom with Quality Weights. For the Bottomly and GTEx datasets, the picture is slightly different. There DESeq and baySeq are clearly separated from the other methods, obviously because their conservative character is very much outspoken for these datasets. LimmaQN is now much closer to limmaVoom and limmaVst. edgeR robust, DESeq2 and limmaVoom with quality weights tend to cluster together.

	edgeR_classic	edgeR_glm	edgeR_rob	DESeq	DESeq2	LimmaQN	LimmaVoom	LimmaVoom_QW	LimmaVst	BaySeq	PoissonSeq	SAMSeq
edgeR_classic	6701	6677	6690	5321	5316	3995	6051	6069	5285	4656	6560	6688
edgeR_glm	6677	6825	6823	5328	5317	4036	6090	6109	5348	4656	6664	6795
edgeR_rob	6690	6823	7009	5343	5331	4047	6149	6173	5356	4656	6765	6975
DESeq	5321	5328	5343	5361	5112	3638	5290	5295	4719	4403	5295	5361
DESeq2	5316	5317	5331	5112	5358	3570	5188	5202	4620	4298	5274	5356
LimmaQN	3995	4036	4047	3638	3570	4416	3944	3937	4138	3688	4130	4245
LimmaVoom	6051	6090	6149	5290	5188	3944	6173	6123	5150	4648	5976	6172
LimmaVoom_QW	6069	6109	6173	5295	5202	3937	6123	6201	5159	4652	5999	6200
LimmaVst	5285	5348	5356	4719	4620	4138	5150	5159	5551	4549	5499	5509
BaySeq	4656	4656	4656	4403	4298	3688	4648	4652	4549	4656	4646	4656
PoissonSeq	6560	6664	6765	5295	5274	4130	5976	5999	5499	4646	7684	7426
SAMSeq	6688	6795	6975	5361	5356	4245	6172	6200	5509	4656	7426	10392

Table 3.2: Overlap DE genes between methods at nominal FDR of 5% for the Hammer dataset. The numbers on the diagonal indicate the number of DE genes found by the respective methods, the numbers off-diagonal show the number of DE genes that are shared between each pair of methods

	edgeR_classic	edgeR_glm	edgeR_rob	DESeq	DESeq2	LimmaQN	LimmaVoom	LimmaVoom_QW	LimmaVst	BaySeq	PoissonSeq	SAMSeq
edgeR_classic	1222	1208	1198	604	1163	975	998	1095	1040	504	854	1186
edgeR_glm	1208	1245	1224	604	1173	984	1009	1117	1041	504	854	1199
edgeR_rob	1198	1224	1550	604	1234	1013	1033	1275	1070	505	883	1395
DESeq	604	604	604	604	604	584	597	600	596	463	601	603
DESeq2	1163	1173	1234	604	1329	974	1003	1151	1051	504	847	1297
LimmaQN	975	984	1013	584	974	1029	936	959	985	501	726	1009
LimmaVoom	998	1009	1033	597	1003	936	1036	1010	972	504	765	1024
LimmaVoom_QW	1095	1117	1275	600	1151	959	1010	1401	1017	505	853	1301
LimmaVst	1040	1041	1070	596	1051	985	972	1017	1090	503	780	1082
BaySeq	504	504	505	463	504	501	504	505	503	506	489	505
PoissonSeq	854	854	883	601	847	726	765	853	780	489	961	880
SAMSeq	1186	1199	1395	603	1297	1009	1024	1301	1082	505	880	1894

Table 3.3: Overlap DE genes between methods at nominal FDR of 5% for the Bottomly dataset. The numbers on the diagonal indicate the number of DE genes found by the respective methods, the numbers off-diagonal show the number of DE genes that are shared between each pair of methods

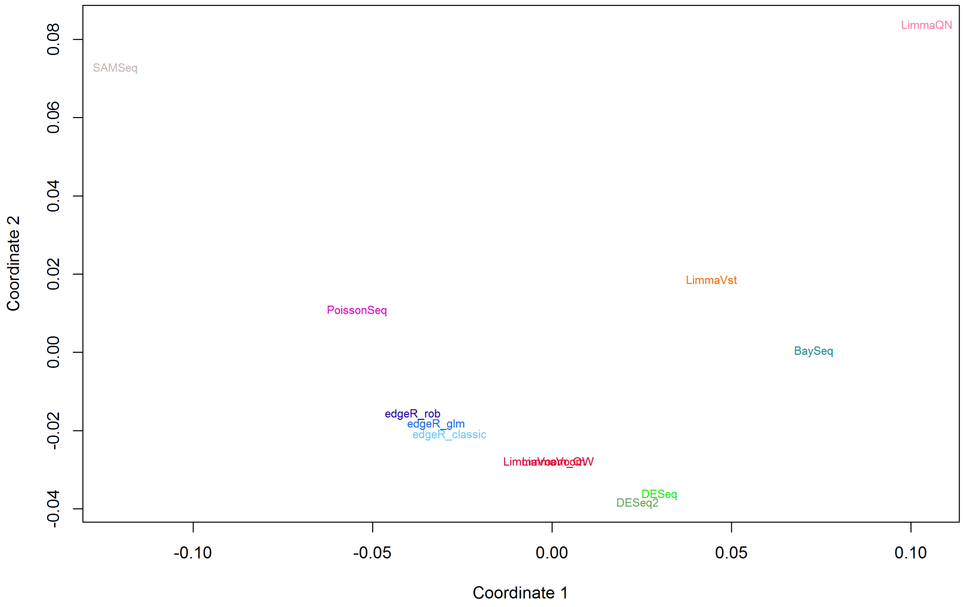


Figure 3.3: MDS plot DE overlap between methods for Hammer dataset - Construction of the MDS-plot is based on the chi-square distance. The variance explained in the plot is 83%.

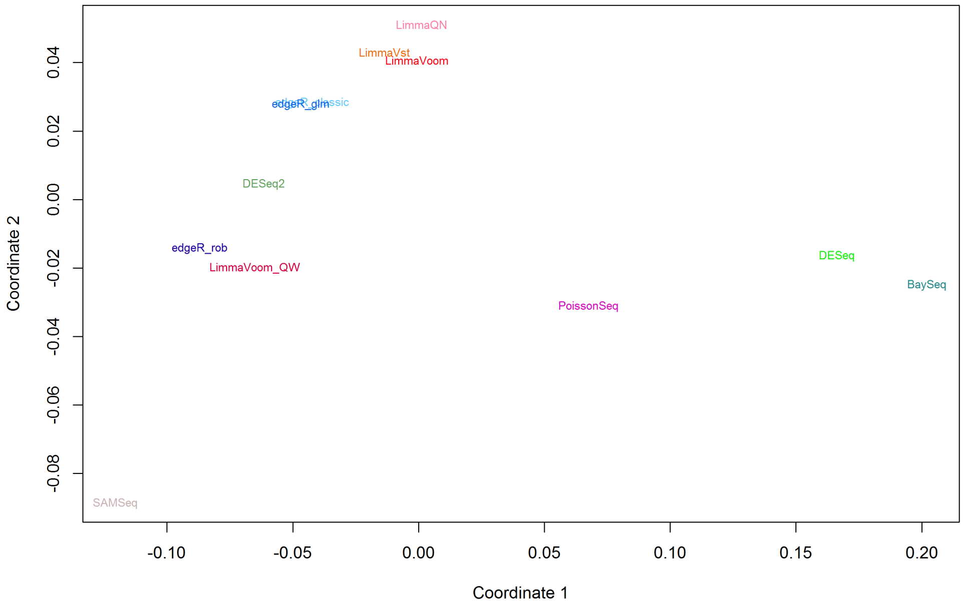


Figure 3.4: MDS plot DE overlap between methods for Bottomly dataset - Construction of the MDS-plot is based on the chi-square distance. The variance explained in the plot is 81%.

Subsequently we investigate the relationship between **the adjusted p-values as generated by the different methods**. Figure 3.5 gives the Spearman rank correlations between the adjusted p-values for each pair of methods. The scatterplots plot the $-\log_{10}$ transformed adjusted p-values of the methods versus each other. The monotone transformation does not impact the Spearman rank correlations, but it allows to zoom in on the smallest p-values. Similar figures have been produced for the other datasets: S22 (Bottomly), S34 (Hammer), S46 (GTEx) end S58 (Rapaport).

Overall, we see that the ranking of the genes is very similar between the methods. For the CRC AZA and Hammer datasets, SAMSeq and limmaQN are the exceptions with considerably lower Spearman rank correlations. This confirms the conclusions we drew based on the MDS-plots. The Rapaport dataset leads to similar conclusions, but the lower correlation of limmaQN with the other methods is now a bit less outspoken. For the Bottomly and GTEx datasets, baySeq has the lowest rank correlation with the other methods. Note that for all datasets, the relation between the $-\log_{10}$ transformed adjusted p-values of baySeq, PoissonSeq and SAMSeq with the other methods might look a bit strange. For these methods, the adjusted p-values (or posterior probabilities for baySeq) are based on either resampling from the data or performing permutations. As such, the adjusted p-values of these methods go to zero in a less 'gradual' manner compared to the other methods.

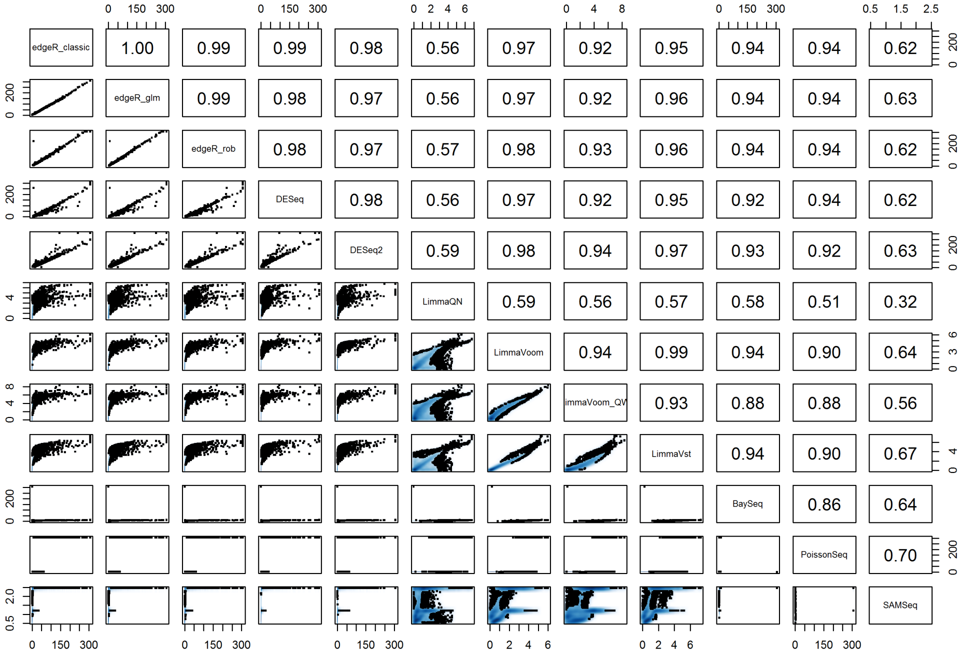


Figure 3.5: Scatter matrix $-\log_{10}$ adjusted p-values between methods for the CRC AZA dataset. Because of the transformation, the smallest p-values are displayed in the upper right corner of each scatter plot. The correlations above the diagonal are Spearman rank correlations. limmaQN and SAMSeq have the lowest rank correlation with the other methods.

