Statistical modelling and analyses of DNA sequence data with applications to metagenomics

Mariana Buongermino Pereira

CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

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

Statistical modelling and analyses of DNA sequence data with applications to metagenomics
Mariana Buongermino Pereira
Gothenburg 2017
ISBN 978-91-7597-607-5

Doktorshavhandlingar vid Chalmers tekniska högskola
Ny serie nr 4288
ISSN 0346-718X

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

Cover illustration: *The study of DNA provides a better understanding of bacterial communities* by Mariana Buongermino Pereira

Typeset with LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2017

To my parents

Statistical modelling and analyses of DNA sequence data with applications to metagenomics

Mariana Buongermino Pereira

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

Abstract

Microorganisms are organised in complex communities and are ubiquitous in all ecosystems, including natural environments and inside the human gut. Metagenomics, which is the direct sequencing of DNA from a sample, enables studying the collective genomes of the organisms that are there present. However, the resulting data is highly variable, and statistical models are therefore necessary to assure correct biological interpretations.

This thesis aims to develop statistical models that provide an increased understanding of metagenomics data. In Paper **I**, we develop, implement and evaluate HattCI, which is a high-performance generalised hidden Markov model for the identification of integron-associated *attC* sites in DNA sequence data. In Paper **II**, we implement HattCI and other bioinformatics tools into a computational method to identify and characterise the biological functions of integron-mediated genes. The method is used to identify 13,397 integron-mediated genes present in metagenomic data. In Paper **III**, we provide a conceptual overview of the computational and statistical challenges involved in analysing gene abundance data. In Paper **IV**, we perform a comprehensive evaluation of nine normalisation methods for metagenomic gene abundance data. Our results highlight the importance of using a suitable method to avoid introducing an unacceptably high rate of false positives.

The methods presented in this thesis improve the analysis of metagenomic data and thereby the understanding of microbial communities. Specifically, this thesis highlights the importance of statistical modelling in addressing the large variability of high-dimensional biological data and ensuring its sound interpretation.

Keywords: generalised hidden Markov models, normalisation, statistical modelling, metagenomics, DNA sequence data, gene abundance data, bioinformatics.

List of publications

This thesis is based on the work represented by the following papers:

I. **Pereira, M.B.**, Wallroth, M., Kristiansson, E., Axelson-Fisk, M. (2016). HattCI: fast and accurate *attC* site identification using hidden Markov models. *Journal of Computational Biology*, **23**(11):891-902, doi: 10.1089/cmb.2016.0024.

II. **Pereira, M.B.**, Österlund, T., Eriksson, K.M., Backhaus, T., Axelson-Fisk, M., Kristiansson, E. (2017). A comprehensive survey of integron-mediated genes present in metagenomes. *Manuscript*.

III. Boulund, F., **Pereira, M.B.**, Jonsson, V., Kristiansson, E. (2017). Computational and statistical considerations in the analysis of metagenomic data. In M. Nagarajan (Ed.), *Metagenomics: perspectives, methods, and applications*. Elsevier. *Accepted*.

IV. **Pereira, M.B.**, Wallroth, M., Jonsson, V., Kristiansson, E.. (2017). Comparison of normalization methods for the analysis of metagenomic gene abundance data. *Manuscript*.

Additional papers not included in this thesis:

V. Boulund, F., Johnning, A., **Pereira, M.B.**, Larsson, D.G.J., Kristiansson, E. (2012). A novel method to discover fluoroquinolone antibiotic resistance (qnr) genes in fragmented nucleotide sequences. *BMC Genomics, **13**:695*, doi: 10.1186/1471-2164-13-695.

VI. **Pereira, M.B.**, Verma, C.S., Fuentes, G. (2013). Differences in the binding affinities of ErbB family: Heterogeneity in the prediction of resistance mutants. *PLoS ONE, **8**(10):e77054*, doi: 10.1371/journal.pone.0077054.

VII. **Pereira, M.B.**, Sato, K., Kristiansson, E., Axelson-Fisk, M. Improved identification of *attC* sites by secondary structure modeling. *Manuscript in Licentiate thesis*, 2015.

