Viktor Jonsson did his doctoral studies in mathematical statistics at the Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg.

ISBN 978-91-629-0089-2 (PRINT)
ISBN 978-91-629-0090-8 (PDF)

Ph.D. thesis

Statistical Analysis and Modelling of Gene Count Data in Metagenomics | Viktor Jonsson 2017

PH.D. THESIS

Statistical Analysis and Modelling of Gene Count Data in Metagenomics

Viktor Jonsson

DEPARTMENT OF MATHEMATICAL SCIENCES

UNIVERSITY OF GOTHENBURG

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Statistical analysis and modelling of gene count data in metagenomics

Viktor Jonsson

CHALMERS | GÖTEBORGS UNIVERSITET

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden 2017

Statistical analysis and modelling of gene count data in metagenomics
Viktor Jonsson
Göteborg 2017
ISBN 978-91-629-0089-2 (printed version)
ISBN 978-91-629-0090-8 (electronic version)

Cover illustration: *Microorganisms* by Kajsa Andersson

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Typeset with LaTeX
Printed by Ineko AB
Göteborg, Sweden 2017

Statistical analysis and modelling of gene count data in metagenomics

Viktor Jonsson

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg

Abstract

Microorganisms form complex communities that play an integral part of all ecosystems on Earth. Metagenomics enables the study of microbial communities through sequencing of random DNA fragments from the collective genome of all present organisms. Metagenomic data is discrete, high-dimensional and contains excessive levels of both biological and technical variability, which makes the statistical analysis challenging.

This thesis aims to improve the statistical analysis of metagenomic data in two ways; by characterising the variance structure present in metagenomic data, and by developing and evaluating methods for identification of differentially abundant genes between experimental conditions. In Paper I we evaluate and compare the statistical performance of 14 methods previously used for metagenomic data. In Paper II we implement an overdispersed Poisson model and use it to show that the biological variability varies considerably between genes. The model is used to evaluate a range of assumptions for the variance parameter, and we show that correct modelling of the variance is vital for reducing the number of false positives. In Paper III we extend the model used in Paper II to incorporate zero-inflation. Using the extended model, we show that metagenomic data does indeed contain substantial levels of zero-inflation. We demonstrate that the new model has a high power to detect differentially abundant genes. In Paper IV we suggest improvements to the annotation and quantification of gene content in metagenomic data. Our proposed method, HirBin, uses a data-centric approach to identify effects at a finer resolution, which in turn allows for more accurate biological conclusions.

This thesis highlights the importance of statistical modelling and the use of appropriate assumptions in the analysis of metagenomic data. The presented results may also guides researchers to select and further refine statistical tools for reliable analysis of metagenomic data.

Keywords: metagenomics, statistical modelling, hierarchical statistical models, gene ranking, overdispersion, zero-inflation, false discovery rate, receiver operating characteristic curves.

List of papers

Paper I : **Jonsson, V.**, Österlund, T., Nerman, O., Kristiansson, E. (2016). Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC genomics*, 17(1), 1. DOI: 10.1186/s12864-016-2386-y

Paper II : **Jonsson, V.**, Österlund, T., Nerman, O., Kristiansson, E. (2016). Variability in metagenomic count data and its influence on the identification of differentially abundant genes. *Journal of Computational Biology*, ahead of print. DOI:10.1089/cmb.2016.0180

Paper III : **Jonsson, V.**, Österlund, T., Nerman, O., Kristiansson, E. (2017). A zero-inflated model for improved inference of metagenomic gene count data. *Manuscript*.

Paper IV : Österlund, T., **Jonsson, V.**, Kristiansson, E. (2017). HirBin: High-resolution identification of differentially abundant functions in metagenomes. *Submitted*.

Additional published paper not included in thesis

Paper V : Bengtsson-Palme, J., Boulund, F., Edström, R., Feizi, A., Johnning, A., **Jonsson, V. A.**, Karlsson, F. H., Pal, C., Pereira, M. B., Rehammar, A., Sanchez, J., Sanli, K. and Thorell, K. (2016). Strategies to improve usability and preserve accuracy in biological sequence databases. *Proteomics*, 16: 2454-2460. DOI: 10.1002/pmic.201600034

Author contributions

Paper I : Participated in study design, implemented the previously proposed methods in R, wrote the framework for data resampling and methods evaluation, performed all comparisons and analysed the results, created all figures, drafted and edited the manuscript.

Paper II : Participated in study design, developed and implemented the Bayesian models, analysed model performance and convergence, performed the analysis on real data, performed all comparisons and analysed the results, created all figures, drafted and edited the manuscript.

Paper III : Participated in study design, developed and implemented the zero-inflated model, analysed model performance and convergence, performed the analysis on real data, performed all comparisons and analysed the results, created all figures, drafted and edited the manuscript.

Paper IV : Participated in study design, assisted in data analysis and the interpretation of results, edited the manuscript.

Acknowledgements

First I would like thank the people who have collaborated with me and who made this work possible. First, **Erik Kristiansson**, my supervisor, you have been a source of constant inspiration through your unceasing enthusiasm. Your ability to see the progress that I have made when I have felt stuck, and your ability to put everything into perspective, have been invaluable for me. You taught me how to do science but you also gave me so much more. I would also like to thank **Olle Nerman**, my co-supervisor, for always being there to answer the trickiest questions. Your ability to quickly grasp a problem and ask the right questions has helped me many a times. I also thank you for your timely anecdotes, with topics ranging from growing up in Värmland to phase-type distributions. Finally, I would like to thank **Tobias Österlund**, your clear minded insights have helped me refine countless of arguments. In sharing your own experiences of science you made my own journey that much easier.

My years at the department would not have been so bright if not for the wonderful people here. Some people deserve a special mention for everything that they have done for me. **Mariana Pereira**, from Berlin to the Rover you were always there for me, thank you for making my life a bit more crazy (in a good way!), and always being there to brighten my day. **Fredrik Boulund**, by my side since almost forever, now my personal online support group for anything nerdy, and when you play your guitar know that I am your number one fan. **Anna Rehammar**, I could not have dreamt of a better office mate, our long discussions on life and science has meant more to me than you know, I only wish you were here a little more often. **Anna Johnning**, you personify the word epic, you have an incredible ability to see people, and I also thank you for introducing me to ever so many cultural activities. **Johan Bengtsson-Palme**, a better source of energy and enthusiasms is hard to find, I also remember a certain python project featuring kaminmannen, thank you for that. **Fanny Berglund**, from pole vaulting to minecraft you never cease to amaze me, good times were had both when teaching together and at lunch. **Olle Elias**, you might be even nerdier than me, no kappa. **Jonatan Kallus**, a better lunch companion for deep discussions is hard to find. Thank you all!

I would also like to thank some of the remaining, current and former, members of the bioinformatics group: **Marija Cvijovic, Anders Sjögren, Marina Axelson-Fisk, Emma Wijkmark, Johannes Borgqvist, and Henrik Imberg**. Special thanks to, **Mikael Wallroth** for being the perfect master thesis student.

For exciting collaborations in the far north of Sweden I extend my thanks to **Martin Rosvall** and **Alcides Viamontes Esquivel**.

I would also like to thank all my fellow PhD students throughout the years for making this a wonderful place to work, encouraging me to take hour-long coffee breaks by providing excellent company, or spontaneously bumping in to me and saying hi on late evenings. There are far too many fantastic people and simply not enough time for all of you. Thank you: **Henrike, Claes, Ivar, Sandra, Alexei, Anton, Magnus Ö, Anders M, Tobias A, John, Magnus R, Dawan, Peter, Jonas, Oscar, Emil, Hossein, Richard, Jakob, Malin PF, Åse, Åsa, Linnea, Marco, Ronny, Kristin, Carl, Valentina, Roza, Anders H, Zuzana, Anna P, Elizabeth, Elin, Andreas, Christoffer, Malin Ö, Medet, Oscar Hammar, Fredrik L, Adam, Urban, Emilio and everyone I might have missed!**. A special thanks to everyone else at the department, junior or senior, that have assisted me in countless ways during my time here.

To the GoBig crowd, thank you for providing a wonderful arena for the exchange ideas beyond the walls of the maths department. Some of you have already been mentioned above but I would also like to extend thanks to: **Leif, Franscesco, Kaisa, Kemal, Amir, Robert, Chandan, and Fredrik**. Many of which also collaborated on the database paper.

A special thanks to my friends outside the department who continuously have reminded me that there is a world beyond the math building. I will not be able to mention you all here but I would like to extend a special thanks to **Erik Sterner**, my personal guide when it comes to new adventures, both in life and science, and an endless source of energy, enthusiasm, chocolate, and pomegranate. Remember Cochran. **Lucas Gren**, my number one dance partner (sorry **Johan**), I hope for many a fruitful collaborations in the future, I especially want more breakfast meetings.

Extra special thanks to my family who has been there for me all these years. **Mom** and **Dad**, when times are rough, I always long for a trip back to home, thank you for everything. **Olle** and **Sigrid**, I could not have wished for more awesome siblings, I look forward to finally being able to spend more time with you.

Finally, **Kajsa**, you made these last few years the best of my life, thank you for being you and for helping me be me. With all my heart, I love you!

Göteborg, januari, 2017

Viktor Jonsson

Contents

Abstract **iii**

List of publications **v**

Acknowledgements **vii**

Contents **ix**

1 Background **1**

1.1 Metagenomics 1

1.2 Quantification of genes 2

1.3 Statistical challenges 3

2 Aims **7**

3 Statistical analysis of metagenomic data **9**

3.1 Identification of differentially abundant genes 9

3.2 Overview of statistical methods 10

3.3 Evaluation of statistical performance 16

4 Summary of papers **23**

4.1 Paper I 23

4.2 Paper II 26

4.3 Paper III 28

4.4 Paper IV 31

5 Conclusion and outlook **35**

Bibliography **39**

Papers I-IV

1 Background

1.1 Metagenomics

Microorganisms, such as bacteria, viruses and fungi, are ubiquitously present everywhere around us (Pace, 1997; Whitman et al., 1998). Contrary to popular belief, most microorganisms are not harmful to humans; rather, they constitute an integral part of all living ecosystems. Microorganisms are typically studied by isolating and cultivating single isolates. However, microorganisms form complex communities that contain thousands of species (Roesch et al., 2007). In addition, a vast proportion of microorganisms are difficult to cultivate using standard protocols (Schloss and Handelsman, 2005), thus making culture-dependent approaches unsuitable for capturing the intricacies of most microorganisms. Metagenomics was introduced as a culture-independent method for studying microbial communities at the genomic level (Handelsman et al., 1998). In short, the DNA of all cells present within a sampled microbial community is extracted. The DNA is then randomly sheared and sequenced, resulting in a large set of short DNA fragments, called reads. These reads represent a random sample from the metagenome, i.e. the collective genome of all organisms present in the community (Rondon et al., 2000). Metagenomics, the study of the metagenome, therefore provides a way to analyse both the structure and functional capability of metagenomic communities.