The **overlap of the top 100 most significant genes** is a different way of analyzing whether the different methods have a similar ranking of the genes by significance. However, this time we only include the 100 most significant genes instead of the full set of genes. Table 3.4 shows the pairwise overlap in the top 100 most significant genes for the CRC AZA dataset (similar tables for the other datasets are given by S5, S8, S11 and S14). For the CRC AZA dataset, the edgeR and DESeq variants return almost exactly the same set of 100 most significant genes, while the overlap of these methods with the limma-based methods, baySeq and PoissonSeq is lower but still in the range of thirty to seventy genes. For the Hammer, Bottomly and GTEx datasets the overlap between the edgeR variants and the DESeq variants is still high but already further away from perfect overlap, while the overlap between these methods with the other methods (except SAMSeq) is higher. At first sight, it seems like SAMSeq is the method that overlaps the least with the other methods in terms of the 100 most significant genes. It needs to be noted however that this is at least partly driven by the fact that SAMSeq is characterized by large ties in the lowest p-values. For all datasets, 500 to 5000 genes correspond with the lowest SAMseq p-value. As a result, if we just take the first 100 genes of the complete list of genes that correspond with this lowest adjusted p-value, the overlap with the other methods whose adjusted p-value better allows to discriminate the top 100 most significant genes is limited. For the same reason the Rapaport dataset is less suited to compare overlap of the top 100 most significant genes. There are so many DE genes in this dataset and the differential expression is so outspoken that for many of the methods there are many more than 100 genes for which the adjusted p-value is zero (or at least smaller than the smallest positive integer different from zero that can be read by our machine). Further research is needed to compare the overlap in the set of most significant genes if we increase the size of this set to 200, 300 etc. genes. For us the most important takeaway is that for SAMSeq, the adjusted p-value is insufficient to differentiate the most differentially expressed genes.

In terms of **LFC estimation** (Figure 3.6), the relationship between the estimated values is close to the identity line for most combinations of the methods. DESeq2 and limmaVst form exceptions on this general rule. When plotting the estimated LFC for these latter methods versus the other methods, the cloud of points seems to be located between the identity line and the x-axis (when DESeq2 or limmaVst is put on the y-axis). This is another illustration that overall DESeq2 and limmaVst tend to estimate lower values for the LFC compared to the other methods. Another exception is SAMSeq for which the cloud of points seems to follow three parallel lines. The lines parallel with the identity line correspond to the extreme LFC-estimates for some of the low count genes we observed earlier in the MA-plots. limmaQN is also less correlated with the other methods for the CRC AZA dataset, but this is less the case for the other dataets. The estimated LFC scattermatrices for the other datasets can be found in Figures S23, S35, S47 and S59.

	edgeR_classic	edgeR_glm	edgeR_rob	DESeq	DESeq2	LimmaQN	LimmaVoom	LimmaVoom_QW	LimmaVst	BaySeq	PoissonSeq	SAMSeq
edgeR_classic	100	99	97	90	93	36	58	61	53	74	39	1
edgeR_glm	99	100	98	90	92	37	58	62	54	75	40	1
edgeR_rob	97	98	100	90	91	37	59	63	54	77	41	1
DESeq	90	90	90	100	85	37	61	65	54	79	41	1
DESeq2	93	92	91	85	100	35	61	64	50	70	42	1
LimmaQN	36	37	37	37	35	100	58	40	78	48	31	4
LimmaVoom	58	58	59	61	61	58	100	75	72	60	40	2
LimmaVoom_QW	61	62	63	65	64	40	75	100	56	61	37	1
LimmaVst	53	54	54	54	50	78	72	56	100	62	31	3
BaySeq	74	75	77	79	70	48	60	61	62	100	37	3
PoissonSeq	39	40	41	41	42	31	40	37	31	37	100	2
SAMSeq	1	1	1	1	1	4	2	1	3	3	2	100

Table 3.4: Overlap top 100 most significant genes for the CRC AZA dataset. The numbers off-diagonal indicate the number of the top 100 most significant genes that are shared between each pair of methods. The low values for SAMSeq are at least partly driven by the large ties in the SAMseq adjusted p-values

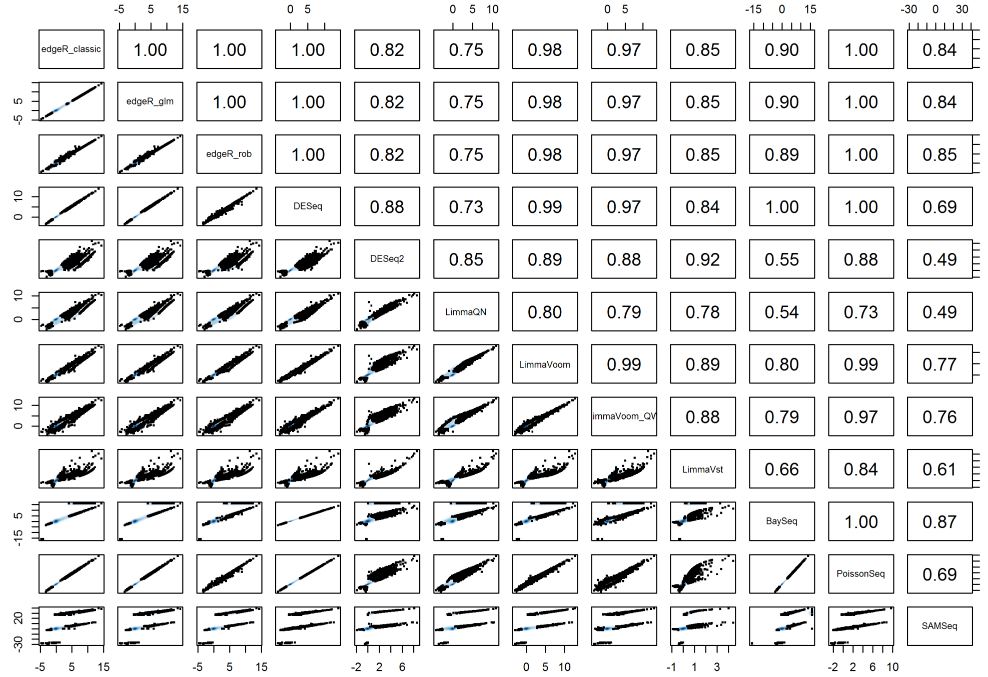


Figure 3.6: Scatter matrix estimated LFC between methods for the CRC AZA dataset. The correlations above the diagonal are Pearson correlations. DESeq2, limmaQN, limmaVst and SAMSeq are less correlated with the other methods.

For each of the datasets the normalized counts per sample and per condition have been visualized for the top 3 most significant genes (Figures S12, S24, S36, S48, S60). In order to better understand which methods are more likely to classify which type of genes as DE, Summary Table 3.5 which includes some **key metrics on the top 100 most significant genes** for each of the methods and each of the datasets, has been constructed according to the methodology we described in section 2.3.2. The data suggest following trends:

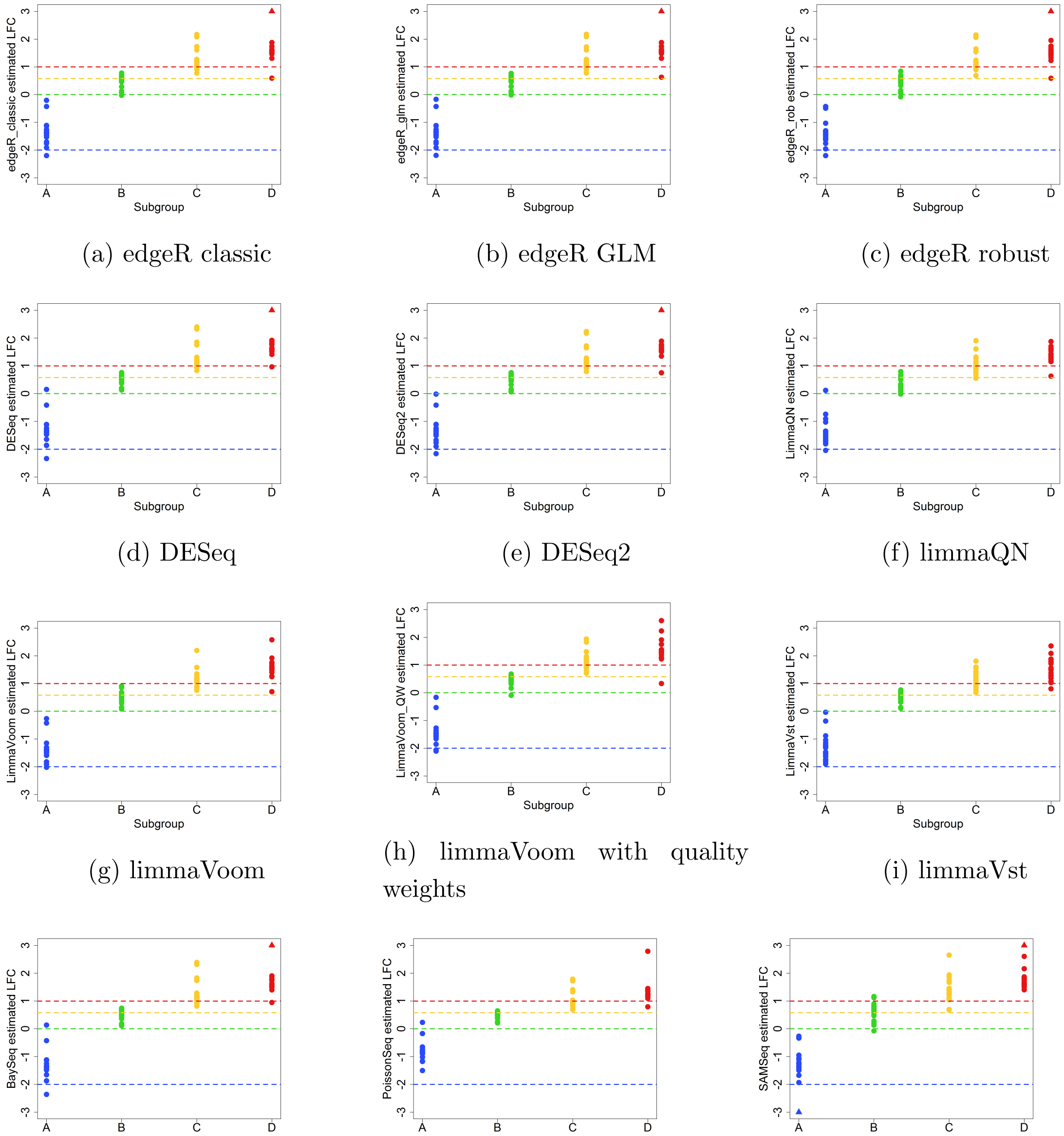
- limmaQN and limmaVst have the most genes with low expression in one of the conditions in their top 100 of most significant genes
- PoissonSeq, the only model that assumes a Poisson distribution, tends to have more genes with high dispersion in its top 100 of most significant genes, in particular for the the Bottomly and GTEx datasets which are the two datasets with the highest biological variation
- in datasets with low biological variation, baySeq tends to select genes with high counts even if the fold change is limited. This is particularly well illustrated by Figure S12j: in the CRC AZA dataset, the gene that is detected as most significant by baySeq is also the gene with the highest average cpm despite the fact that its estimated LFC is small (0.0204). This gene, EEF1A1, is by the other methods even not classified as differentially expressed.

As mentioned before, a methodological issue is that there are a lot of ties in the adjusted p-values of SAMSeq and to a lesser extent for PoissonSeq. A same phenomenon is observed for most of the methods in the Rapaport dataset: for eight out of the twelve methods, the method's top 100 most significant genes have exactly the same p-value. As such, these top 100 genes are only a subset of the method's most significant genes as they are way more than 100 genes with exactly the same p-value. This impedes a fair comparison of the methods in terms of the metrics currently used. Further research would be needed to investigate these differences in more depth.