VIII. 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., Thorell, K. (2016). Strategies to improve usability and preserve accuracy in biological sequence databases. *PROTEOMICS, **16**(18):2454-60*, doi: 10.1002/pmic.201600034.

IX. Johnning, A., Karami, N., Hallbäck, E.T., Müller, V., Nyberg, L., **Pereira, M.B.**, Stewarte, C., Ambjörnssone, T., Westerlund, F., Adlerberth, I., Kristiansson, E. A detailed analysis of the resistomes of six carbapenem-resistant pathogens. *Submitted*.

Author contributions

I. Created the model, implemented the first version of the algorithm in R, supervised the C implementation, developed model improvement study design, manually curated the dataset used to train and test the HMM, participated in the performance study design, implemented and executed performance tests, performed data analysis, wrote the online documentation, and drafted and edited the manuscript.

II. Created and tested the model for the *attC* site secondary structure. Implemented the computational pipeline. Participated in the planning of the analyses to be conducted. Performed functional annotation of the integron-mediated genes and all the other analyses presented. Drafted and edited the manuscript.

III. Planned and edited the statistical part of the manuscript. Drafted the part about normalisation. Participated in deciding the methods to include. Performed the comparison of normalisation methods.

IV. Participated in the design of an evaluation framework that preserves the structure of data variability. Contributed to the study design, including the choice of methods to evaluate and performance measures to use. Supervised the execution of the preliminary analysis. Implemented the final workflow in R. Performed the analysis of the results. Drafted and edited the manuscript.

Acknowledgements

Pursuing a PhD is fantastic 5-year journey, and as in any fantastic adventure, you need support to overcome the obstacles! I am very thankful to have had so many wonderful people around me to help me reach the end of this journey. First of all, I wish to thank **Erik Kristiansson**, my supervisor and an overall role model as a scientist! Your constant enthusiasm has always been a source of inspiration and motivation to proceed. Our interesting discussions and how you showed me the joy of even the less interesting tasks that I encountered were essential to my education. This thesis is the result of your constant support and encouragement through these years, thank you! Next, I would like to thank **Marija Cvijovic**, my co-supervisor and mentor, for always encouraging me to go further. Your advice about life after the PhD and how you believed in me lead me to the next step in my career! Additionally, to **Olle Nerman**, my examiner, thank you for always having a moment to answer my questions and for caring. Thank you, **Marina Axelson-Fisk**, for the supervision during the first years of my PhD studies.

A very special thank you is dedicated to my best friend and collaborator **Viktor Jonsson**; I am so happy to have gotten to know you. You definitely taught me much about the detection of differentially abundant genes, but above all, you taught me about life! These 5 years would not have been the same without you and our daily conversations about virtually anything. Thank you for being you! Moreover, thank you, **Tobias Österlund**, my friend, co-author, and neighbour; your kindness has always been a great joy. I learned much with you when we supervised the bachelor project together and anytime I had questions about GO terms. Thank you, **Anna Johnning**, for the loads of energy you always radiate and for all answering my questions on the most diverse topics. Thank you, **Anna Rehammar**, for complementing me so well in the leukaemia collaboration and all the teaching that we did together. Thank you, **Fredrik Boulund**, for introducing me to metagenomics. Thank you, **Fanny Berglund**, for your happiness and support with our server Terra. Thank you, **Jonatan Kallus**, for your calm and pleasant lunch company. Thanks, **Johannes Dröge**, for interesting conversations about scientific ideas. Thank you also to some former and new members of the bioinformatics group: **Anders Sjögren**, **Emma Wijkmark**, **Henrik Imberg**, and **Mikael Gustavsson**.