Originally, metagenomics was performed using slow and expensive Sanger sequencing (Sanger and Coulson, 1975), with the first data sets consisting only of thousands of reads (Healy et al., 1995). With the introduction of modern high-throughput sequencing methods, which have the ability to sequence multiple DNA fragments in parallel, the potential of metagenomics has considerably increased (van Dijk et al., 2014; Scholz et al., 2012). Large sequencing efforts, such as the Human Microbiome Project (Turnbaugh et al., 2007; Human Microbiome Project Consortium, 2012) and the Earth Microbiome Project (Gilbert et al., 2014), have recently generated data sets consisting of billions of reads corresponding to trillion of nucleotides (base pairs). In addition the number of applications of metagenomics to fields in the life sciences is continuously increasing. Within medicine, metagenomics has been used to link changes in the microbial communities living inside and on our bodies to common diseases, for example, type-II diabetes (Karlsson et al., 2013) and Crohn's disease (Manichanh et al., 2006). In ecology, metagenomics has been applied to study the microbial communities in a wide range of ecosystems from prairie soil (Howe et al., 2014), ocean water (Sunagawa et al., 2015; DeLong et al., 2006) and the cow rumen (Ross et al., 2013). In ecotoxicology, it has been used to understand the role of microbial ecosystems in biodegradation in waste water treatment (Fang et al., 2013), and to characterize the spread of antibiotic resistance genes in the environment (Kristiansson et al., 2011; Bengtsson-Palme et al., 2014).

1.2 Quantification of genes

Gene-centric metagenomics is the study of the functional capability of a microbial community (Hugenholtz and Tyson, 2008). A gene is a short stretch of DNA that encodes a protein which in turn performs a specific biological function. The genome of a typical bacterial species contains thousands of genes, which means that the number of genes in a microbial community is in the order of millions. Genes may exist in several variants across multiple species yet perform similar functions. To facilitate the biological interpretation, genes are often grouped together in protein domains, gene families or orthologous groups that correspond to a similar biological function in different species (e.g , eggNOG (Huerta-Cepas et al., 2015), KEGG (Kanehisa et al., 2008), TIGRFAM (Haft et al., 2013) and SEED (Overbeek et al., 2005)). The choice of resolution used, from single genes to wide groups of genes, depends on the biological question. For consistency, the term gene will be used throughout this thesis to refer to each of these different options.

Beyond the experimental steps of sample preparation, DNA extraction and sequencing, several steps are necessary to quantify the gene content of a sample (Wooley et al., 2010). The steps are summarized in Figure 1. The raw data from the sequencing machine consist of a large number of short DNA fragments called reads, which are typically 75-400 nucleotides in length depending on the technology used (van Dijk et al., 2014; Scholz et al., 2012). However, DNA sequencing is not exact and the raw reads often contain sequencing errors, for example, misidentified nucleotides or insertion of extra nucleotides. Depending on the outcome of the sequencing and the sequence technology used, this error rate can be as high as 1% of the sequenced nucleotides (Quail et al., 2012). Sequencers provide a quality score for each nucleotide that reflects the probability of an error, and this score is used to identify and remove low-quality reads (Schmieder and Edwards, 2011). Next, the aim is to determine the genetic origin of the remaining high-quality reads. To facilitate this process, each read is mapped to an annotated reference database. In the mapping step, reads are aligned to the sequences within the reference database in order to identify the possible matches (Scholz et al., 2012). A read may match to several positions within the reference database, but only the best match is generally kept.

The reference database can be constructed in several different ways. For example, the reference database can be a collection of previously characterized genes or microbial genomes. Alternatively, the database can be constructed by directly assembling the reads from the sequenced metagenome into longer stretches of DNA (Mende et al., 2012). The sequences in the reference database are typically annotated based on their biological function, which is often predicted based on sequence similarity to previously studied genes. The annotation can either be single genes or groups of genes predicted to have a similar structure or function depending on the desired level of resolution (see Paper IV). The reference database typically contains several variants of every gene, for example, different bacterial species share core functionality (Human Microbiome Project Consortium, 2012). Every read that maps to a specific gene, regardless of location in the database, is counted as an occurrence of that gene. In this way, all reads are "binned" together resulting in a list that contains the number of occurrences of each gene. The final gene counts are thus measures of the relative abundance of each gene in each sample.

1.3 Statistical challenges

An essential aspect of gene-centric metagenomics is detecting changes in relative gene abundance in relation to experimental parameters. Examples of such parameters are the health status of the human host, the temperature along a gradient and the presence or absence of an anti-microbial compound. Differentially abundant genes are detected by statistically assessing whether specific genes differ in relative abundance between communities. However, metagenomic gene count data i) is discrete and undersampled, ii) is high-dimensional, iii) contains high levels of biological and technical variability and iv) is often represented by few biological replicates. These characteristics make the statistical analysis of metagenomic data challenging on many levels.

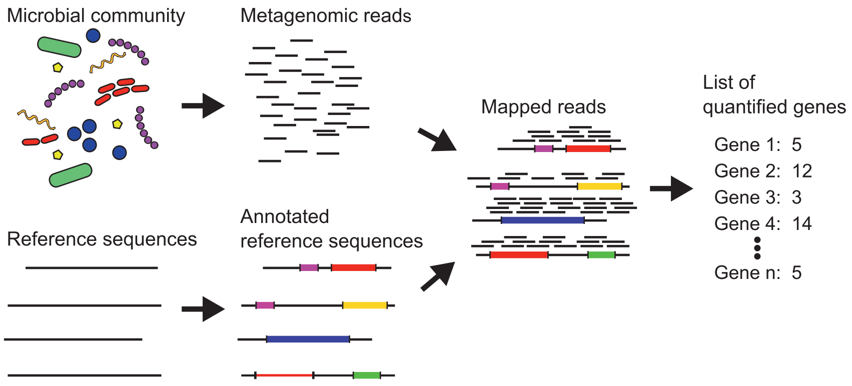

Figure 1: Overview of gene quantification in metagenomics. DNA is extracted and randomly sequenced from a microbial community. The resulting reads are then mapped to reference sequences that have been annotated according to their gene content. Each read that matches a gene is counted as an occurrence of that gene. The end result is a list of counts for each sample providing the relative abundance of each gene.

Because genes are quantified by counting the number of reads matching specific genes, metagenomic data becomes discrete, asymmetric and has a dependency between the expected value and the variance. This means that standard statistical methods that rely on normality assumptions are not suitable and can lose power to detect differences (Law et al., 2014). Metagenomes are also undersampled, meaning that the sequencing depth is not sufficient to reliably capture the full genetic diversity present in a community (Paulson et al., 2013; Unterseher et al., 2011). This means that a large proportion of genes may be represented with only a few reads, making the statistical analysis challenging. Rare genes with low abundance can also fall below the detection limit while still being present in the community, making it difficult to compare abundances between samples.

Metagenomic gene count data is high-dimensional and often thousands of genes are tested for differential abundance simultaneously. Each test may result in a false positive making it difficult to distinguish the few truly differentially abundant genes. Correction for multiple-testing is therefore needed to control the type-I error rate, which reduces the power to detect the true differences (Dudoit et al., 2003). Thus, methods with high power and low type-I error rate are needed for accurate analysis of metagenomic data.

Metagenomic data is affected by high levels of biological and technical variability. The biological variability reflects the variation in gene abundances between microbial communities. The variation is induced by differences in uncontrolled environmental factors between samples, for example temperature, salinity, nutrient availability, pH and host age (Fierer et al., 2012; Lozupone et al., 2012). Because the composition and abundance of species change due to these factors the abundance of their genes also change. Furthermore, bacteria, which constitute a major part of many bacterial communities, have plastic genomes, and gene content can vary between individual strains of the same species (Greenblum et al., 2015). For example, the genome of *Escherichia coli* contains 3188 genes within the core genome, while the plastic genome, which contains approximately 1500 genes, can vary between strains and 90,000 possible genes variants have been characterized to date (Land et al., 2015). Genes can also be present on horizontal gene transfer elements, such as plasmids, where it is possible for a single bacterium to contain several copies of the same plasmid. The horizontal gene transfer elements further increase the variation in gene abundances between samples. Finally, microbial species are not omnipresent and can be entirely missing in samples, causing observations represented by zero reads (Sohn et al., 2015). These characteristics make the biological variability substantial in most metagenomic data sets.

Technical variability is introduced due to differences in sample preparation sample preparation (Morgan et al., 2010), sequencing errors (Quail et al., 2012), differences in sequencing depth between samples (McMurdie and Holmes, 2014) and incorrectly mapped reads in the gene quantification step (Wooley and Ye, 2009). The variability between technical replicates has been shown to be smaller than the biological variability (Nayfach and Pollard, 2016). However, an unknown factor of technical errors is the bias introduced from biological databases. Metagenomics aims to study previously unknown microorganisms; however, all databases are based on previously identified genes and species (Rinke et al., 2013). In 2015, the number of completed bacterial genomes was 4000 (Land et al., 2015), which is only a small proportion of the total number of bacterial species, estimated to be at least 10 million (Curtis et al., 2002; Pedrós-Alió, 2006). The lack of accurate reference sequences can cause genes to be incorrectly annotated or missed completely (Wooley and Ye, 2009).

Metagenomic data sets often represented by few biological replicates due to the high sequencing costs (Knight et al., 2012; Prosser, 2010). The lack of replicates worsens the problems induced by the other challenges. For example, correctly estimating the variability of a gene is difficult when only a few samples are available. The obvious solution is to encourage replicated experimental designs. However, data sets with a low number of samples are still being produced and statistical methods that can provide robust estimates even when few samples are available are therefore vital.

2 Aims

The aim of this thesis is to improve the statistical analysis of metagenomic gene count data. The many challenges that are present in the analysis of metagenomic data require special care to be taken in model development to maintain high power to detect differences and a low proportion of false positives. The papers included in this thesis cover the evaluation of previously proposed methods, the investigation of the data itself and the development of new methods. Specifically, the aims are as follows:

- Evaluate previously proposed statistical methods for detecting differentially abundant genes in metagenomic gene count data (Paper I).
- Investigate and characterize the variance structures present in metagenomic gene count data (Papers II-III).
- Develop statistical models for improved detection of differentially abundant genes (Papers II-III).
- Extend the binning process to increase the power to detect biologically relevant effects in metagenomic gene count data (Paper IV).

3 Statistical analysis of metagenomic data

The following sections outline the statistical analysis of metagenomic gene count data and define the mathematical notation used throughout this thesis. This section also provides an overview of previously suggested statistical methods. Note that this section does not include a discussion of the performance of each method; for such a discussion, see papers I and III.

3.1 Identification of differentially abundant genes

In the typical gene-centric metagenomic experiment considered, the aim is to identify differentially abundant genes associated to experimental factors. Throughout this thesis we will focus on the comparison between two groups of samples. Note that more complicated experimental designs are possible, such as regression or comparisons between multiple conditions. For a comparison between two groups, the data takes the form of a matrix of counts with n rows corresponding to genes and $m_1 + m_2 = m$ columns representing samples. Let Y_{ij} be the counts of gene i in sample j, and let N_j denote the total sum of counts within each sample j.