A final comment of this section concerns the spike-in genes in the Rapaport dataset. Our original intent was to benchmark performance of the methods based on these 92 spike-in genes for which the log expression change is predefined. However, when calculating performance metrics based on the ERCC-genes, we found terrible results for all methods (Figure S-61) with FDRs between 12% and 20% and FPRs exceeding 50%. When plotting the estimated LFCs versus the theoretical LFCs, we found that all methods were consistently overestimating the LFC as shown in Figure 3.7. In the literature we found that the ERCC spike-in read counts are not independent from the biological factor of interest (Risso et al., 2014): the proportion of reads mapping to spike-ins may be consistently larger in one condition than in another. As a consequence, ERCC spike-ins are not suited to assess performance of DE methods.

Dataset	Method	Nr Low Expr	UQ Disp	Nr top 10 Expr	Unique adj. pval
	minimal value	0	0	0	0
	maximal value	100	100	10	100
CRCAZA	edgeR_classic	69	3	0	92
	edgeR_glm	70	3	0	92
	edgeR_rob	68	3	0	91
	DESeq	69	0	0	95
	DESeq2	63	3	0	91
	LimmaQN	89	6	0	60
	LimmaVoom	60	0	0	40
	LimmaVoom_QW	42	0	0	55
	LimmaVst	81	2	0	62
	BaySeq	77	1	3	100
	PoissonSeq	61	18	0	1
	SAMSeq	35	27	1	1
Bottomly	edgeR_classic	39	23	0	96
	edgeR_glm	39	24	0	95
	edgeR_rob	39	22	0	99
	DESeq	37	26	0	96
	DESeq2	29	22	0	96
	LimmaQN	45	24	0	77
	LimmaVoom	36	20	0	80
	LimmaVoom_QW	34	20	0	87
	LimmaVst	42	22	0	86
	BaySeq	39	21	0	100
	PoissonSeq	29	34	0	32
	SAMSeq	22	12	0	1
Hammer	edgeR_classic	2	0	0	95
	edgeR_glm	2	0	0	97
	edgeR_rob	2	0	0	96
	DESeq	0	2	2	97
	DESeq2	0	0	2	96
	LimmaQN	28	32	0	1
	LimmaVoom	0	0	0	14
	LimmaVoom_QW	0	0	0	18
	LimmaVst	10	7	0	1
	BaySeq	5	2	4	100
	PoissonSeq	2	9	0	1
	SAMSeq	4	13	0	1
GTEx	edgeR_classic	13	37	0	89
	edgeR_glm	15	38	0	94
	edgeR_rob	12	32	0	98
	DESeq	12	24	0	90
	DESeq2	11	33	0	93
	LimmaQN	11	22	0	67
	LimmaVoom	7	13	0	74
	LimmaVoom_QW	7	15	0	72
	LimmaVst	10	17	0	64
	BaySeq	7	30	0	100
	PoissonSeq	7	64	0	16
	SAMSeq	3	12	0	1
Rapaport	edgeR_classic	6	0	0	1
	edgeR_glm	6	0	0	1
	edgeR_rob	11	0	0	1
	DESeq	10	0	0	1
	DESeq2	0	0	0	1
	LimmaQN	16	0	1	56
	LimmaVoom	0	0	2	46
	LimmaVoom_QW	0	0	2	54
	LimmaVst	17	0	1	47
	BaySeq	2	0	0	1
	PoissonSeq	14	19	0	1
	SAMSeq	4	12	0	1

Table 3.5: Overview key metrics 100 most significant genes by dataset and method; 'Nr Low Expr' and 'UQ Disp' express respectively the number of top 100 genes with low expression in one condition and with a dispersion factor in the upper quartile of the dataset's disperion distribution. 'Nr top 10 Expr' indicates how many of the dataset's top 10 most expressed genes occur in the top 100 of most significant genes. 'Unique adj. pval' indicates the number of unique adjusted p-values in the top 100 of most significant genes.



(a) edgeR classic

(b) edgeR GLM

(c) edgeR robust

(d) DESeq

(e) DESeq2

(f) limmaQN

(g) limmaVoom

(h) limmaVoom with quality weights

(i) limmaVst

(j) baySeq

(k) PoissonSeq

(l) SAMSeq

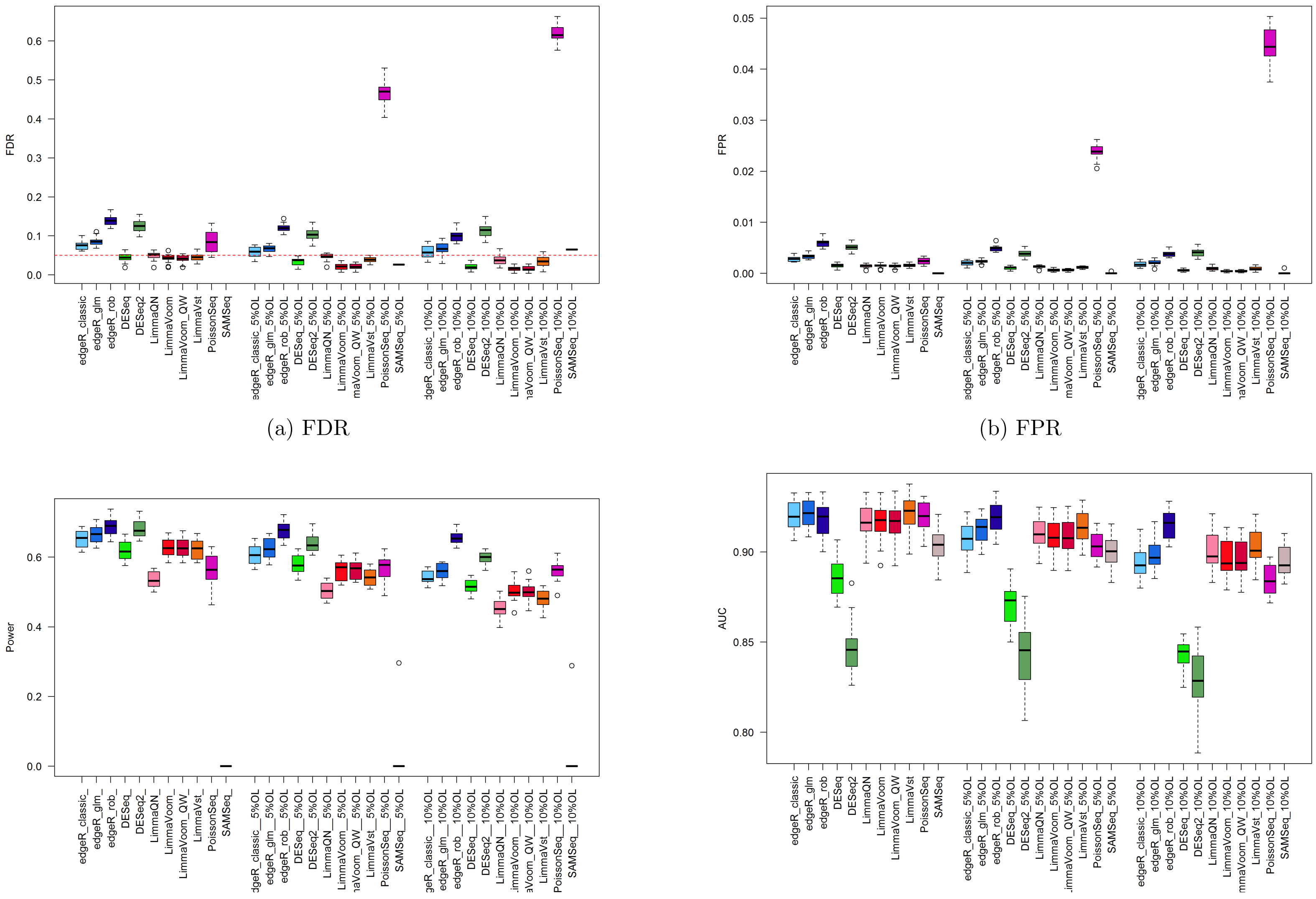
Figure 3.7: LFC estimates for the ERCC spiked-in genes in the Rapaport dataset. The ERCC genes consist of four subgroups each having a different theoretical LFC: -2 (A), 0 (B), 0.58 (C) and 1 (D). The estimated LFCs for the individual genes are shown by the dots (or triangles for values that would fall outside the plotting window), the dashed lines represent the theoretical LFC for the respective groups. All methods seem to overestimate the LFC.

3.2 Simulations

While the concordance analysis helps to understand similarities and dissimilarities between methods, it does not give insight in which of the methods perform best. We use simulations to understand the different methods' performance in terms of relevant metrics as FDR, FPR, power and AUC. This is done under varying circumstances: we investigate the impact of the fraction of outliers, fraction of differentially expressed genes, (a)symmetry in the data, number of samples per condition, LFC of the DE genes and average expression level.

Figure 3.8 gives a comparison of methods in function of the **fraction of outliers**. Overall we see that, except for PoissonSeq, the FDR and FPR remain stable or even slightly decrease with the fraction of outliers. The power and the AUC are negatively impacted by the number of outliers. At the level of the methods, following observations are made:

- In general, the order of the methods according to a particular performance metric does not change dramatically in function of the fraction of outliers.
- In terms of the FDR, DESeq and the limma-based methods are around the nominal level when there are no outliers and go below the nominal level in case of outliers. edgeR classic and edgeR glm are slightly above the nominal FDR when there are no outliers and around the nominal FDR with outliers. Robust edgeR and DESeq2 are consistently above the nominal FDR, with values exceeding 10%. PoissonSeq is also above the nominal FDR, with FDR values quickly inflating with the percentage of outliers.
- The order of the methods for the FPR is similar as for the FDR, for all levels of the percentage of outliers.
- Looking at power, it is observed that robust edgeR and to a lesser extent DESeq2 outperform the other methods. This effect becomes more outspoken when there are more outliers. Note that in the setting of Figure 3.8 (5+5 samples, pDiff=0.05 and pUp=0.50), SAMSeq has no power whatsoever.
- With respect to AUC most methods have comparable performance, except for DESeq and DESeq2, in case there are no outliers. As soon as outliers are introduced, robust edgeR has a higher AUC compared to the other methods. Two further remarks need to be made here, which are nicely illustrated by Figure 3.9. First, DESeq and DESeq2 clearly have a lower AUC compared to the other methods. At least part of the reason is that DESeq and DESeq2 exclude part of the genes from analysis and that we assigned these genes a p-value of 1 in order to calculate the AUC for the same set of features across all methods. Second, despite the fact that SAMSeq has no power in the setting we used, it does have an AUC that is in line with the other methods. This is explained by the fact that the DE genes have p-values higher than the nominal 5%, but still tend to have p-values that are smaller than for the non-DE genes.



(a) FDR

(b) FPR

(c) Power

(d) AUC

Figure 3.8: Performance metrics by method (at nominal FDR 5%) for 0%, 5% and 10% outliers (5+5 samples, pDiff=0.05, pUp=50%) - Limma-based methods best control the FDR; EdgeR robust has highest power, but shows weak FDR control. PoissonSeq is most sensitive to presence of outliers.