Thank you to my collaborators, who are such an important part of this thesis. Specially, thank you, **Mikael Wallroth**, for being a perfect student; the work you did on two different projects was great! Thank you, **Erik Malmberg** and **Linda Fogelstrand**, for the fruitful collaboration and for teaching me about leukaemia research. Thank you, **Kengo Sato**, for your hospitality when I was in Japan and for teaching me about stochastic context-free grammar. Thank you, **Joakim Larsson** and your research group, for the after-works and exchange of ideas about metagenomics and bacteria. In particular, thank you, **Johan Bengtsson-Palme**, for your friendship, patience with my Brazilianness, and endless enthusiasm about research and academic life! Thank you, **Patricia Huijbers**, for so quickly becoming a friend. Thank you, **Martin Eriksson**

and **Thomas Backhaus**, for providing part of the data used in Paper **II**. Thank you, the GoBig team, for nice discussions on bioinformatics, which led to the database paper; some of you were mentioned before, but let me extend thanks to **Amir Feizi**, **Chandan Pal**, **Francesco Gatto**, **Fredrik Karlsson**, **Kaisa Thorell**, **Kemar Sanli**, and **Robert Edström**. Special thanks to you, **Jose Sanchez**, GoBig member, and friend, for teaching me statistics during my masters and for all the friendship, dinners, parties, and travels. In addition, thank you, **Chandra Verma** and **Gloria Fuentes**, for the collaboration on structural bioinformatics; what I learned from you helps my work until today! Thank you, bachelor and master students that I supervised, **Andreas Nilsson**, **Anna Källsgård**, **Anton Martinsson**, **Deimante Neimantaite**, **Emily Curry**,
Rikard Isaksson, and **Sara Finati**; I learned much with you. Finally, thank you guys in Cvijovic's lab, **Damiano Ognissanti**, **Felix Held**, **Jacob Leander**, **Johannes Borgqvist**, **Niek Welkenhuysen**, and **Qasim Ali**, for accepting a statistician as a member of your research group.

Thank you, **Cecilia Gelin**, **Jovan Pankovski**, **Lotta Fernström**, and **Marie Kühn**, for the friendly and efficient support with all administrative questions around my studies.

Very importantly, thank you my dear friends! You guys have added variation to my life, and after all, what is the life of a statistician without some variability to play with? In the department, you certainly made it possible for me to keep smiling even on the greyest of days! First of all, thank you, **Magnus Ö**, you have *always* been there: to defeat me in a board game, to listen and to laugh: "Delad glädje är dubbel glädje, delad sorg är halv sorg!". **Henrike**, thank you for all the adventures that we shared! **Claes**, thank you for all the music shared, concerts and warm conversations. **Malin**, thank you for sharing with me your joy of having little Hedvig. **Vera**, thank you for the piano lessons, ping pong and afternoons studying together. Thank you, **David**, **Ivar**, and **Tuomas**, for very fun (board) gaming. **Magnus R**, **Anton** and **Natasha**, thank you for the good times outside work. **Robert**, thank you for making these good times even more interesting with the curiosities you find on Wikipedia. Additionally, thank you, **Anders**, **Anna-Kaisa**, **Edvin**, **Elizabeth**, **Maud**, **Olle**, **Sandra**, and **Valentina**, for company and nice conversations during lunch, fika, and after-works.

Thank you also to my friends outside the department. You reminded me that there is life out there! Thank you my little sister, **Juna**, for all the countable but infinitely many moments of happiness we shared, in particular, any time singing around town! **Parastoo** and **Peiman**, thank you for always protecting me. **Renan** and **Mauro**, thank you for sharing memories about our home country. **Ninon** and **Fredrik**, thank you for always having a warm smile and delicious food. Thank you, **Håkan**, my first boss in Sweden, for believing in me and for teaching me so much about Swedishness.

Thank you, my parents, **Regina** and **Carlos**, sister, **Priscilla**, and brother-in-law, **Theo**. Even far, you are always so present. Your unconditional love and support are the basis for all my achievements! *Saudades*!

Mariana Buongermino Pereira
Gothenburg, August 2, 2017

The eternal mystery of the world is its comprehensibility.