Before the genes in the data are tested for differential abundance, the data is normalized to make the samples comparable and to reduce the variability in the data (Nayfach and Pollard, 2016). Commonly, the total sum of counts within each sample, N_j, is used for normalization to correct for differences in sequencing depth between samples. Other normalization methods have been proposed but evaluation of these are beyond the scope of this thesis, for more information see McMurdie and Holmes (2014). Throughout this thesis, the total sum, N_j, will be used for most methods, either to normalize the data or as an off-set specified within a model; the exception is software packages that by default use other normalization methods.

Each gene (row), i, is tested independently against the null hypothesis of no difference in relative abundance between groups, i.e.,

$$H_{0i} : \text{No difference in relative abundance for gene } i \text{ between groups,}$$
$$H_{ai} : \text{Difference in relative abundance for gene } i \text{ between groups.}$$

The exact formulation of the hypotheses varies depending on the underlying distributional assumption used by the specific method. Here we consider two-sided hypotheses for differential abundance but one-sided variants are also possible. The null hypotheses can then be rejected or not based on the outcome of the test statistic used. In most situations, a p-value is calculated for each gene and used to rank genes based on the significance of their differential abundance. For Bayesian methods (for example in paper II and III) other decision rules or metrics are used to rank the genes, such as the posterior probability for the relative abundance to differ between experimental conditions. The genes on top of this ranking list are then considered likely candidates to be differentially abundant and investigated further. Ideally, all of the truly differentially abundant genes would end up in the top of the ranking list, but this is never the case. These errors are often caused by small effect sizes, high variability in the data and a lack of biological replicates.

3.2 Overview of statistical methods

3.2.1 Tests for comparing pairs of samples

A number of classical statistical methods have been proposed for and applied to metagenomic data. In the early days of metagenomics, when sequencing was expensive and replicated experimental designs were rare, methods for comparing pairs of samples were used. Among the most prevalent is **Fisher's exact test** (Fisher, 1922, 1925). The test uses a 2×2 contingency table to test independence and tests whether there is an association between the number of matching fragments and the experimental conditions (Smith et al., 2012). **Fisher's exact test** has also been used to compare groups of metagenomes when the samples in each group have been pooled per gene, i.e. summing the counts per gene within each group (Parks and Beiko, 2010). Let y_{i1}^{sum} and y_{i2}^{sum} denote the sum of counts in the two groups for gene i, let r_{i1} and r_{i2} denote the sum of counts for all genes excluding gene i, and let N_1^{sum} and N_2^{sum} denote the sum of the total counts in the two pooled groups (see Table 1).

Table 1: Contingency table for a pooled analysis with Fisher's exact test. y_{i1}^{sum} and y_{i2}^{sum} is the sum of counts in each group, r_{i1} and r_{i2} the sum of counts in all genes excluding gene i and N_1^{sum} and N_2^{sum} are the sum of the total counts in the two pooled groups.

	Group 1	Group 2
Counts in gene i:	y_{i1}^{sum}	y_{i2}^{sum}
Counts in other genes:	r_{i1}	r_{i2}
Total counts:	N_1^{sum}	N_2^{sum}

Assuming that the margins are fixed, the probability of observing a specific outcome in the pooled case can be calculated via the hyper-geometric distribution as,

$$\mathbf{P}(Y_{i1}^{sum} = y_{i1}^{sum}) = \frac{\binom{y_{i1}^{sum}+y_{i2}^{sum}}{y_{i1}^{sum}}\binom{r_{i1}+r_{i2}}{r_{i1}}}{\binom{N_1^{sum}+N_2^{sum}}{N_1^{sum}}}. \tag{3.1}$$

The p-value for a two-sided **Fisher's exact test** is then calculated as the sum of the probabilities for all 2 x 2 tables with a lower probability than the observed table.

Another commonly applied test used for pairwise comparisons between samples is the binomial test, which has also been applied to metagenomics (Kristiansson et al., 2009; Mackelprang et al., 2011). This test compares the proportion of gene i to the total counts in each group. Under the null hypothesis of an equal proportion of gene i across both groups, the test statistic X follows the binomial distribution:

$$X \sim \text{Binomial}(y_{i1}^{sum} + y_{i2}^{sum}, \frac{N_{tot1}}{N_{tot1} + N_{tot2}}). \tag{3.2}$$

The p-values are derived directly from the binomial distribution using a two-sided alternative hypothesis. When the total number of fragments is large the binomial test is approximately equal to Fisher's exact test assuming that the counts for different genes are independent.

Another early method specifically designed for the analysis of metagenomic data was **XIPE-TOTEC** (Rodriguez-Brito et al., 2006), which is still being used today (Jeffries et al., 2015). **XIPE-TOTEC** focuses on pairwise comparisons between samples and assesses significance by bootstrapping the counts within each sample. In short, the algorithm works as follows. First, the counts in each of the two samples are redrawn with replacement to generate a large number of mock data sets. The difference between samples for each gene in every mock data set is calculated, and then the median difference for every gene is calculated. Next, a new set of mock data sets is created, but fragments belonging to any gene in any of the two original data sets are selected and a new set of gene-wise differences are calculated. This new set of differences constitutes a reference distribution. To determine the significance of each gene, the median difference of the original data set is compared with the reference distribution. At the 90% confidence level, a gene is considered significant if its median difference is smaller then the 5th percentile or larger than the 95th percentile. Note that **XIPE-TOTEC** was not included in the comparison performed in Paper I of this thesis due to the lack of an easily accessible implementation.

3.2.2 Methods based on normality assumptions

The t-tests are among the most commonly used statistical tests and have been applied in various forms to metagenomic gene count data (Grzymski et al., 2012; Turnbaugh et al., 2009; Ward et al., 2013). This includes both Student's t-test assuming equal variances in both groups and Welch's t-test assuming non-equal variances. However, t-tests rely on the assumption that the data is normally distributed. Metagenomic count data is discrete, often skewed and has a dependency between the mean and the variance. Thus, normality assumptions do not hold. To make the data symmetric and remove the dependency between the mean and the variance, the data is generally transformed before testing (Anscombe, 1948). Often used examples of transformations are the square root and $\log(Y_{ij} + \epsilon)$ where ϵ is a number e.g. 1 or 0.1.

The method **metaStats** is also based on a t-statistic but derives p-values using permutations (White et al., 2009). The raw counts are transformed using the base-two logarithm, and Welch's two sample t-statistics is computed for each gene. A null distribution for the statistic is then computed by permuting samples between groups and recalculating the t-statistic for each permutation. The p-values are then derived as the proportion of permuted t-statistics greater than the observed statistic as,

$$p_i = \frac{1}{B} \sum_{b=1}^{B} I\{|t_i^{0b}| \geq |t_i|\}, \tag{3.3}$$

where B denotes the number of permutations, t_i is the observed t-statistic of gene i and t_i^{0b} are the permuted t-statistics. At low sample sizes (≤ 8), where there are too few permutations to accurately estimate the p-values, **metaStats** pools the resampled t-statistics for all genes to form a reference distribution and calculates the p-values according to

$$p_i = \frac{1}{nB} \sum_{j=1}^{n} \sum_{b=1}^{B} I\{|t_j^{0b}| \geq |t_i|\}, \tag{3.4}$$

where the first sum is taken over all genes. The authors argue that the above tests are inaccurate for genes with very low abundance, i.e. less than 1 count on average across all samples. For this case, they pool the counts within each group and use **Fishers's exact test** to derive the p-value (see above).

The method **metagenomeSeq** assumes that gene abundances in metagenomes follow a log-normal distribution (Paulson et al., 2013). However, metagenomic data is often sparse due to undersampling and contains excess zeros. Because the log-normal distribution does not support zeros, the authors extend the model to include zero-inflation. The distribution of the log-transformed counts, $x_{ij} = log_2(1 + y_{ij})$, is defined as a mixture distribution

$$f_{zig}(x_{ij}; N_j, \mu_i, \sigma_i) = \pi_j(N_j)I_0(x_{ij}) + (1 - \pi_j(N_j))f_{count}(x_{ij}; \mu_i, \sigma_i), \tag{3.5}$$

where $\pi_j(N_j)$ is a mixture parameter depending on the total counts (N_j), $I_0(x_{ij})$ is a point mass at zero, and $f_{count}(x_{ij}; \mu_i, \sigma_i)$ is a log-count distribution approximated by a normal distribution. The estimates of differential abundance derived using the mixture distribution are tested using the moderated t-statistic implemented in limma (Smyth, 2004). The limma package stabilizes the variance estimates of each gene by sharing information between genes using an empirical Bayes approach. The zero-inflation is modelled on a per sample basis as a function of the sequencing depth within each sample with the motivation to correct for undersampling. Note that metagenomeSeq was primarily developed for analysis at the species level (i.e. operational taxonomic units (OTUs)), but it has also seen use on gene count data (Noyes et al., 2016). In addition to the statistical model **metagenomeSeq** also implements a new normalization procedure, cumulative sum scaling; for details, see Paulson et al. (2013).

RAIDA (Sohn et al., 2015) features a zero-inflated log-normal model for differential abundance testing. Let y_{kj} denote the sum of counts of the common divisor in sample j. The ratio of the observed counts to the common divisor, $r_{ij} = \frac{y_{ij}}{y_{kj}}$, is then modelled using the zero-inflated log normal distribution as

$$R_{ij}^{\epsilon} \sim \begin{cases} \text{Uniform}(0, \epsilon) & \text{w.p.} \quad \eta_i \\ \text{Log-normal}\,(\mu_i, \sigma_i) & \text{w.p.} \quad 1 - \eta_i \end{cases} \tag{3.6}$$

where ϵ is used as an offset to account for the lack of support in the log-normal distribution and η_i is the zero-inflation probability. In contrast to **metagenomeSeq**, **RAIDA** uses a gene-specific zero-inflation parameter. The model is fitted using the EM algorithm with p-values derived using the moderated t-statistic of limma (Smyth, 2004). **RAIDA** also features a heuristic and robust approach for selecting a set of genes to use as the common divisors. The assumption used is that the proportion of differentially abundant genes between two conditions should be small. Thus, genes are selected to be included in the common divisor by iteratively testing which common divisor that yields the fewest significant genes.

3.2.3 Non-parametric methods

Non-parametric methods are popular for inference because they avoid the problem of making specific assumptions on the distribution of the gene count data. The most commonly used non-parametric test is the Wilcoxon-Mann-Whitney test (WMW) (Karlsson et al., 2013; Sanli et al., 2013). The WMW test assumes that the data originates from distributions with the same shape and scale and tests whether one sample is stochastically larger than the other by comparing the ranks of observations (Mann and Whitney, 1947; Wilcoxon, 1946). The Kruskal-Wallis test which extends the WMW test to more than two groups has also been applied to metagenomic data (Segata et al., 2011).