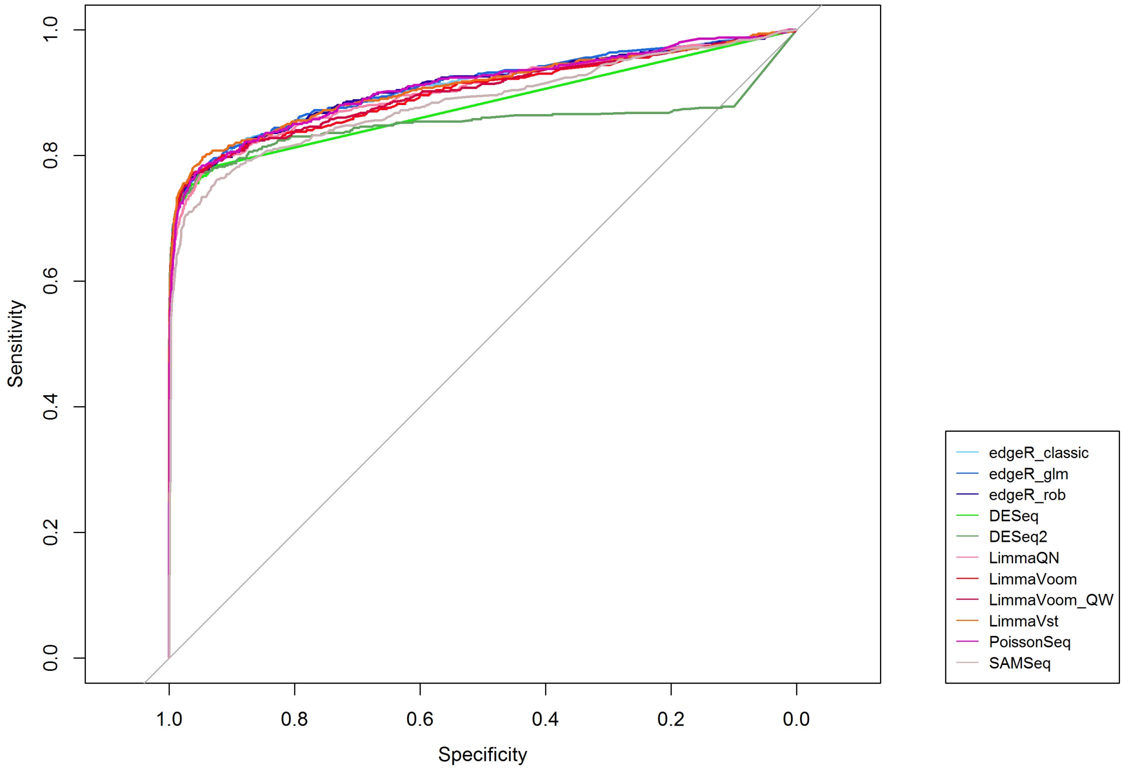
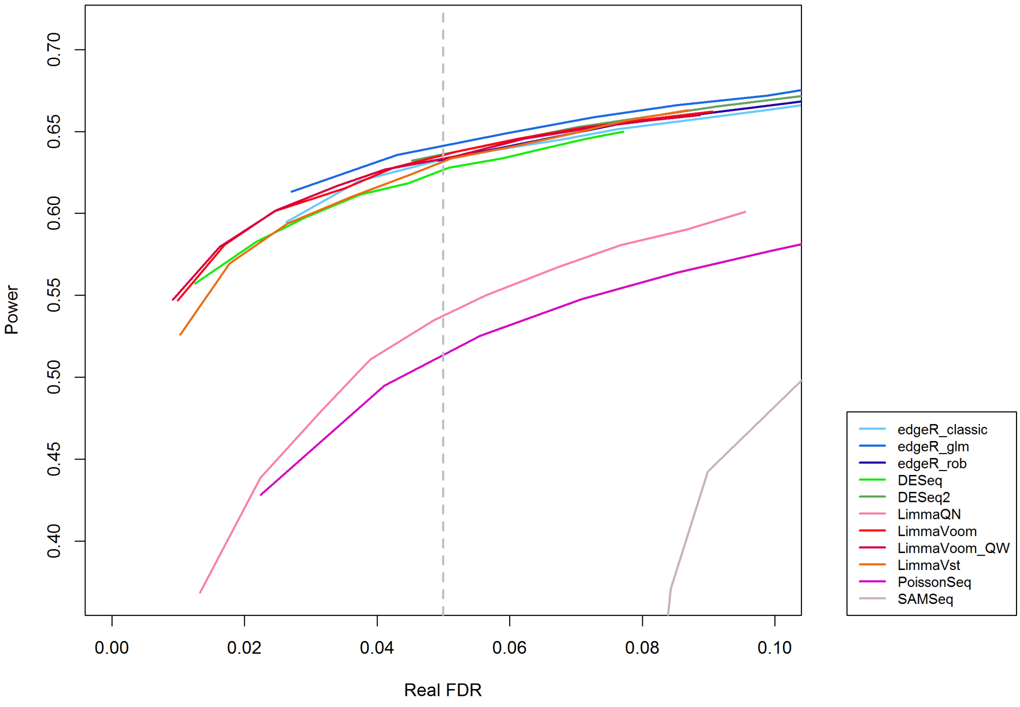
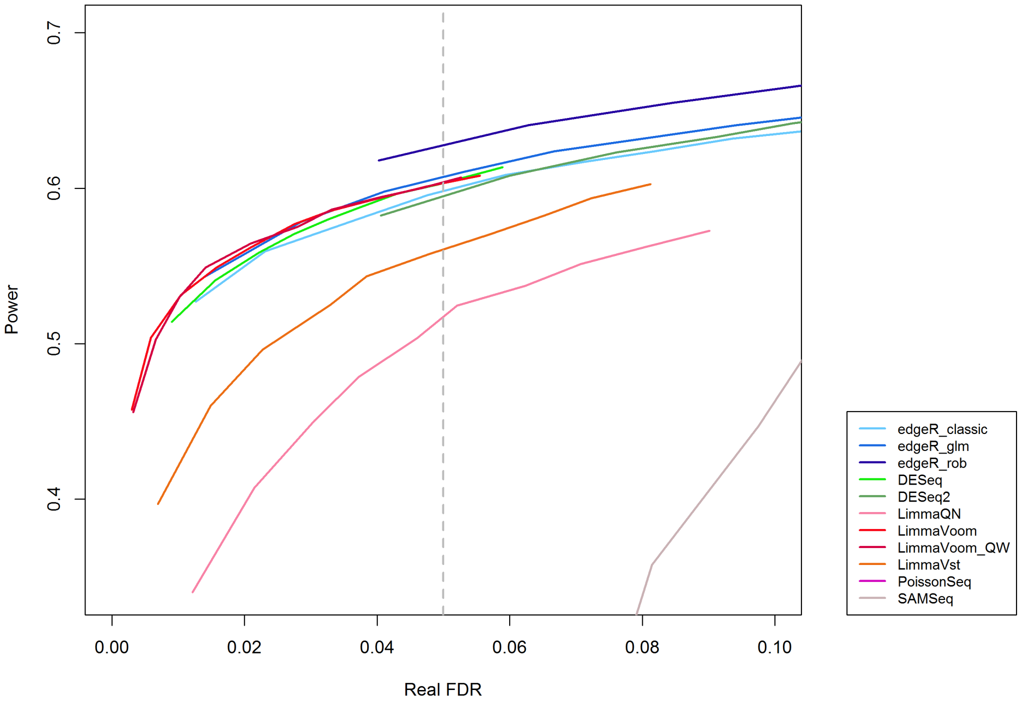


Figure 3.9: Typical ROC curve by method (5+5 samples, pDiff=0.05, pUp=0.5, without outliers) - Similar AUCs across methods are observed, except for DESeq and DESeq2. Lower AUCs of these latter methods are at least partly driven by the fact that they exclude a number of genes from analysis.

Our observation that limma-based methods best control the FDR has been made before (Soneson and Delorenzi, Love et al.). In addition, our analysis confirms the finding that edgeR robust has a higher power, especially in the presence of outliers, but that this comes at the cost of a too high real FDR (Zhou et al.). As not all methods control the FDR equally well, the comparison of the power is not completely fair. Figure 3.10 shows the power in function of the real FDR. Without outliers the edgeR-based methods, DESeq, DESeq2 and the limma-based methods (except for limmaQN) have a comparable real FDR-power tradeoff, with a slight advantage for edgeR glm. As soon as there are outliers, robust edgeR has a markedly better power for the same real FDR compared to the other methods which supports the claim of Zhou et al.



(a) Without outliers



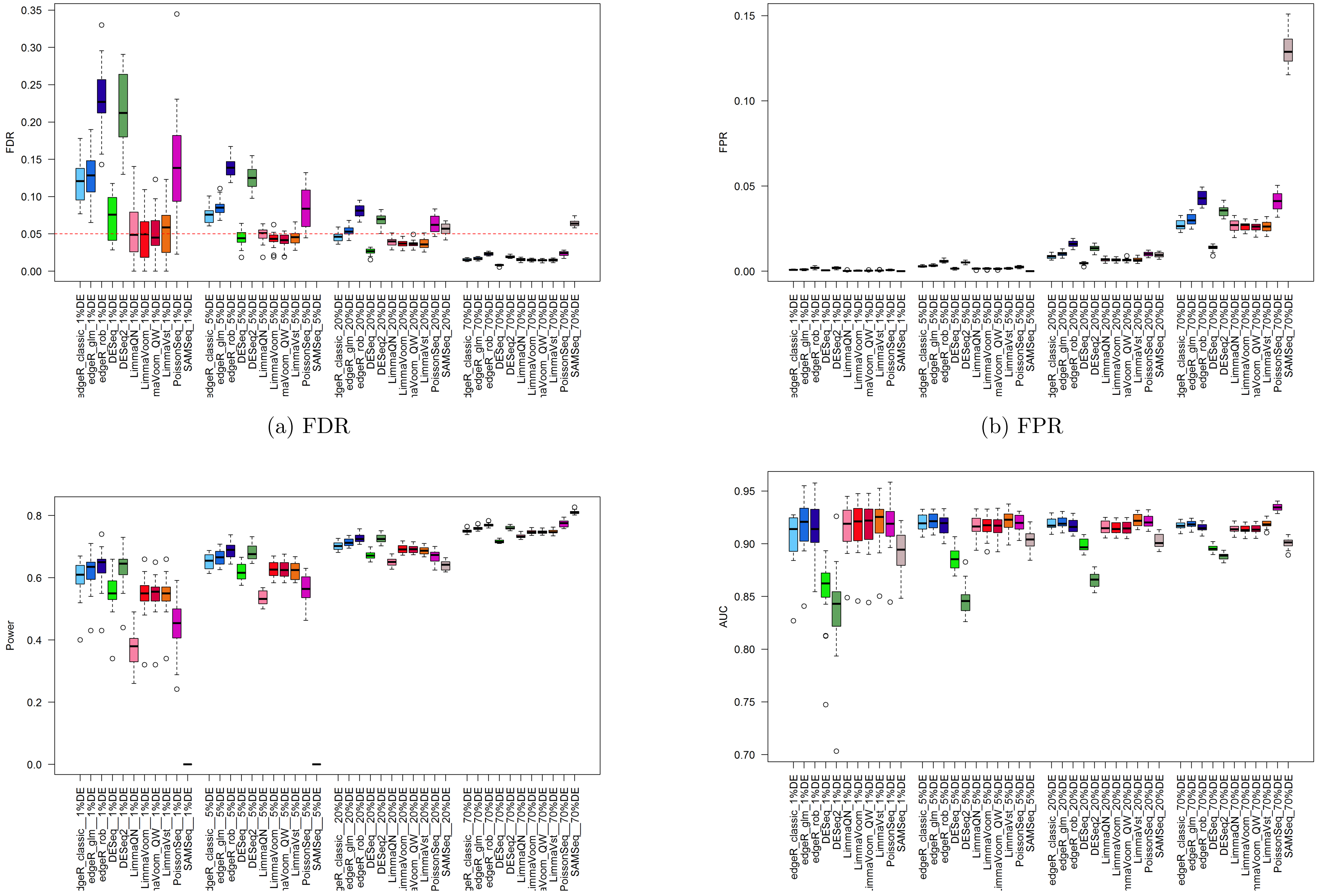
(b) With 5% outliers

Figure 3.10: Trade-off real FDR vs. power by method (5+5 samples, pDiff=0.05, pUp=0.5) - Extremes of each line correspond with nominal FDR of respectively 1% and 10%. Robust edgeR has a markedly better real FDR-power trade-off compared to the other methods in the presence of outliers.