3.2.4 Generalized linear models

Another way to model count data is generalized linear models (GLM) (McCullagh and Nelder, 1989). GLMs is a term used for a wide range of models that extend ordinary linear models beyond the assumptions of normality of residuals and permit other outcomes, e.g. counts or proportions that are often used in metagenomics. The expected outcome of gene i, $\text{E}[Y_i]$, is modelled using a linear predictor via a link function g, i.e.,

$$g(\text{E}[Y_i]) = \mathbf{X}\beta_{\mathbf{i}}, \tag{3.7}$$

where $\mathbf{X}$ denotes a design matrix ans β_i denotes a vector of predictors for gene i. Several different GLMs have been applied to metagenomic data. These include GLMs based on the Poisson distribution, which has the assumption that the mean is equal to the variance ($\text{E}[Y] = \text{var}[Y] = \lambda$) and was previously used in (Yatsunenko et al., 2012). Another is the quasi-Poisson, which includes a scaling factor, θ, allowing for variability beyond the Poisson variability parameter, i.e. $\text{E}[Y] = \lambda$ and $\text{var}[Y] = \lambda\theta$ (Kristiansson et al., 2009). Other examples include the negative-binomial (Zhang et al., 2017) zero-inflated negative binomial (Fang et al., 2016) and beta regression (Peng et al., 2016).

3.2.5 Methods from RNA sequencing

Much of the development of count-based statistical methods for large scale biological data has taken place within the related field of RNA sequencing (RNAseq) (Robles et al., 2012; Soneson and Delorenzi, 2013). Here, the counts represent the expression levels of genes within a single organism, and while the structure of the final data is similar to that in metagenomics, the underlying biological process is very different. However, many of the techniques used, such as overdispersed count models are applicable to metagenomic data. Numerous methods have been proposed, and a subset of methods developed for RNAseq that have been applied to metagenomic data is presented below.

DESeq2 (Love et al., 2014) and **edgeR** (Robinson et al., 2010) are two of the most commonly used methods for RNAseq that have also been applied to metagenomic data (Castro-Nallar et al., 2015). Both methods model the data as overdispersed counts using the negative binomial distribution and stabilize variance estimates using an empirical Bayes approach. However, the methods use slightly different approaches for calculating the amount of variance information to share between genes and determining which genes that should have their variance estimates adjusted. In addition, the methods are implemented in software packages that by default rely on different normalization methods; trimmed mean of m-values (TMM) (Robinson and Oshlack, 2010) for **edgeR** and meadian-of-ratios (Anders and Huber, 2010) for **DESeq2**. **DESeq2** also includes automatic filtering for outliers and genes with low expression.

Voom (Law et al., 2014) was developed to retain the simplicity and ease of use of standard linear models while accounting for the count-based nature of RNAseq data. **Voom** achieves this by modelling the expected value of the log counts per million (log-cpm),

$$c_{ij} = \log_2\left(\frac{y_{ij} + 0.5}{N_j + 1.0} \times 10^6\right) \tag{3.8}$$

, as a standard linear model,

$$E[C_{ij}]) = x_j^T \beta_i \; , \tag{3.9}$$

with x_j being a vector of co-variates and β_i being a vector of coefficients. Using the fit of this simple model, **Voom** identifies the mean-variance dependency found in count data using a trend line between the mean log counts and the square root of the standard deviations. This trend line is translated into a set of precision weights w_{gi}, which, together with the log-cpm counts, are fed into the limma package to detect differential abundance.

3.3 Evaluation of statistical performance

The evaluation of statistical performance is a part of all papers included in this thesis. This section describes how to generate suitable test data that is similar to real metagenomic data. This section also explains several of the performance measures used to evaluate statistical power and control of false positive rates.

3.3.1 Generating test data

To analyse the performance of statistical methods, suitable test data is needed, both to ensure that the type-I error rate is controlled and to evaluate the power to detect differences. However, many different forms of test data can be used, e.g. data simulated from parametric distributions or real metagenomic data. In this thesis we primarily use resampled data, which can be viewed as a mix between real and simulated data.

Resampled data is generated by randomly drawing samples from a real metagenomic data set and then adding simulated effects (see figure 2). The algorithm used in the papers proceeds as follows. Start with a large (>30 samples) metagenomic data set that is representative of some environment, for example, the human gut or an environmental ecosystem. Next, randomly select the desired number of samples without replacement and divide them into two groups. This new data set will represent a null distribution where genes, on average, will have no effect.

Next, effects are added to the resampled data. There are several possible ways to add effects but we argue that downsampling (thinning) the counts has a low impact on the structure of the data. First, the desired proportion

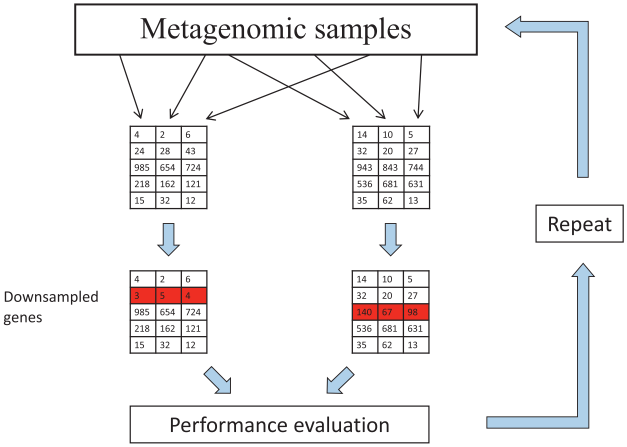

Figure 2: Generating resampled data to evaluate performance. First draw a subset of samples from a large metagenomic data set and randomly assign them to two groups. Add effects to the desired number of genes by randomly downsampling counts in all samples belonging to one of the groups. Evaluate performance and repeat the resampling procedure the desired number of times.

(e.g. 10%) of genes in the resampled data set are randomly selected to have a difference in relative abundance of q (e.g. 5). All of the samples belonging to one group (randomly selected) then have their observed counts y_{ij} replaced with downsampled counts $\tilde{y}_{ij}$ drawn from a binomial distribution according to,

$$\tilde{y}_{ij} \sim \text{Binomial}(y_{ij}, \frac{1}{q}). \tag{3.10}$$

This corresponds to randomly removing DNA fragments with a probability of $(1 - \frac{1}{q})$. Because effects are simulated by removing counts, only negative effects can be added, but by dividing effects equally among both groups, both positive and negative effects can be included. The proportion of affected genes can be different between groups but balanced effects are the main focus of this thesis. For a discussion of unbalanced effects, see the work by Sohn et al. (2015).

Data simulated directly from parametric distributions has the benefit of providing full control over all parameters. However, simulated data relies on many assumptions that most likely are not true in real data. This can result in a strong bias towards models that are based on similar assumptions. Conversely, resampled data provides a more realistic basis for the evaluation of statistical performance and it retains the within-sample variability present in real data. Furthermore, downsampling provides a non-intrusive way to add effects by preserving the variance structure of the affected gene. Note that while resampling typically results in data that is more realistic compared to simulations, it still presents an idealized case. In any real sampling situation, such as comparing a set of control samples with a treatment group, the samples are never a true subset of the population, and there can be several parameters, such as pH, temperature, age of patient, and so forth, that may co-vary with the groups and mask the effect of the treatment. On the other hand, resampled data becomes truly randomized on average, and effects are added orthogonally to all potential co-variates. In addition, effects are added independently and with equal probability to all genes in the data set, disregarding the abundance, variability and proportion of zeros present in the genes. In the current implementation, the added fold-change is equal for all gene with an effect and all samples are are downsampled with the same parameter. In a real comparison, genes would be expected to have different effects and the effect size between samples is likely to vary for a single gene. Furthermore, differences in gene abundances between microbial communities are likely to be strongly correlated. For example, if an organism is favored in an environment, all genes specific to that organism would increase their relative abundance. However, increasing the complexity of the resampled data would make the results less transparent, and we argue that resampling still provides a suitable approach to evaluate the performance of statistical methods.

The benefit of using real metagenomic data for testing is that no distributional assumptions are needed. Real data is often used as a proof of concept that a new method works as intended. When reanalysing real data with a new method, the result can be compared to those previously attained, as was done in Paper IV. The downside is that the true differences, if any, are unknown and the performance of the method can therefore not be correctly estimated. For this reason, well-studied real data sets, such as the reference data sets generated by the sequence quality consortium for RNAseq data (Seqc/Maqc-Iii Consortium, 2014), can be used but no such initiative exists for metagenomic data. Another approach is to create mock communities (Morgan et al., 2010), but these will not capture the full complexity of real metagenomic communities and should therefore be considered as idealized cases. Thus, it is not possible to solely rely on real data for the evaluation of statistical performance within metagenomics.

3.3.2 Measures of statistical performance

A large proportion of this thesis focuses on the evaluation of statistical performance. Throughout these papers, three different aspects have been primarily been considered: the ability to rank genes based on differential abundance by generating receiver operating characteristic (ROC) curves and calculating the area under the curve (AUC) (Fawcett, 2006) (Papers I-III), the ability to control type I errors by investigating the distribution of p-values under the null hypothesis (Paper I) and the ability to control errors in a multiple testing situation through the false discovery rate (FDR) (Benjamini and Hochberg, 1995)(Papers I and IV). This section will outline these different performance measures and their advantages and drawbacks.

Ranking genes based on differential abundance is a common way to analyse metagenomic data. Ideally, the list of ranked genes generated by a statistical test should contain the truly differentially abundant genes at the top and non-differentially abundant genes at the bottom. However, ranking lists are typically far from perfect due to the variability present in the data, lack of replicated samples, small effect sizes and non-optimal model assumptions. To evaluate ranking performance, we use ROC curves which are a common method for visualising the statistical performance of a classifier, in this case the test for differential abundance. In short, a ROC curve is created by going through the ranking list and at each position calculating the true positive rate (TPR) and the false positive rate (FPR). The TPR is defined as the number of true positives above the threshold divided by the total number of true positives in the data. The FPR is similarly defined as the proportion of false positives above the threshold in relation to all the false positives present in the data. The result is a curve for each ranking list, where each point is the FPR (x-value) and TPR (y-value) at a specific position in the ranking list (see Figure 2). Every true positive encountered in the list is a step upwards, and every false positive is a step to the right. The area under the curve (AUC) summarizes the ranking performance into a single value and is calculated as the area under the ROC curve. A perfect ranking result corresponds to a ROC curve that immediately achieves and FPR of 1 and would therefore achieve an AUC of 1. A test that randomly selects between the hypotheses generates a ROC curve that is a straight line with a slope of 1, resulting in an AUC of 0.5. It is most common to use the full AUC, which measures the quality of the entire ranking list. However, within metagenomics, the assumption is often that only a small proportion of genes are truly differentially abundant. Thus, the quality of the top of the ranking list, which hopefully contains the majority of true positives, is more important than the bottom. To provide a more representative value, we calculate the AUC up to some pre-specified FPR cut-off, e.g. the AUC up to a FPR value of 0.1, which we denote as $\text{AUC}_{0.1}$.