Figure 3.11 helps us to assess **the impact of the fraction of differentially expressed genes**. The global picture tells us that a higher percentage of truly DE genes goes with a lower FDR and a higher power but also with a higher FPR, while the AUC is stable. With respect to the methods' (relative) performance, we observe that

- The methods' order of performance for each of the criteria does not change a lot in function of the fraction of differentially expressed genes.
- If the percentage of DE genes is set to 70%, all methods have an FDR that is below the nominal FDR, except for SAMSeq which stays slightly above 5%. However, for small fractions of truly DE genes, the limma-based methods and to some extent DESeq are the only methods that are around the nominal FDR. For percentages of DE genes as small as 1%, edgeR classic, edgeR glm and PoissonSeq on average have FDRs larger than 10%, robust edgeR and DESeq2 even have FDRs larger than 20%.
- The order of performance in terms of the FPR is preserved for various fractions of DE genes. Only for SAMSeq the FPR has a stronger increase in function of the percentage of DE genes compared to the other methods.
- For small (1% and 5%) to medium (20%) fractions of DE genes, robust edgeR and DESeq2 have the highest power. SAMSeq which has no power for low fractions of DE genes, has the highest power for very large fractions of DE genes (70%) but this comes at the cost of a higher FDR an FPR than the other methods. Most important is that for large fractions of DE genes the differences in power between the different methods become way smaller than they are for small fractions of DE genes.
- A similar effect is observed for the AUC where DESeq and DESeq2 close the gap with the other methods for larger fractions of DE genes.

As the fraction of DE genes has a clear impact on FDR control, authors might be tempted to play with this parameter until they have a value for which their method shows good performance. This might explain why Love et al. claims that DESeq2 has the highest power of the methods that control the FDR, while we stated earlier that DESeq2 is characterized by poor FDR control. Indeed, in their simulation exercise Love et al. use 20% of DE genes. For this percentage of DE genes, we also find that the FDR of DESeq2 starts to come close to the nominal FDR, while having the best power (together with edgeR robust). However, for small fractions of DE genes, DESeq2 is far from having good FDR control. It is especially in this setting of low fractions of DE genes, that the limma-based methods show markedly better FDR control than the other methods. As such, it does not surprise that it is exactly this setting that Law et al. pick to demonstrate performance of their limmaVoom method.



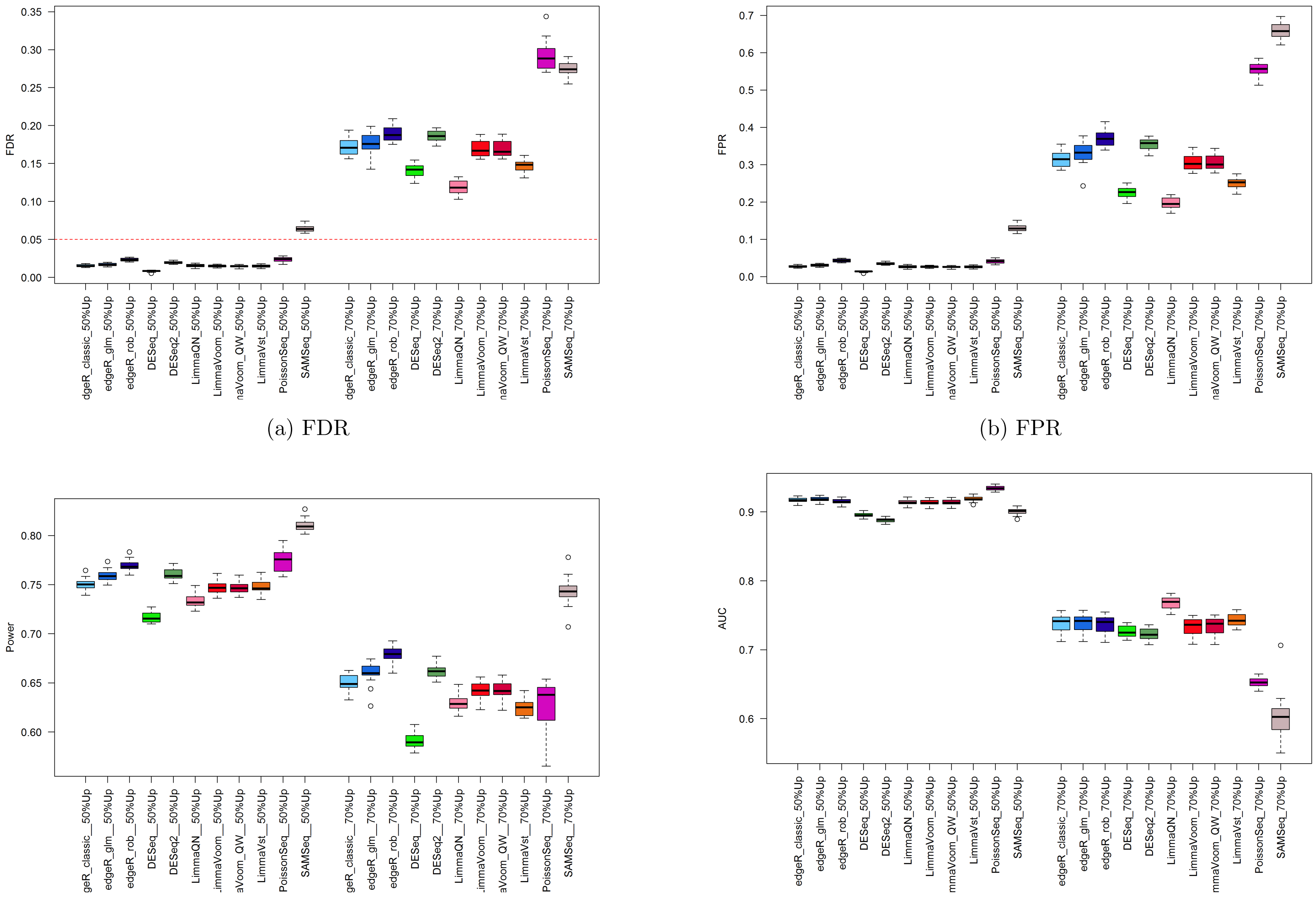
(a) FDR

(b) FPR

(c) Power

(d) AUC

Figure 3.11: Performance metrics by method (at nominal FDR 5%) for pDiff 1%, 5%, 20% and 70% (5+5 samples, pUp=0.5, no outliers) - Fraction of DE genes is highly determining for overall performance of methods. The higher the fraction of outliers, the smaller the difference in performance between the methods.



(a) FDR

(b) FPR

(c) Power

(d) AUC

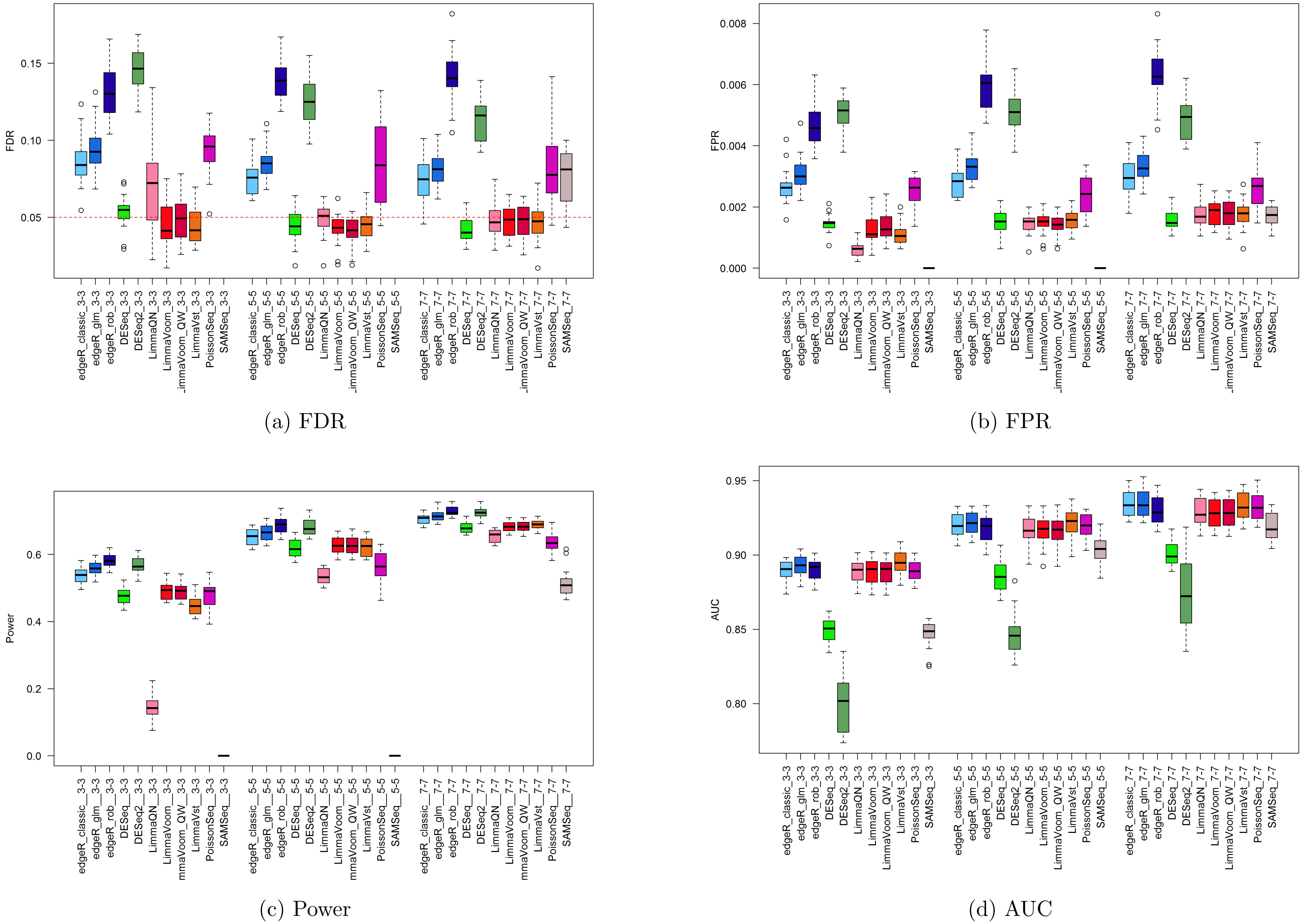
Figure 3.12: Performance metrics by method (at nominal FDR 5%) for pUp 50% and 70% (5+5 samples, pDiff=0.70, no outliers) - Asymmetry in the direction of differential expression has a negative impact on performance if the fraction of DE genes is high.