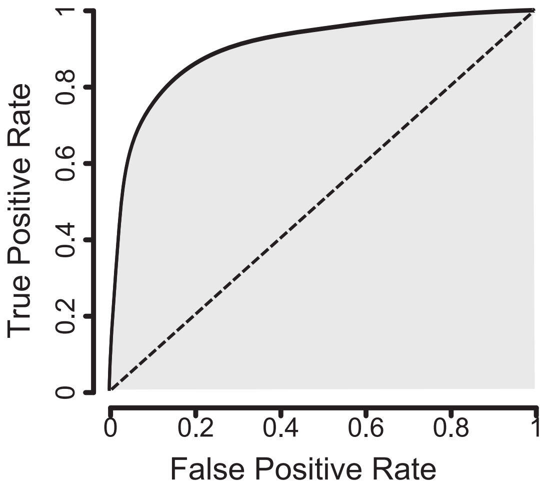

Figure 3: Illustration of a receiver operating characteristic (ROC) curve. A ROC curve measures the quality of a ranking list by measuring the true positive rate and false positive rate along the list. The black line is an example ROC curve. The grey area represents the area under the curve (AUC), which is a measure of the average quality of the ranking list. The dashed line corresponds to the performance of a random classifier.

Gene ranking is often based on p-values, but only reporting ranking performance provides no information regarding the accuracy of the p-values themselves. P-values depend on the validity of the model assumptions. Incorrect model assumptions can lead to strongly biased p-values, resulting either in an overly optimistic classification with too many false positives or a pessimistic classification where true differences are missed. If the model assumptions are correct, the distribution of p-values should be uniformly distributed under the null-hypothesis. In paper I, we evaluated this property using resampled data without added effects to simulate a null distribution. The uniformity of the p-value distribution on the resampled data then provides a measure of how well the assumptions behind each model fit the data and whether there is any risk of excess false positives.

For a single statistical test, the probability of a type-I error, i.e. a false positive, is controlled by specifying the significance level α. When several independent statistical tests are performed simultaneously, e.g. one for each gene in a metagenomic data set, each test can result in a false positive. For this reason, several other measures of type-I error rates along with procedures to control them were introduced that were suitable for multiple-testing situations (Dudoit et al., 2003). One commonly used measure is the family wise error rate (FWER), defined as the probability of getting at least one false positive among all tests. One method to control the FWER is the Bonferroni procedure, which instead of α, uses the more strict significance cut-off $\frac{\alpha}{n}$, where n is the total number of performed tests. However, controlling the FWER is often overly conservative because a small proportion of false positives can be acceptable in order to retain a larger number of true positives. Controlling the false discovery rate (FDR), defined as the expected proportion of false positives, provides such an alternative (Benjamini and Hochberg, 1995). Benjamini and Hochberg introduced a method for estimating the FDR, which has become the most common way to control type-I errors in multiple testing situation arising in the analysis of high-dimensional data within the life-sciences. Given an ordered list of p-values, the FDR at each position k can be estimated as

$$\widehat{FDR}(k) = \frac{np(k)}{k}. \tag{3.11}$$

This creates a new list of values often referred to as q-values. A cut-off in the q-values would on average guarantee that the FDR is controlled below that cut-off given that the assumptions of the Benjamini and Hochberg method are satisfied (see Benjamini and Yekutieli (2001) for details).

The accuracy of the FDR estimation can then be investigated in several ways. Of primary interest is the ability to control the FDR at the specified cut-off which is a fundamental requirement for sound statistical analysis. Given test data with known effects, the true FDR at position k in gene list can be calculated as,

$$FDR(k) = \frac{\text{Number of false calls up until } k}{k}. \tag{3.12}$$

The bias in the estimated FDR compared to the true FDR can give an indication that a method is more or less conservative. Two methods that both are able to control the FDR can still achieve a different number of true positives identified. That is, the ratio of false positives to total called genes is the same but one method has a higher number of called genes. Therefore it is also of interest to measure how many true and false positives are identified by each method at the given FDR cut-off to obtain a sense of the power to detect differentially abundant genes.

The FDR is commonly used within metagenomics to control error rates and to detect differentially abundant genes. Thus, investigating the accuracy of FDR estimates provides important information about the reliability of a statistical method. However, the accuracy of FDR estimates is not a transparent metric on its own. Biases observed in the FDR estimates can depend on the accuracy of the underlying p-values, the quality of the ranking list and of the method used for estimating the FDR. Thus while FDR bias is a useful metric for evaluating the performance of a statistical method, other metrics are needed to show the complete picture.

4 Summary of papers

This section outlines the aims and backgrounds and highlights the main results of each of the four papers included in this thesis.

4.1 Paper I: Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics

The aim of Paper I was to evaluate and compare statistical methods used for detection of differentially abundant genes and provide a guide for the sound statistical analysis of metagenomic data. In total, 14 different methods were included, ranging from classical statistical tests to newly developed methods for metagenomic and RNAseq data. The performance was measured with respect to their ability to rank genes, the control of the type 1 error rate and their ability to control the false discovery rate (FDR). To make the results as realistic as possible, the comparison was based on resampled data generated from two comprehensive metagenomes from the human gut (Qin et al., 2010; Yatsunenko et al., 2012).

Ranking performance was evaluated for each method with regard to group size, effect size and gene abundance. Group size had the largest impact on performance both in terms of overall ranking accuracy and in the relative differences between methods. The overdispersed Poisson GLM had a high performance and was the best method at a group size of 5+5 (see figure 4). DESeq2 and edgeR which use empirical Bayes to stabilize variance estimates had high performance across all group sizes and had the largest advantage at the lowest group size (3+3). Next in terms of performance were a large number of methods that permit a gene-specific variance, such as the t-tests.

These methods had a strong performance at higher group sizes with small differences between methods but tended to have poor performance at lower group sizes. The methods that do not account for between-sample variability, i.e. Fisher's exact test, the binomial-test and the Poisson GLM, had by far the lowest performance across all three group sizes (see Paper I, Figure S1). For details on the performance of specific methods see the full paper.

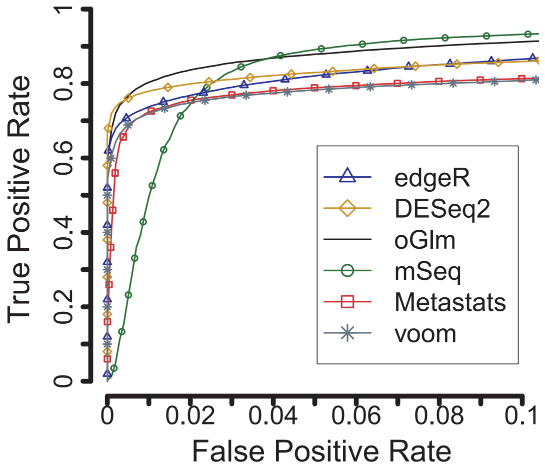

Figure 4: Gene ranking performance on the Qin 2010 data set. Average ROC curves for each of the included methods. The group size was set to 6+6, and the effect size was set to a fold-change of 5. OGLM is short for the overdispersed Poisson GLM and mSeq is short for metagenomeSeq. The results are averaged over 100 sets of resampled data. The plot corresponds to Figure 1 panel b of Paper I.

To investigate the type-I error rate, all methods were applied to resampled data without added effects, i.e. under an empirical null distribution. Ideally, the p-values should be uniformly distributed in this case. Almost all methods satisfied this criterion and had only minor deviations from uniformity. The method metagenomeSeq did show a clear sign of too optimistic p-values. However, the most striking result was again that the methods that do not account for between-sample variability, i.e. Fisher's exact test, the binomial-test and the Poisson GLM, which had very skewed p-value distributions towards low values, leading to a risk of a high number of false positives (see Paper I, Figure S6).

Finally, the ability to control the false discovery rate was investigated by comparing the true FDR at an estimated FDR of 5% for each method (see Figure 5). Most methods were indeed able to control the true FDR at the specified level. However, metagenomeSeq was not able to control the FDR. Additionally, the methods that do not account for between-sample variability were completely unable to control the FDR (see Paper I, Figure S8). Furthermore, the methods that were able to control the FDR varied in the number of true positives detected. The three most powerful methods were edgeR, DESeq2 and the overdispersed Poisson GLM.

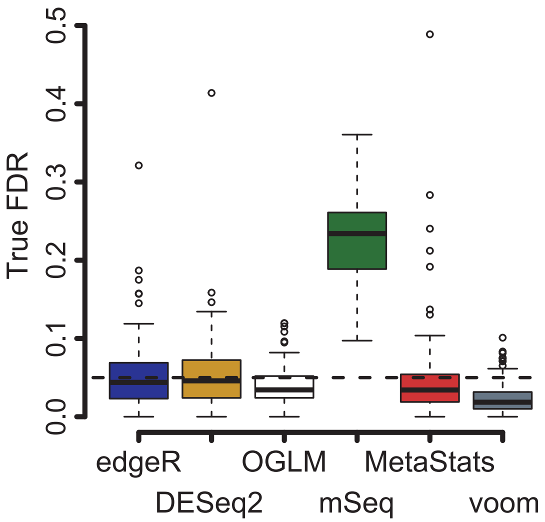

Figure 5: Investigation of FDR control. Box plots showing the true FDR at an estimated FDR of 5% for each method. The group size was set to 6+6, and the effect size was set to a fold change of 5. Each box corresponds to 100 resampled data sets from the Qin 2010 data set. The plot corresponds to Figure 5 panel c of Paper I.

This paper represents the first comprehensive evaluation of statistical methods for metagenomic gene count data. The results showed that several methods developed for RNAseq do indeed also perform well also on metagenomic data and should be recommended in many cases. In addition, the overdispersed Poisson GLM had a high performance and even outperformed the RNAseq methods given that sufficient samples were available. More alarming was the performance of the methods that do not handle the variability of metagenomic data, i.e. Fisher's exact test, the binomial-test and the Poisson GLM, which cause a large number of false positives possibly leading to erroneous biological conclusions. Thus, this paper serves as a guide both for selecting the appropriate methods and for aiding the further development of statistical methods for metagenomic data.

4.2 Paper II: Variability in metagenomic count data and its influence on the identification of differentially abundant genes

The aim of Paper II was to investigate the extent of variability present in metagenomic count data and evaluate different ways of modelling this variability. This work centers around a hierarchical Bayesian generalized linear model based on a Poisson distribution with a gene-specific variance parameter, σ_i^2. Letting Y_{ij} denote the count of gene i in sample j, the model is formulated for a comparison between two groups as

$$\log(E[Y_{ij}|\alpha_i, \beta_i, u_{ij}]) = \alpha_i + \beta_i I_G(j) + u_{ij} + \log(N_j),$$

where α_i is the baseline, β_i is the difference between groups determined by the indicator function $I_G(j)$, u_{ij} are random effects, and N_j is a sample-specific normalization factor. Conditioned on the parameters Y_{ij} is assumed to be independent and Poisson distributed according to

$$Y_{ij}|\alpha_i, \beta_i, u_{ij} \sim \text{Poisson}(N_j e^{\alpha_i + \beta_i I_G(j) + u_{ij}}).$$

The random effects, u_{ij}, are assumed to follow a normal distribution with the variance parameter σ_i^2, $u_{ij} \sim \text{Normal}(0, \sigma_i^2)$, making the distribution of Y_{ij} conditioned on α_i, β_i and σ_i^2 a Poisson log-normal. This paper then focuses on two main questions regarding σ_i^2. First, what does the distribution of σ_i^2 look like in metagenomic data, and second, how do modelling assumptions on σ_i^2 impact the power to detect differentially abundant genes?