If the percentage of DE genes is medium to low, **symmetry or asymmetry** in the direction of differential expression does not seem to have an impact on performance, on the absolute level of the metrics nor on the methods' order of performance (Figure S153). However, when the percentage of DE genes is high, asymmetry in the direction of differential expression has a major negative impact on performance as can be seen in Figure 3.12. The same observation has been made by Soneson and Delorenzi. The reason for this is that in this setting normalization methods start to fail as these methods assume that the majority of genes is not differentially expressed and that there is symmetry in the direction of differential expression. Soneson and Delorenzi states that SAMSeq is more robust against this asymmetry, but this is not confirmed by our simulations. We find that in this setting, SAMSeq indeed has the best power, but it also has an unacceptable FDR.

Figure 3.13 shows that in general **the number of samples** only has a limited impact on the absolute level of the FDR and FPR. The power and the AUC on the other hand experience a positive impact of a growing number of samples. Again the methods' relative performance remains largely unchanged by an increasing number of samples. A few exceptions are:

- The FDR of DESeq and DESeq2 slightly decreases with the number of samples. As such DESeq2 is the worst performer in terms of FDR control in the 3+3 samples case, but this position is taken over by robust edgeR in case of more samples as the FDR of robust edgeR does not seem to change with the number of samples.
- Something similar is observed for the FPR. While the FPR of DESeq and DESeq2 is stable when the number of samples increases, the FPR of the other methods slightly increases. In this way, robust edgeR surpasses DESeq2 as the method with the highest FPR when the number of samples increases.
- While limma QN is far behind the other methods in terms of power in the case of 3+3 samples, it closes this gap when the number of samples increases. Similarly, in the setting described (pDiff=0.05, pUp=50% and no outliers), SAMSeq only starts to have power as of 7+7 samples.
- DESeq and DESeq2 get closer but stay behind the other methods in terms of AUC with an increasing number of samples.

We also tested whether the behavior of the methods differs in function of the **fold change** (Figure 3.14). Obviously, DE genes with a higher fold change have a higher probability to be detected compared to DE genes with a lower fold change. While there are no big shifts in the methods' relative power with increasing fold changes, we do observe some subtle changes. First, DESeq2 performs slightly better than edgeR classic and edgeR glm for lower absolute fold changes (1.5-2), but slightly worse for higher absolute fold changes (>3), which is consistent with the observations of Zhou et al. Second, limmaQN catches up with the other limma-based methods when the fold changes get larger. Third, for PoissonSeq, the increase



(a) FDR

(b) FPR

(c) Power

(d) AUC

Figure 3.13: Performance metrics by method (at nominal FDR 5%) for 3+3, 5+5, 7+7 samples (pDiff=0.05, pUp=50%, no outliers) - The order of edgeR robust and DESeq2 in terms of FDR and FPR changes when going from 3 to 5 replicates per condition. SAMSeq only starts to have power as of 7 replicates per condition. Similarly, the performance of limmaQN is extremely sensitive to the number of samples.

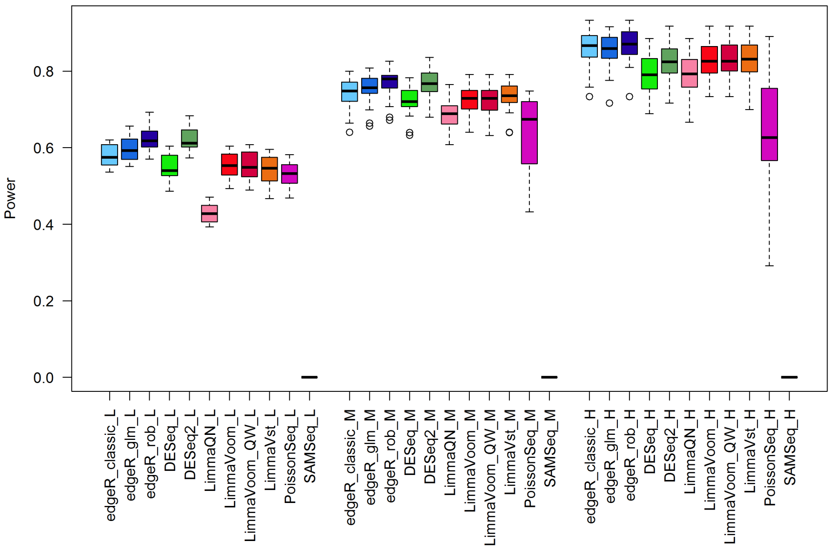
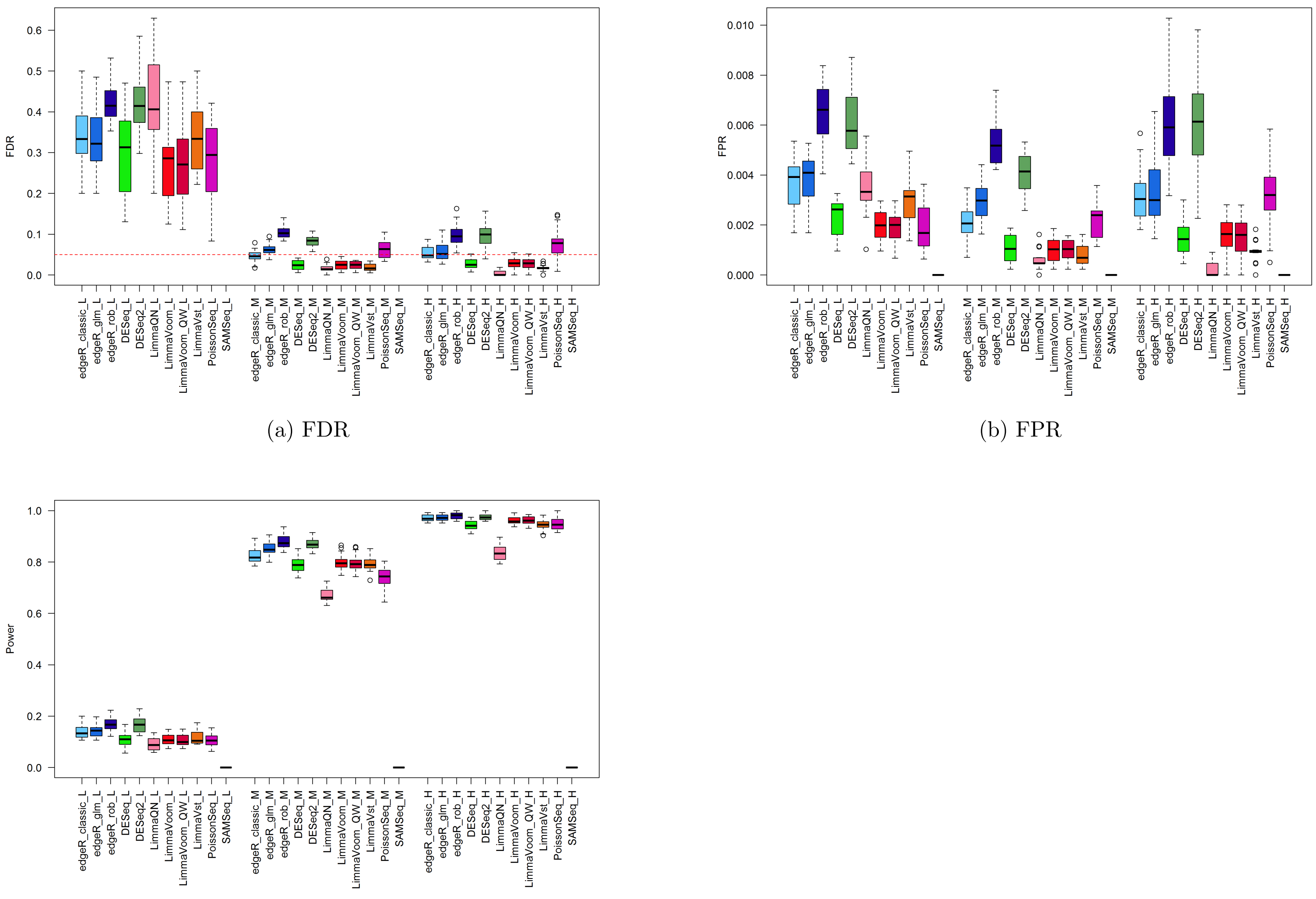


Figure 3.14: Power by method (at nominal FDR 5%) for DE genes with low (1.5-2), medium (2-3) and high (>3) true fold change (5+5 samples, pDiff=0.05, pUp=0.5, without outliers) - DESeq2 is one of the more powerful methods for low fold changes, but looses this edge for larger fold changes; LimmaQN has lower power compared to the other methods for low fold changes.

in power is limited and the spread over the different simulated datasets grows with increasing fold changes.

Figure 3.15 analyzes the results from the perspective of the **average count** of the genes. The general picture, making abstraction of the individual methods, shows that the FPR does not vary a lot for different levels of the average count. On the contrary, the FDR is significantly higher for the lowest count group (<10 cpm), strongly exceeding the nominal 5% for all methods. The power jumps from levels below 20% for the low count group, to levels around 80% for the medium count group (10-100 cpm) and values around 100% for the highest count group (>100 cpm). As for most of the other factors we discussed, the average count has only limited impact on the methods' order of performance. Only notable exception is limmaQN that has the highest FDR for low counts, but is the most conservative for the largest counts. This, together with our previous observation of poor limmaQN performance for small numbers of replicates, corresponds with Rapaport et al. who state that the largest difference between limmaQN and limmaVoom is in the number of false detections at low counts and in the sensitivity as a function of the number of samples.



(a) FDR

(b) FPR

(c) Power

Figure 3.15: Performance metrics by method (at nominal FDR 5%) for low (<10 cpm), medium (10-100 cpm) and high (>100 cpm) average count genes (5+5 samples, pDiff=0.05, pUp=0.5, without outliers) - Absolute performance is highly dependent on the average expression level of the genes, but except for limmaQN relative performance does not change a lot with the average count.

So far we have been evaluating the methods by the FDR, the FPR, the power and the AUC. All these metrics depend on the methods' ability to correctly classify genes as DE or non-DE. Related to this, but still something different, is the methods' **ability to correctly estimate the (log) fold changes**. Table 3.6 shows that the average bias is reasonably close to zero for all methods, but that there are quite some differences in terms of the RMSE. LimmaVst and DESeq2 have a RMSE that is considerably lower than the other methods. However, we don't agree with Love et al. who claims, based on a comparison of the RMSE between methods, that DESeq2 more precisely estimates LFCs versus the other methods. The low RMSE of DESeq2 and limmaVst is driven by a low RMSE for the non-DE genes as visualized in Figure 3.16. In the setting used, non-DE genes represent 95% of all genes and thus have a way larger impact on the total RMSE compared to the DE genes. When visualizing the true versus the estimated LFC for all genes (Figure 3.17), we see that DESeq2 and limmaVst structurally underestimate the absolute LFC in the low count group and to a lesser extent in the medium count group. The overall average bias of DESeq2 and limmaVst is close to zero because the biases for overexpressed genes and underexpressed genes neutralize each other. However, it is clear that for every individual non-zero level of the true LFC, DESeq2 and limmaVst do have a bias. PoissonSeq is subject to the same problem, but here the issue is larger for the medium and the large count groups. The high RMSE of SAMSeq is the consequence of a number of low-count genes for which the estimated LFC is extremely far of the truth. All the other methods show unbiased estimation of the LFC for all true LFC levels. Obviously, the higher the average count, the more accurate LFC estimation becomes.

	Bias			RMSE		
Method	Non-DE	DE	Total	Non-DE	DE	Total
edgeR_classic	7.29e-04	3.25e-03	8.55e-04	0.423	0.429	0.424
edgeR_glm	7.32e-04	3.25e-03	8.58e-04	0.423	0.429	0.424
edgeR_rob	1.70e-04	2.88e-03	3.05e-04	0.458	0.458	0.458
DESeq	2.54e-05	2.98e-03	1.73e-04	0.432	0.436	0.432
DESeq2	1.79e-04	-3.47e-03	-3.75e-06	0.159	0.475	0.188
LimmaQN	6.37e-05	3.72e-03	2.47e-04	0.390	0.407	0.391
LimmaVoom	3.16e-04	4.51e-03	5.26e-04	0.429	0.435	0.429
LimmaVoom_QW	3.57e-04	4.46e-03	5.62e-04	0.428	0.434	0.428
LimmaVst	2.36e-04	-1.04e-03	1.73e-04	0.133	0.545	0.178
PoissonSeq	1.50e-04	3.21e-03	3.03e-04	0.299	0.452	0.309
SAMSeq	8.67e-03	1.74e-02	9.11e-03	2.846	3.169	2.863

Table 3.6: Average bias and RMSE of LFC-estimation by method and type of genes (5+5 samples, pDiff=0.05, pUp=0.5, without outliers) - DESeq2 and limmaVst have the lowest overall RMSE, but there is a big difference between the RMSE for DE and non-DE genes

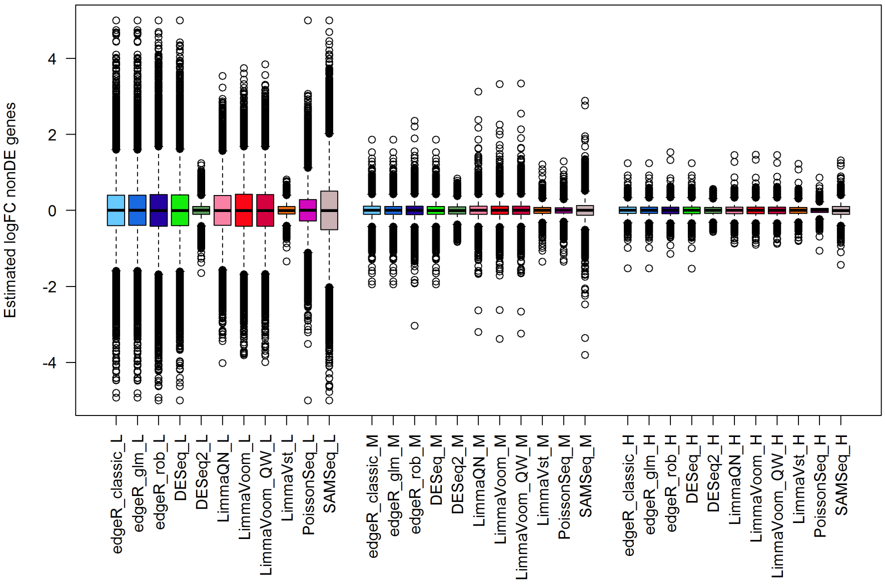
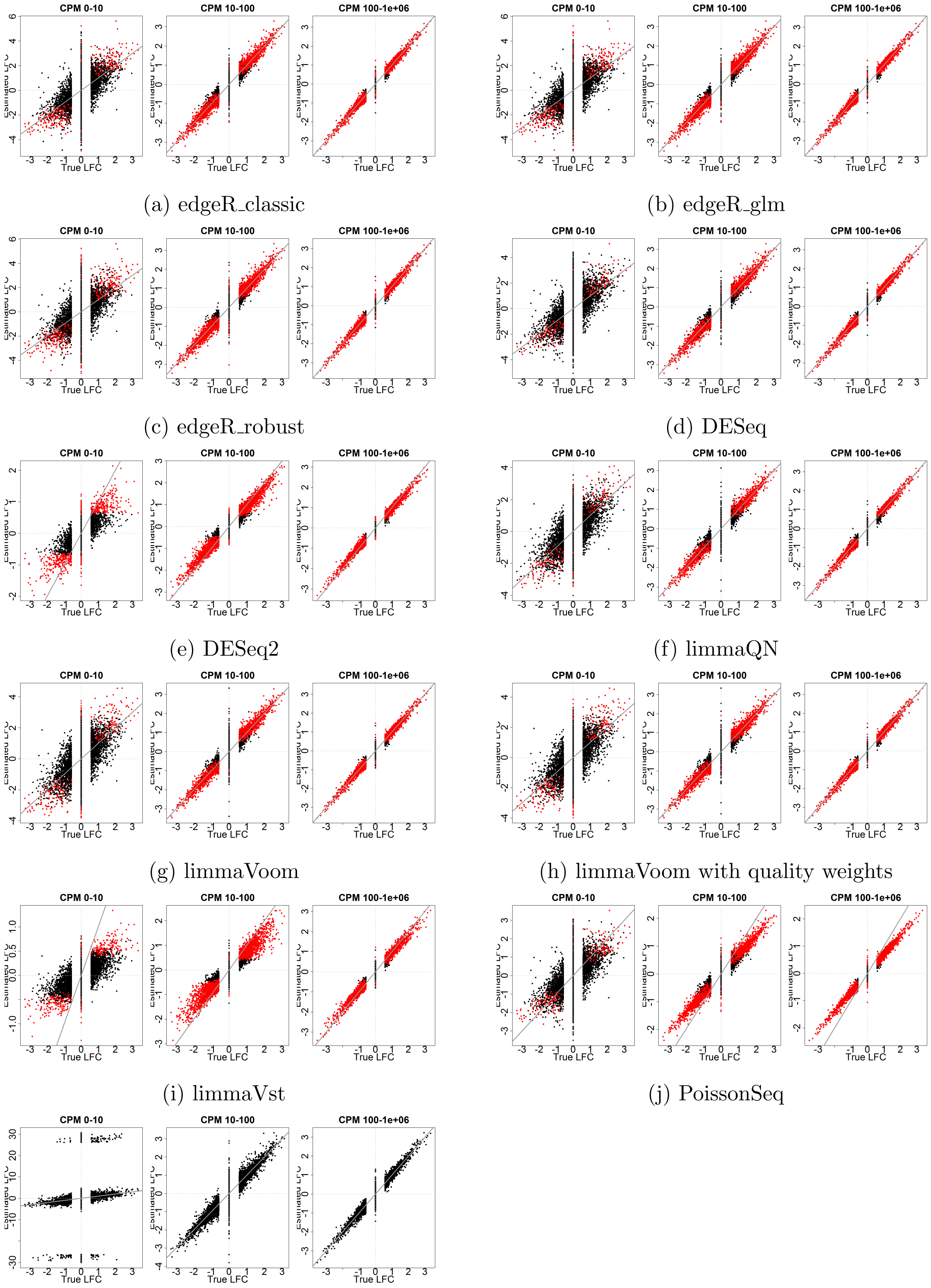


Figure 3.16: Distribution of the estimated LFC of non-DE genes by method for low (<10 cpm), medium (10-100 cpm) and high (>100 cpm) average count (5-5 samples, pDiff=0.05, pUp=0.5, without outliers) - All methods give unbiased estimates for the non-DE genes. For low counts, DESeq2 and limmaVst have a considerably lower variance



(a) edgeR_classic

(b) edgeR_glm

(c) edgeR_robust

(d) DESeq

(e) DESeq2

(f) limmaQN

(g) limmaVoom

(h) limmaVoom with quality weights

(i) limmaVst

(j) PoissonSeq

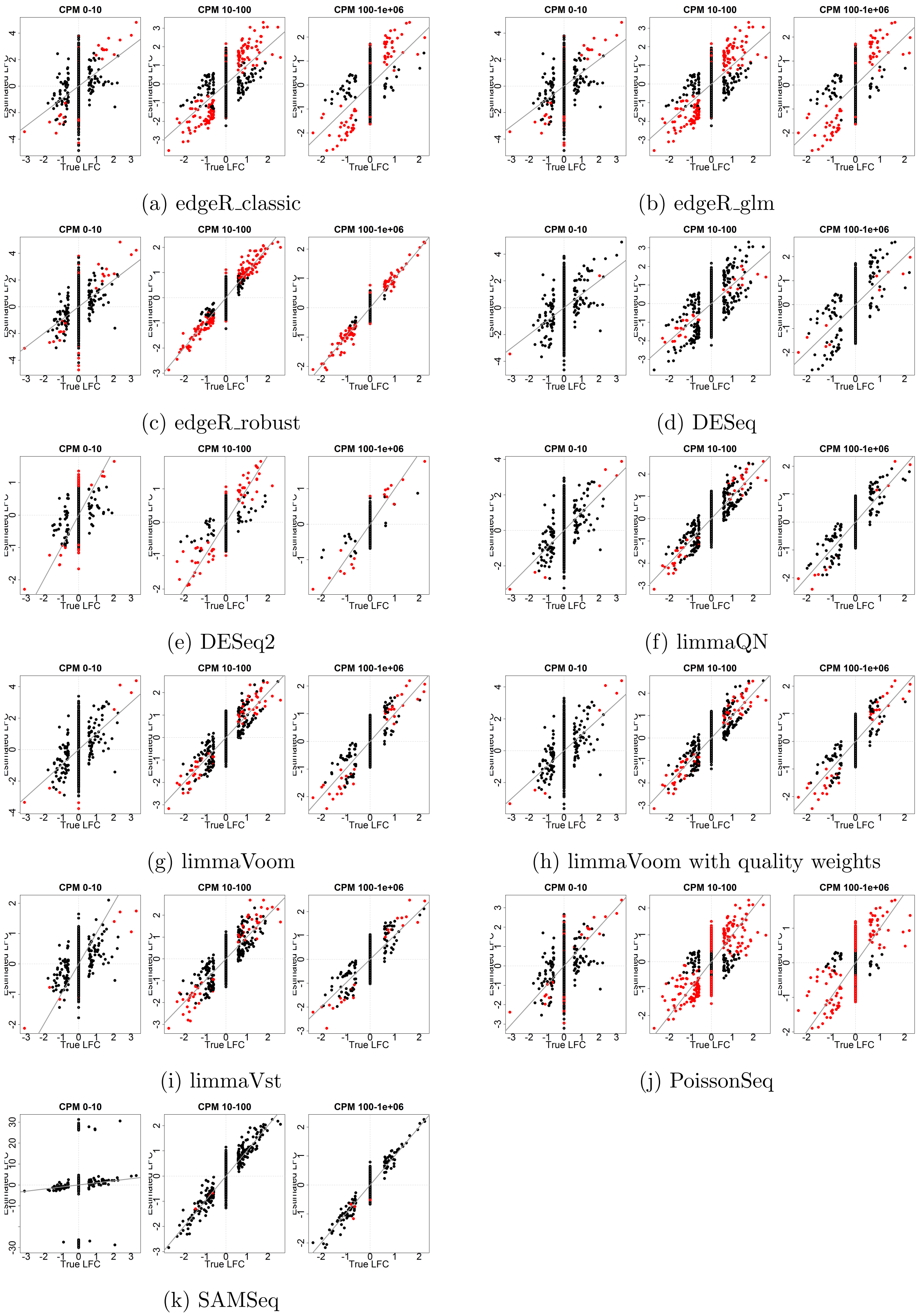
(k) SAMSeq

Figure 3.17: True vs. estimated LFC of all genes by method for low (<10 cpm), medium (10-100 cpm) and high (>100 cpm) average count (5-5 samples, pDiff=0.05, pUp=0.5, without outliers); Red points indicate genes with an adjusted p-value <0.05, black points indicate genes with an adjusted p-value≥0.05 - DESeq2 and limmaVst structurally underestimate the absolute LFC for low and medium counts. PoissonSeq has a similar issue for medium and high counts

As we are particularly interested in how accurately the **LFC** is estimated for **genes with outliers**, we create a similar picture for the outlier genes only. Table 3.7 reports the average bias and the RMSE for outlier genes. Comparison with Table 3.6 shows that the RMSE is way higher for outlier genes, but that the RMSE increase for edgeR robust is limited compared to the increase for the other methods. Figure 3.18 visualizes the estimated versus the true LFC of the outlier genes for different levels of average counts. For the low count groups there is quite some scatter around the identity line for all methods. For the medium and high count groups the differences are more outspoken. For these groups edgeR robust clearly shows less scatter around the identity line, confirming our conclusion that edgeR robust estimates the LFC of outlier genes more accurately compared to the other methods. It also classifies more of the DE genes correctly (i.e. higher power), but this comes at the price of some false discoveries. limmaQN, limmaVoom and limmaVoom with quality weights give a decent LFC estimation for outlier genes. These methods have a lower power, but also a lower FDR; A final remark is that for DESeq2, quite some of the outlier genes are excluded from the analysis, especially in the high count group.