To answer the first question, three comprehensive metagenomic data sets were included: two from the human gut (denoted Human Gut I (Qin et al., 2012) and Human Gut II (Yatsunenko et al., 2012)), and one sampled from oceanic surface water (Sunagawa et al., 2015) (denoted Marine). The model was fit to each of these data sets, and the posterior mean of the overdispersion defined as $\phi_{ij} = e^{\sigma_i^2} - 1$ was calculated (see Figure 6). As expected, the overdispersion parameter varied widely between different genes within each data set. However, the distributions showed large similarities between data sets. Furthermore, the correlation of the gene-specific overdispersion between data sets was high, indicating that overdispersion has a link to the properties of each gene. The overdispersion was also shown to be linked to the biological properties of each gene by mapping every gene to their corresponding gene ontology (GO) term (Consortium, 2015). GO terms related to basal cell functions showed a significant enrichment of genes with low overdispersion and 60 significant GO terms overlapped between all three data sets. This indicates that the gene-specific variability is indeed linked to biological function.

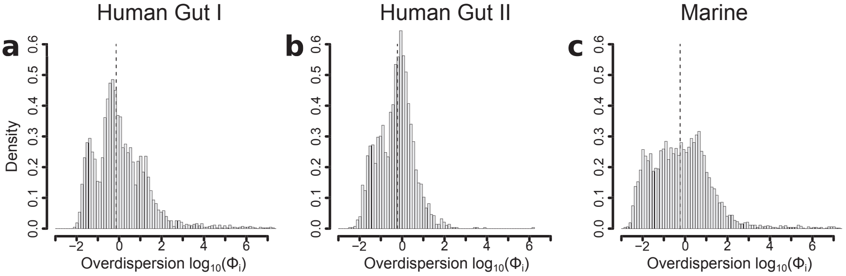

Figure 6: Histograms of the posterior mean of the overdispersion parameter ϕ_{ij} for each gene in each dataset. The dashed line shows the median overdispersion. Panel a shows the results for the Human Gut I data set, panel b shows the results for the Human Gut II dataset, and panel c shows the results for the Marine data set. The figure corresponds to Figure 1 of Paper II.

The second part of this paper targeted modelling of the gene-specific variability σ_i^2 and how this impacts the ability to detect differentially abundant genes through ranking. First, the three models assumed 1) $\sigma_i^2 = 0$ (no between sample variability beyond the Poisson variance), 2) $\sigma_i^2 = \sigma^2$ (same variance for all genes) and 3) σ_i^2 (gene-specific variance). The models were evaluated on resampled data generated from the three data sets, and the ranking performance was measured (see Figure 7). The model using gene-specific variability outperformed the other models on all but the third data set, where the model assuming a single variance parameter for all genes had a slightly higher performance. Model 1, which did not account for between sample variability, had the lowest performance in all data sets and had up to 80% false positives among the top 10% of the ranking list. This shows the importance of modelling the gene-specific variability in metagenomics.

Model 3 was extended with a prior on the gene-specific variance parameter that was shared between genes to stabilize the variance estimates. Three different shared priors for σ_i^2 were evaluated: a gamma distribution, an inverse-gamma distribution, and a log-normal distribution. For comparison, the model with a gene-specific flat prior was also included. The choice of prior did have a large impact on the ranking performance, and all three prior distributions were preferred to the gene-specific prior. However, none of the three priors were clearly better than the others. The inverse-gamma had the best overall ranking performance, but the gamma prior showed a better performance when only

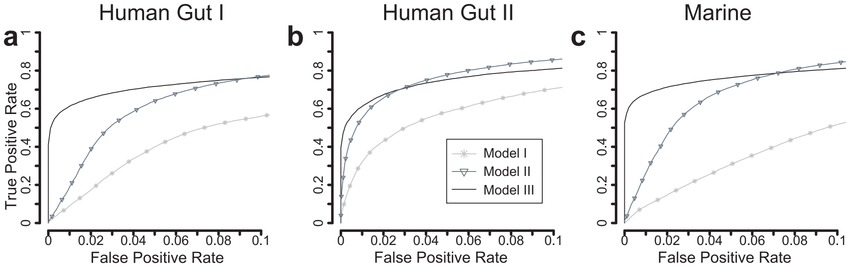

Figure 7: Evaluation of modelling assumptions on the gene-specific variability σ_i^2. Model 1 assumed no between sample variability, Model 2 assumed the same variance for all genes, and Model 3 allowed for gene-specific variance. The group size was 10 + 10 and the effect size was set to 3. The results were averaged over 50 sets of resampled data from each of the three data sets. Panel a shows the results for the Human Gut I data set, panel b shows the results for the Human Gut II dataset, and panel c shows the results for the Marine data set. The figure corresponds to Figure 3 of Paper II.

considering genes with a high overdispersion. This result showed that the choice of prior is important and can in some cases have as large of an impact as the gain from using a shared prior to begin with.

The results in Paper II highlight the importance of modelling the variability present in metagenomic data. It showed that there is biological information present in the overdispersion of a gene and that incorrect modelling of the variance will increase the number of false positives.

4.3 Paper III: A zero-inflated model for improved inference of metagenomic gene count data

The aim of Paper III was to improve the analysis of metagenomic data both by investigating the presence of zero-inflation in the data and evaluating the impact of modelling zero-inflation on the statistical analysis. In this context, zero-inflation means that the data is expected to contain more zeros than what would be predicted by a standard distribution, in this case the Poisson distribution. There are many possible causes for the excess zeros found in metagenomic data. In general, a zero occurs when the abundance of a gene is below the detection limit. However in metagenomic data the diversity between bacterial communities is so large that genes might be entirely absent from samples while still present in other samples from that environment. For this reason, we extended the model used in Paper II to incorporate zero-inflation.

The new model was formulated as

$$Y_{ij}|\alpha_i, \beta_i, u_{ij} \sim \begin{cases} 0 & \text{w.p.} \quad p_i, \\ \text{Poisson}\left(N_j e^{\alpha_i + \beta_i I_G(j) + u_{ij}}\right) & \text{w.p.} \quad 1 - p_i, \end{cases} \tag{4.1}$$

where p_i is a gene-specific zero-inflation parameter. The result is that the observed count is modelled to originate from either the previously defined Poisson-log-normal distribution or an independent zero-inflating processes governed by p_i. The overdispersion parameter σ_i^2 is modelled via a global gamma prior ($\sigma_i^2 \sim \text{Gamma}(\eta, \kappa)$ to permit sharing of variance between genes. For an overview of the model see Figure 8.

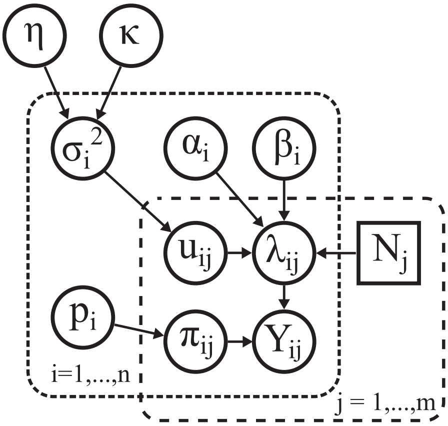

Figure 8: Illustration of model structure. The dashed boxes show which parameters are defined per gene, i, and sample, j, where n is the total number of genes and m is the total number of samples. η and κ are parameters for the global prior on the gene-specific overdispersion σ_i^2. u_{ij} are the random effects, α_i denotes the baseline abundance, β_i denotes the difference in abundance, N_j is a sample-specific normalization factor, λ_{ij} is the raw gene abundance, and Y_{ij} is the sampled gene abundance. p_i is the gene-specific zero-inflation parameter controlling π_{ij} which indicates whether an observation is zero-inflated. The figure corresponds to Figure 1 of Paper III.

When zero-inflation is included in the model, the excess variability in the data is divided into two parts: zero-inflation for any extra zeros present in the data and overdispersion to capture the between-sample variability. Without zero-inflation both forms of variability have to be captured by the overdispersion parameter which results in biased estimates. To examine this, we fitted the model both with zero-inflation (denoted ZoP) and the model without zero-inflation (denoted oP) to simulated data, with and without added zeros, and examined the posterior distributions of σ_i^2 for all genes (see Figure 9). When no extra zeros were added to the data, both models were similar to the true distribution. However, when zeros were added to the data, the non-zero inflated model showed a large increase in overdispersion estimates with more than a 300% increase at the highest level. The zero-inflated model generated almost unbiased variance estimates regardless of the level of zero-inflation added.

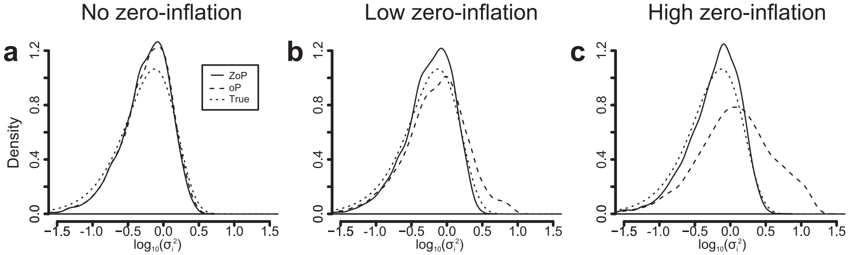

Figure 9: Effect of zero-inflation on overdispersion estimates in simulated data. Posterior means of the overdispersion parameter σ_i^2 for the model with zero-inflation (ZoP, solid line) and the model without zero-inflation (oP, dashed line). The levels of zero-inflation added were a) expected $p_i = 0$ (no zeros added), panel b) expected $p_i = 0.034$ and panel c) expected $p_i = 0.11$. The dotted line indicates the true distribution for σ_i^2. The figure corresponds to Figure 2 of Paper III.

The zero-inflated model was then fit to three real metagenomic data sets from the human gut (Qin et al., 2010, 2012; Yatsunenko et al., 2012), and two of them showed a large amount of excess zeros. The same pattern of increasing overdispersion estimates was observed on these two data sets, where there were large differences between the estimates of the ZoP and oP.

The ZoP model was then evaluated on resampled data together with five other methods: RAIDA and metagenomeSeq, which both model excess zeros, and edgeR, DESeq2 and voom, which were originally developed for RNAseq data and do not incorporate zero-inflation. Incorporating zero-inflation did provide a large increase in performance on resampled data, and the ZoP model and RAIDA had an overall high performance (see figure 10). The impact of adding zero-inflation was the largest at the higher group sizes. On resampled data from the third data set, which had a lower amount of zero-inflation, the ZoP model still performed on par with the best RNAseq methods. The two other zero-inflated methods had a lower performance when no excess zeros were present. Thus, the ZoP model had the highest performance overall.