	Bias			RMSE		
Method	Non-DE	DE	Total	Non-DE	DE	Total
edgeR_classic	1.13e-03	2.89e-02	2.43e-03	0.952	0.964	0.953
edgeR_glm	1.13e-03	2.89e-02	2.43e-03	0.952	0.964	0.953
edgeR_rob	1.70e-03	2.06e-02	2.58e-03	0.517	0.526	0.517
DESeq	1.39e-03	2.90e-02	2.68e-03	0.955	0.971	0.956
DESeq2	-3.33e-04	-3.91e-04	-3.36e-04	0.466	0.706	0.480
LimmaQN	2.62e-03	2.15e-02	3.50e-03	0.590	0.613	0.591
LimmaVoom	2.81e-03	3.14e-02	4.15e-03	0.621	0.647	0.622
LimmaVoom_QW	2.38e-03	3.06e-02	3.70e-03	0.620	0.647	0.621
LimmaVst	1.27e-03	7.75e-03	1.57e-03	0.474	0.609	0.481
PoissonSeq	1.17e-03	2.37e-02	2.22e-03	0.662	0.752	0.667
SAMSeq	4.89e-02	8.09e-02	5.04e-02	2.993	3.414	3.014

Table 3.7: Average bias and RMSE of LFC-estimation by method for outlier genes only (5+5 samples, pDiff=0.05, pUp=0.5, 5% outliers) - RMSE is higher for outlier genes. edgeR robust experiences the smallest increase in the RMSE.



(a) edgeR_classic

(b) edgeR_glm

(c) edgeR_robust

(d) DESeq

(e) DESeq2

(f) limmaQN

(g) limmaVoom

(h) limmaVoom with quality weights

(i) limmaVst

(j) PoissonSeq

(k) SAMSeq

Figure 3.18: True vs. estimated LFC of outlier genes (both DE and non-DE) by method for low (<10 cpm), medium (10-100 cpm) and high (>100 cpm) average count (5-5 samples, pDiff=0.05, pUp=0.5, 5% outliers). Red points indicate genes with an adjusted p-value <0.05, black points indicate genes with an adjusted p-value≥0.05 - For medium and high counts, robust edgeR estimates the LFC more accurately compared to the other methods

Conclusions

In this master thesis we compared several statistical methods for differential gene expression analysis based on RNA-seq data. We focused on the most frequently used methods that are available in R: edgeR (classic, glm and robust), DESeq and DESeq2, limma-based methods (limmaVoom with and without quality weights, limmaQN and limmaVst), baySeq, PoissonSeq and SAMSeq. Our key purpose was to come to a sound recommendation on which methods perform best under varying conditions. Whereas earlier research states that no method is optimal under all circumstances, we claim that generally speaking **two methods outperform the others**: edgeR robust and limmaVoom (with or without quality weights). The choice for one of both depends on the performance criterion that is considered as most important by the researcher.

EdgeR robust, which augments edgeR glm with a methodology to downweight outlying observations in order to increase robustness, proves to have the **best trade-off between power and real FDR in the presence of outliers**. If outliers are present, edgeR robust has a higher power compared to the other methods for a given level of the real FDR. However, a major disadvantage of edgeR robust is that it is characterized by **poor FDR control**. Especially for datasets where the percentage of truly differentially expressed genes is low, the real FDR is way higher than the nominal FDR.

limma-based methods allow better control of the FDR. In particular **limmaVoom (with and without quality weights)** shows strong performance: the trade-off between power and real FDR is good and the **real FDR is around or below the nominal FDR**. This is remarkable as this method only transforms the data and calculates weights before inputting it into the limma pipeline that was originally developed for analysis of microarray data. In our simulations limmaVoom with and without sample quality weights performed equally well. However, our simulations did not really test the impact of differences in sample quality, the setting that limmaVoom with quality weights was designed for. As such, in situations with differences in sample quality, limmaVoom with quality weights might even be the better option.

In general, **we would not recommend to use the other methods**: either they don't show a clear strength over the other methods or they even have an important disadvantage. In the absence of outliers, edgeR classic, edgeR glm, DESeq, DESeq2 and LimmaVst have a similar trade-off between power and real FDR compared to edgeR robust and limmaVoom. However, they have a less favorable trade-off compared to edgeR robust when outliers are present, with limmaVst experiencing the largest deterioration in the trade-off. Except for limmaVst, these methods also underperform versus limmaVoom in terms of FDR control: edgeR classic, edgeR glm and DESeq2 are too liberal, while DESeq is overly conservative. In addition DESeq2 and LimmaVst give biased estimates of the log fold changes for low and medium gene counts. The real FDR-power trade-off of limmaQN is less advantageous compared to the other methods. In addition, relative to the other limma-based methods, it has a lower power for smaller fold changes and a smaller number of samples and a higher real FDR for low count genes. baySeq has a tendency to overselect genes with the highest average counts and is computationally slow. PoissonSeq tends to overselect genes with high dispersion and it sees the number of false discoveries quickly inflating when outliers are introduced. In addition it results in biased estimates for the log fold changes of genes with medium and high counts. SAMSeq shows poor performance when the number of samples is small. In the literature, twelve replicates per condition is mentioned as a minimum before the nonparametric approach starts to work well. This is a major drawback, knowing that in practice the number of replicates per conditions is often not larger than two or three.

A number of factors have not been considered in our simulation study and could be used as **areas for further research**. First, while our concordance analysis indicated that the behaviour of the methods depends on the characteristics of the underlying dataset, this has not been further explored in the simulation analysis where we only performed simulations based on the CRC AZA dataset. Repeating the simulations on other datasets with different characteristics of which the amount of biological variation and differences in sample quality are probably the most important ones, could be useful to further refine the insights from this master thesis. Second, as part of the concordance analysis we explored whether different methods are more likely to select specific types of genes. Analyzing the set of genes classified as differentially expressed by the different methods could help to gain a deeper understanding hereof. Third, in our simulations, counts have been generated by means of a negative binomial distribution. It could be of interest to check if our results still hold if the counts are generated by different mechanisms. Finally, the scope of our research was limited to single-factor designs and mRNA. As such, further research could evaluate the performance of the methods for multi-factor designs and for other types of RNA.

Bibliography

S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10), 2010.

P. L. Auer and R. W. Doerge. A Two-Stage Poisson Model for Testing RNA-Seq Data. *Statistical Applications in Genetics and Molecular Biology*, 10(1), 2011.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, 57(1):289–300, 1995. ISSN 0035-9246.

D. Bottomly, N. A. R. Walter, J. E. Hunter, P. Darakjian, S. Kawane, K. J. Buck, R. P. Searles, M. Mooney, S. K. McWeeney, and R. Hitzemann. Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays. *Plos One*, 6 (3), 2011.

J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11, 2010.

N. Cloonan, A. R. R. Forrest, G. Kolle, B. B. A. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, A. J. Robertson, A. C. Perkins, S. J. Bruce, C. C. Lee, S. S. Ranade, H. E. Peckham, J. M. Manning, K. J. McKernan, and S. M. Grimmond. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, 5(7):613–619, 2008.

Y. Di, D. W. Schafer, J. S. Cumbie, and J. H. Chang. The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. *Statistical Applications in Genetics and Molecular Biology*, 10(1), 2011.

M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloe, C. Le Gall, B. Schaeffer, S. Le Crom, M. Guedj, F. Jaffrezic, and F. S. Consortium. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013.

P. Hammer, M. S. Banck, R. Amberg, C. Wang, G. Petznick, S. Luo, I. Khrebtukova, G. P. Schroth, P. Beyerlein, and A. S. Beutler. mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain. *Genome Research*, 20(6):847–860, 2010.

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

C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2), 2014.

N. Leng, J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. G. Smits, J. D. Haag, M. N. Gould, R. M. Stewart, and C. Kendziorski. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 29(8):1035–1043, 2013.

J. Li and R. Tibshirani. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research*, 22 (5, SI):519–536, 2013.

J. Li, D. M. Witten, I. M. Johnstone, and R. Tibshirani. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, 13(3):523–538, 2012.

R. Liu, A. Z. Holik, S. Su, N. Jansz, K. Chen, H. S. Leong, M. E. Blewitt, M.-L. Asselin-Labat, G. K. Smyth, and M. E. Ritchie. Why weight? modelling sample and observational level variability improves power in rna-seq analyses. *Nucleic Acids Research*, 2015.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 2014.

J. Lu, J. Tomfohr, and T. Kepler. Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, 6, 2005.

J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, 2008.

D. J. McCarthy, Y. Chen, and G. K. Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10): 4288–4297, 2012.

T. T. Perkins, R. A. Kingsley, M. C. Fookes, P. P. Gardner, K. D. James, L. Yu, S. A. Assefa, M. He, N. J. Croucher, D. J. Pickard, D. J. Maskell, J. Parkhill, J. Choudhary, N. R. Thomson, and G. Dougan. A Strand-Specific RNA-Seq Analysis of the Transcriptome of the Typhoid Bacillus Salmonella Typhi. *Plos Genetics*, 5(7), 2009.

J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289): 768–772, 2010.

F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. Mason, N. Socci, and D. Betel. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome Biology*, 14(9):R95, 2013.

D. Risso, J. Ngai, T. P. Speed, and S. Dudoit. The role of spike-in standards in the normalization of rna-seq. In S. Datta and D. Nettleton, editors, *Statistical Analysis of Next Generation Sequencing Data*. Springer International Publishing, 2014.

M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3), 2010.

M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

M. D. Robinson and G. K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332, 2008.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

C. Soneson and M. Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, 14(1):91, 2013.

S. Tarazona, F. Garcia-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa. Differential expression in RNA-seq: A matter of depth. *Genome Research*, 21(12):2213–2223, 2011.

The GTEx project. `http://www.gtexportal.org/home/`. Accessed: 2015-06-30.

C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, 31(1):46+, 2013.

V. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–5121, 2001.

M. A. Van De Wiel, G. G. R. Leday, L. Pardo, H. Rue, A. W. Van Der Vaart, and W. N. Van Wieringen. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, 14(1):113–128, 2013.

L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26(1):136–138, 2010.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

H. Wu, C. Wang, and Z. Wu. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, 14(2):232–243, 2013.

H. Xiong, J. B. Brown, N. Boley, P. J. Bickel, and H. Huang. De-fpca: Testing gene differential expression and exon usage through functional principal component analysis. In S. Datta and D. Nettleton, editors, *Statistical Analysis of Next Generation Sequencing Data*. Springer International Publishing, 2014.

X. Zhou, H. Lindsay, and M. D. Robinson. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Research*, 42(11), 2014.