The results in Paper III show that excess zeros are a natural part of metagenomic data. If not accounted for, excess zeros in the data will lower the power to detect

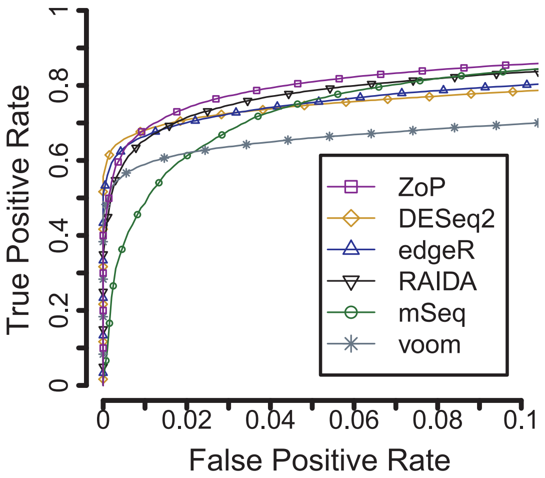

Figure 10: Ranking performance on resampled data from the Qin 2012 data set. The group size was 10+10, and the effect size was 3. The results were averaged over 100 resampled data sets. The figure corresponds to Figure 6 panel e of Paper III.

differentially abundant genes. However, care should be taken that sufficient samples are available to accurately identify zero-inflated observations. Taken together, the model proposed in this paper further improved the ability to identify differentially abundant genes and provides insights into the variance structure of metagenomic data.

4.4 Paper IV: HirBin: High-resolution identification of differentially abundant functions in metagenomes.

In Paper IV, we present a new method for quantifying the gene content in metagenomic data (binning) with the aim of identify more specific biological effects. The bins typically used in metagenomic data analysis are defined as groups of genes of similar function or structure and defined in various databases, e.g. TIGRFAM (Haft et al., 2013), eggNOG (Huerta-Cepas et al., 2015) and KEGG (Kanehisa et al., 2008). However, the definitions of a bin inside the databases are designed to cover many genes from multiple species and are therefore not necessarily able to discern more specific functions. The main idea behind this paper is that biological effects can act on several different levels, from single genes to whole bins. If an effect acts on only a subset of the genes inside a bin, that effect will likely be diluted when viewed from the full bin. For example, if a specific gene variant is beneficial for surviving in a polluted environment, it would become more prevalent in that environments. However, if that gene variant is binned together with other related genes that do not confer this benefit, the effect on the gene will be reduced and its differential abundance harder to identify. To solve this problem, we have developed HirBin, which uses a two-stage procedure for binning reads to improve the resolution of the analysis (see Figure 11). In the first supervised step, the data is annotated using standard methods against a pre-existing database. In the second unsupervised step, the sequences matching each bin are further divided into sub-bins by clustering based on sequence similarity. The result is a new set of sub-bins that represent more specific biological functions . The statistical analysis can then be performed at the sub-bin level at a pre-specified sequence similarity cut-off.

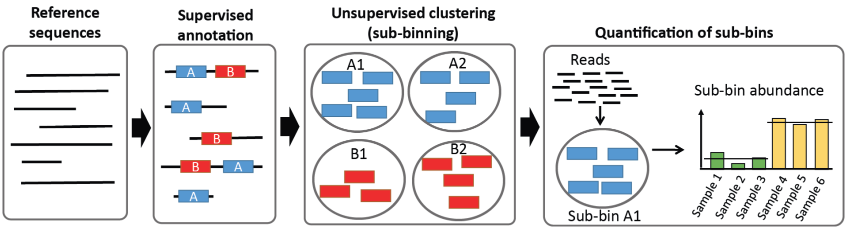

Figure 11: Overview of the HirBin method. The reference sequences are first annotated according to gene content. For each gene, the sequences matching that gene are clustered, forming sub-bins. Finally, the gene content is quantified according to the sequences within each sub-bin. The figure corresponds to Figure 1 of Paper IV.

To show the benefit of this methodology, HirBin was applied to a metagenomic data set from a study of type II diabetes (Qin et al., 2012) at both 50% and 75% sequence similarity cut-offs. At the 50% cut-off, the total number of observed sub-bins increased to 15,740 compared to 2,465 observed without sub-binning. After bin-wise statistical testing between individuals with and without type-II diabetes, 4,436 sub-bins at 50% sequeance similary were deemed significant (FDR< 0.05) compared to 457 of the original bins. This corresponded to an increase in the proportion of significant bins from 18.5% on the full bins to 28.2% at the 50% sequence similarity cut-off. Considering the full bins, 987 bins that were not significant had at least one significant sub-bin, while 112 of the previously significant bins were lost at the 50% sequence similarly cut-off. This means that the increase in the number of significant bins did indeed detect previously undetected functions.

To further explain the observed dilution of effects when analysing the full bins, HirBin was applied to resampled data from the same data set. Effects were added to 10% of the sub-bins at the 80% sequence similarity level. HirBin was then applied to this data and the ability to detect these effects wes analysed at 50% and 75% sequence similarity levels along with the analysis of the full-bin level (see Figure 12). The results showed that both the power to detect differences and the estimated fold-change substantially decreased at the less precise binning levels as a consequence of the dilution effect. At the full-bin level the effects were almost completely diluted but were still detectable at the sub-bin levels.

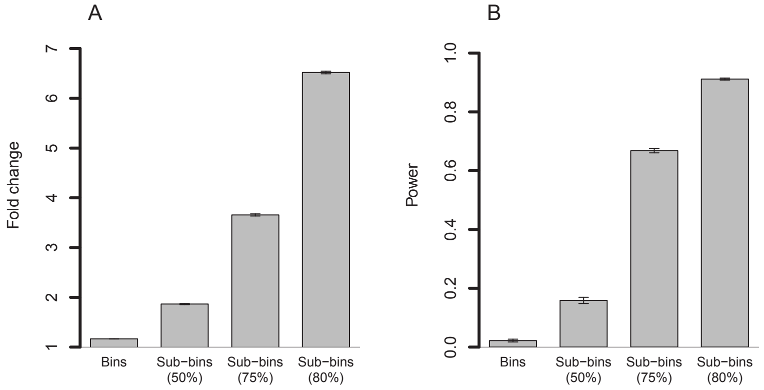

Figure 12: Analysis of achieved fold-change and power when effects are added at a sub bin level. Both panels show the results on resampled data where an effect has been added at the 80% sequence similarity level. Panel A shows the average estimated fold-change at each sequence similarity cut-off. Panel B shows the average power to detect the effect at FDR< 0.05 at different sequence-similarity cut-offs. The figure corresponds to Figure 4 of Paper IV.

In conclusion, HirBin provides a novel data-centric approach to binning that makes it possible to detect differences at finer resolution, effects which would be missed using standard approaches to binning. This enables more accurate biological interpretations, which will further our understanding of microbial communities.

5 Conclusion and outlook

This thesis targeted the statistical analysis and modelling of metagenomic gene count data. The work was centered around four main aims. Paper I provided the first broad evaluation of statistical methods for metagenomic data based on resampled data. It also gave a first indication of which modelling aspects are the most important to consider. Paper II presented a more detailed evaluation of the variability present in metagenomic data from the perspective of an overdispersed Poisson model. The paper also explicitly evaluated the impact of modelling this variability and the consequences of not properly accounting for the between-sample variability. During the work on the first two papers and through the work of Paulson et al. (2013) and Sohn et al. (2015), it became apparent that excess zeros and sparsity might be an integral part of metagenomic data. In Paper III, zero-inflation was introduced as an extension to the model used in Paper II. Using this zero-inflated model, it was shown that metagenomic data does indeed contain more zeros than predicted by most standard models. Furthermore, the proposed model showed a considerable increase in ranking performance both on simulated and resampled data. Finally, Paper IV explored the impact of sub-clustering genes (bins) to identify differences in biological functions at a more detailed level. It was shown that effects can indeed be missed when bins are too broad in their definition. This hints at the complexity of the underlying biological processes, and HirBin provides a useful data-centric approach to identify effects at a more specific functional level.

Throughout this work, it has been shown that statistical modelling have a large impact on the analysis of metagenomic data. Using appropriate models can substantially improve the power to detect differentially abundant genes, while non-optimal models can cause an excess number of false positives. In summary, four main aspects of modelling metagenomic data have been observed. The first is the discrete nature of the data. Count-based methods, for example the overdispersed Poisson GLM, were shown to have higher power to detect differences than e.g. the t-tests and non-parametric methods in Paper I. The impact of modelling the counts is larger when the average observed count is lower. As data becomes cheaper and easier to produce, the average sequencing depth will increase, which will lower the impact of modelling the data as counts. Still, when targeting less abundant genes or more specific functions, for example, the sub-bins generated by HirBin in Paper IV, the average count will be lower, and using count-based methods is therefore recommended.

The second aspect was modelling of the between-sample variability and permitting a gene-specific variance. Not accounting for this variability was shown to cause a large increase in false positives in both Paper I and Paper II. Most standard methods do permit gene-specific variability, for example, t-tests and the overdispersed Poisson GLM. Unfortunately, the lack of replicates in the early metagenomic era invited the use of too simplistic methods, and many of these are still in use. Thus, an essential step to improve the overall reliability of metagenomic data analysis is to completely stop using methods that do not account for between-sample variability.

The third aspect was extending the modelling of variance to permit sharing of information between genes. This results in variance estimates shrinking towards some common value. Shrinkage of the gene-specific variance estimates has been an integral part in the analysis of high-dimensional genomic data since the introduction in microarrays and was further developed for count-data in RNAseq data. DESeq2 and edgeR, which do permit shrinkage were shown to have a very high performance on metagenomic data in Paper I. Shrinkage has the largest impact when the group size is small as very little information is available for gene-specific estimates and outliers are more likely to occur.

Finally, metagenomic data has been observed to be sparse with an excess number of zeros. Using the zero-inflated model of Paper III we confirmed that metagenomic data does contain excess zeros compared to what is expected from the underlying model assumptions used today. When not accounted for, extra zeros cause the variance estimates of affected genes to be highly biased. Thus, not incorporating for zero-inflation will reduce the power to detect differences in abundance for affected genes. The gain in performance from using zero-inflated methods is especially large when the number of samples is high.

The analysis of metagenomic data is maturing to a point where standard protocols and software packages are available. However, many parts are still under constant development. Improvements in sequencing technologies will lead to larger amounts of higher quality data and the potential to study even more specific functions. Reference databases are continually and rapidly expanding as new sequence data is added. Although care has to be taken in the curation of these databases (Bengtsson-Palme et al., 2016), it will still lead to more gene families being identified and correctly annotated. New algorithms are being developed to process the ever expanding quantities of data. This thesis has shown that the statistical analysis of metagenomic data also requires further work. First, this thesis has highlighted that using flawed modelling assumptions will cause unreliable results. To stop using those models that do not account for between-sample variability and to encourage replicated experimental designs, is of vital importance to establish a basis for sound statistical analysis. Second, the work presented in this thesis shows that statistical modelling of metagenomic data can provide large benefits in increasing the power to detect differentially abundant genes. However, further research is needed to fully understand and utilize the complex variance structure exhibited by metagenomic data.

Bibliography

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106.

Anscombe, F. (1948). The transformation of poisson, binomial and negative-binomial data. *Biometrika*, 35.

Bengtsson-Palme, J., Boulund, F., Edström, R., Feizi, A., Johnning, A., Jonsson, V. A., Karlsson, F. H., Pal, C., Pereira, M. B., Rehammar, A., Sanchez, J., Sanli, K., and Thorell, K. (2016). Strategies to improve usability and preserve accuracy in biological sequence databases. *PROTEOMICS*, 16(18):2454–2460.

Bengtsson-Palme, J., Boulund, F., Fick, J., Kristiansson, E., and Larsson, D. (2014). Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in india. *Frontiers in microbiology*, 5:648.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Castro-Nallar, E., Shen, Y., Freishtat, R. J., Pérez-Losada, M., Manimaran, S., Liu, G., Johnson, W. E., and Crandall, K. A. (2015). Integrating microbial and host transcriptomics to characterize asthma-associated microbial communities. *BMC medical genomics*, 8(1):1.

Consortium, G. O. (2015). Gene ontology consortium: going forward. *Nucleic acids research*, 43(D1):D1049–D1056.

Curtis, T., Sloan, W., and Scannell, J. (2002). Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences of the United States of America*, 99(16):10494–9.

DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N.-U., Martinez, A., Sullivan, M. B., Edwards, R., Brito, B. R., et al. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science*, 311(5760):496–503.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, pages 71–103.

Fang, H., Cai, L., Yu, Y., and Zhang, T. (2013). Metagenomic analysis reveals the prevalence of biodegradation genes for organic pollutants in activated sludge. *Bioresource technology*, 129:209–218.

Fang, R., Wagner, B., Harris, J., and Fillon, S. (2016). Zero-inflated negative binomial mixed model: an application to two microbial organisms important in oesophagitis. *Epidemiology and infection*, pages 1–9.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., Owens, S., Gilbert, J. A., Wall, D. H., and Caporaso, J. G. (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences*, 109(52):21390–21395.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94.

Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Gilbert, J. A., Jansson, J. K., and Knight, R. (2014). The earth microbiome project: successes and aspirations. *BMC biology*, 12(1):1.

Greenblum, S., Carr, R., and Borenstein, E. (2015). Extensive strain-level copy-number variation across human gut microbiome species. *Cell*, 160(4):583–594.

Grzymski, J. J., Riesenfeld, C. S., Williams, T. J., Dussaq, A. M., Ducklow, H., Erickson, M., Cavicchioli, R., and Murray, A. E. (2012). A metagenomic assessment of winter and summer bacterioplankton from antarctica peninsula coastal surface waters. *The ISME journal*, 6(10):1901–1915.

Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K., and Beck, E. (2013). Tigrfams and genome properties in 2013. *Nucleic acids research*, 41(D1):D387–D395.

Handelsman, J., Rondon, M., Brady, S., Clardy, J., and Goodman, R. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry & Biology*, 5.

Healy, F., Ray, R., Aldrich, H., Wilkie, A., Ingram, L., and Shanmugam, K. (1995). Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. *Applied microbiology and biotechnology*, 43(4):667–674.

Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., and Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences*, 111(13):4904–4909.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., et al. (2015). eggnog 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic acids research*, page gkv1248.

Hugenholtz, P. and Tyson, G. (2008). Microbiology: metagenomics. *Nature*, 455(7212):481–3.

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214.

Jeffries, T. C., Ostrowski, M., Williams, R. B., Xie, C., Jensen, R. M., Grzymski, J. J., Senstius, S. J., Givskov, M., Hoeke, R., Philip, G. K., et al. (2015). Spatially extensive microbial biogeography of the indian ocean provides insights into the unique community structure of a pristine coral atoll. *Scientific reports*, 5.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., et al. (2008). Kegg for linking genomes to life and the environment. *Nucleic acids research*, 36(suppl 1):D480–D484.

Karlsson, F., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C., Fagerberg, B., J, N., and F, B. (2013). Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature*, 498:99–103.

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, Nand Fuhrman, J., Hugenholtz, P., van der Lelie, D., Meyer, F., Stevens, R., Bailey, M., Gordon, J., Kowalchuk, G., and Gilbert, J. (2012). Unlocking the potential of metagenomics through replicated experimental design. *Nature biotechnology*, 30(6):513–20.

Kristiansson, E., Fick, J., Janzon, A., Grabic, R., Rutgersson, C., Weijdegå rd, B., Söderström, H., and Larsson, D. (2011). Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. *PLoS One*, 6(2):e17038.

Kristiansson, E., Hugenholtz, P., and Dalevi, D. (2009). ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics (Oxford, England)*, 25(20):2737–8.

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., et al. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics*, 15(2):141–161.

Law, C., Chen, Y., Shi, W., and Smyth, G. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2):R29.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21.

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature*, 489(7415):220–230.

Mackelprang, R., Waldrop, M. P., DeAngelis, K. M., David, M. M., Chavarria, K. L., Blazewicz, S. J., Rubin, E. M., and Jansson, J. K. (2011). Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature*, 480(7377):368–371.

Manichanh, C., Rigottier-Gois, L., Bonnaud, E., Gloux, K., Pelletier, E., Frangeul, L., Nalin, R., Jarrin, C., Chardon, P., Marteau, P., et al. (2006). Reduced diversity of faecal microbiota in crohn's disease revealed by a metagenomic approach. *Gut*, 55(2):205–211.

Mann, H. B. and Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. 18:50–60.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. London England Chapman and Hall, 2 edition.

McMurdie, P. J. and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol*, 10(4):e1003531.

Mende, D. R., Waller, A. S., Sunagawa, S., Järvelin, A. I., Chan, M. M., Arumugam, M., Raes, J., and Bork, P. (2012). Assessment of metagenomic assembly using simulated next generation sequencing data. *PloS one*, 7(2):e31386.

Morgan, J., Darling, A., and Eisen, J. (2010). Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS ONE*, 5:e10209.

Nayfach, S. and Pollard, K. S. (2016). Toward accurate and quantitative comparative metagenomics. *Cell*, 166(5):1103–1116.

Noyes, N. R., Yang, X., Linke, L. M., Magnuson, R. J., Cook, S. R., Zaheer, R., Yang, H., Woerner, D. R., Geornaras, I., McArt, J. A., et al. (2016). Characterization of the resistome in manure, soil and wastewater from dairy and beef production systems. *Scientific reports*, 6.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic acids research*, 33(17):5691–5702.

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, 276(5313):734–740.

Parks, D. and Beiko, R. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*, 26(6):715–21.

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10(12):1200–2.

Pedrós-Alió, C. (2006). Marine microbial diversity: can it be determined? *Trends in microbiology*, 14(6):257–263.

Peng, X., Li, G., and Liu, Z. (2016). Zero-inflated beta regression for differential abundance analysis with metagenomics data. *Journal of Computational Biology*, 23(2):102–110.

Prosser, J. I. (2010). Replicate or lie. *Environmental microbiology*, 12(7):1806–1810.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D., Li, J., Xu, J., Li, S., Li, D., Cao1, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Paslier, D., Linneberg, A., Nielsen, H., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill1, J., Weissenbach, J., MetaHIT Consortium, Bork, P., Ehrlich, D., and Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464:59–65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E., Renault, P., Pons, N., Batto, J.-M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li, S., Yang, H., Wang, J., Ehrlich, S., Nielsen, R., Pedersen, O., Kristiansen, K., and Wang, J. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers. *BMC genomics*, 13(1):1.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N., Anderson, I., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B., Gies, E., Dodsworth, J., Hedlund, B., Tsiamis, G., Sievert, S., Liu, W.-T., Eisen, J., Hallam, S., Kyrpides, N., Stepanauskas, R., Rubin, E., Hugenholtz, P., and Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499:431–437.

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

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):1.

Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J., and Taylor, J. M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using rna-sequencing. *BMC genomics*, 13(1):1.

Rodriguez-Brito, B., Rohwer, F., and Edwards, R. (2006). An application of statistics to comparative metagenomics. *BMC bioinformatics*, 7:162.

Roesch, L. F., Fulthorpe, R. R., Riva, A., Casella, G., Hadwin, A. K., Kent, A. D., Daroub, S. H., Camargo, F. A., Farmerie, W. G., and Triplett, E. W. (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME journal*, 1(4):283–290.

Rondon, M. R., August, P. R., Bettermann, A. D., Brady, S. F., Grossman, T. H., Liles, M. R., Loiacono, K. A., Lynch, B. A., MacNeil, I. A., Minor, C., et al. (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Applied and environmental microbiology*, 66(6):2541–2547.

Ross, E., Moate, P., Marett, L., Cocks, B., and Hayes, B. (2013). Investigating the effect of two methane-mitigating diets on the rumen microbiome using massively parallel sequencing. *Journal of dairy science*, 96(9):6030–6046.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of molecular biology*, 94(3):441–448.

Sanli, K., Karlsson, F., Nookaew, I., and Nielsen, J. (2013). FANTOM: Functional and taxonomic analysis of metagenomes. *BMC bioinformatics*, 14(1):38.

Schloss, P. and Handelsman, J. (2005). Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome biology*, 6(8):229.

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864.

Scholz, M. B., Lo, C.-C., and Chain, P. S. (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current opinion in biotechnology*, 23(1):9–15.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome biology*, 12(6):R60.

Seqc/Maqc-Iii Consortium (2014). A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature biotechnology*, 32(9):903–914.

Smith, R. J., Jeffries, T. C., Roudnew, B., Fitch, A. J., Seymour, J. R., Delpin, M. W., Newton, K., Brown, M. H., and Mitchell, J. G. (2012). Metagenomic comparison of microbial communities inhabiting confined and unconfined aquifer ecosystems. *Environmental microbiology*, 14(1):240–253.

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

Sohn, M. B., Du, R., and An, L. (2015). A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics*, page btv165.

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

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., et al. (2015). Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359.

Turnbaugh, P., Hamady, M., Yatsunenko, T., Cantarel, B., Duncan, A., Ley, R., Sogin, M., Jones, W., Roe, B., Affourtit, J., Egholm, M., Henrissat, B., Heath, A., Knight, R., and Gordon, J. (2009). A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–4.

Turnbaugh, P., Ley, R., Hamady, M., Liggett, C., Knight, R., and Gordon, J. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, 449:804–810.

Unterseher, M., Jumpponen, A., Öpik, M., Tedersoo, L., Moora, M., Dormann, C. F., and Schnittler, M. (2011). Species abundance distributions and richness estimations in fungal metagenomics–lessons learned from community ecology. *Molecular Ecology*, 20(2):275–285.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426.

Ward, T. L., Hosid, S., Ioshikhes, I., and Altosaar, I. (2013). Human milk metagenome: a functional capacity analysis. *BMC microbiology*, 13(1):1.

White, J., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS computational biology*, 5(4):e1000352.

Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*, 95(12):6578–6583.

Wilcoxon, F. (1946). Individual comparisons of grouped data by ranking methods. *Journal of economic entomology*, 39:269.

Wooley, J., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS computational biology*, 6(2):e1000667.

Wooley, J. and Ye, Y. (2009). Metagenomics: Facts and Artifacts, and Computational Challenges*. *Journal of computer science and technology*, 25(1):71–81.

Yatsunenko, T., Rey, F., Manary, M., Trehan, I., Dominguez-Bello, M., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R., Anokhin, A., Heath, A., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J., Lozupone, C., Lauber, C., Clemente, J., Knights, D., Knight, R., and Gordon, J. (2012). Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–7.

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., and Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, 18(1):4.