ALBERT EINSTEIN
Physics and Reality, 1936

Contents

1 Background

This chapter introduces the biological concepts used in this thesis. In particular, integrons, which are a bacterial genomic element used as a mechanism to share genetic material between cells, are introduced along with the principles of metagenomics, a tool for studying the collective DNA of microorganism communities.

1.1 Microorganisms

Microorganisms are organisms that are so small that they cannot be seen by the naked eye. The study of microorganisms, known as microbiology, began in 1676 with Antonie van Leeuwenhoek's introduction of the microscope. Currently, the advances in DNA sequencing technology enable a considerably more detailed study of microorganisms, their genomes and genes. Microorganisms can be unicellular or multicellular and are found across the three domains of life: Archaea, Bacteria and Eukarya (Woese et al., 1990). Eukaryotes are uni- or multicellular organisms whose cells have their genetic material organised in a nucleus. Archaea and Bacteria are unicellular and prokaryotes, i.e. their only cell does not contain a nucleus. While the origin of life is debatable, it is known that Eukarya is evolutionary closer to Archaea than to Bacteria (Gribaldo et al., 2010). Bacteria are found almost everywhere. For example, they inhabit the human body, including the gut, where they live in symbiosis with us, participating in our digestion and constituting vital parts of other metabolic processes. Similarly, bacteria can live in symbiotic relationships with other animals and plants. Some bacterial species, however, are pathogenic and can be harmful, causing diseases in their hosts. Bacteria are also ubiquitous in environmental ecosystems, e.g. in water and soil. In the environment, bacteria are essential in the recycling of nutrients, for example, by decomposing dead bodies or by fixating nitrogen from the atmosphere. Some bacterial species have evolved to live in extreme conditions, such as in hydro-thermal vents or in the deepest parts of the oceans (Glud et al., 2013). In such environments, bacteria have developed special metabolisms to sustain life. Despite their

importance, most bacterial species have not yet been characterised (Hug et al., 2016; Solden et al., 2016). Since more than than 98% of bacterial species cannot easily be cultivated in the laboratory (Amann et al., 1995; Hugenholtz et al., 1998; Rinke et al., 2013), it is important to continue studying bacteria using techniques that avoid cultivation.

1.2 Integrons and *attC* sites

Bacterial cells are able to share genetic material through a process known as horizontal gene transfer (HGT). In contrast to vertical exchange, in which genetic material is inherited by the offspring from parents, HGT enables bacteria to share genetic material directly between cells. This ability to horizontally exchange genes has enabled bacteria to rapidly adapt to environmental changes. For instance, at least six bacterial phyla have a photosynthesis-based metabolism, which uses light to generate energy, and it has been shown that HGT supports the spread of photosynthesis among different bacterial species (Swithers et al., 2012; Raymond et al., 2002; Xiong et al., 2002). More recently, with the introduction of the human use of antibiotics, genes that confer resistance against antibiotics have been acquired in clinical pathogens from environmental bacteria via HGT (Von Wintersdorff et al., 2016).

One common mechanism of HGT is acquiring exogenous mobile genes in the genome through a genomic element known as an integron. Each gene mobilised by an integron is organised into a gene cassette that can be incised (Hall et al., 1991), excised (Collis and Hall, 1992) or rearranged (MacDonald et al., 2006) by the integron using site-specific recombination. Integrons are found in a wide range of bacterial species and share the same common structure (Figure 1.1). They all carry one gene that codes for an integrase, which is the enzyme that mediates the transfer process; a common recombination site *attI* used by the integrase during the transfer process and where the incoming gene is incorporated; and one integron-associated promoter (Pc) that regulates the expression of incorporated genes. In addition, an array of gene cassettes is found to be sequentially incorporated downstream of the promoter. Each cassette typically contains one gene and an *attC* site, which is the other recombination site recognised by the integrase during the transfer (Stokes and Hall, 1989; Mazel, 2006).

Integrons are ancient structures that are estimated to be present in the genomes of approximately 6% of bacterial species (Cury et al., 2016). These structures can be found both on conjugative elements, as in pathogens such as *Escherichia coli* and *Salmonella enterica* (Partridge et al., 2009), or on chromosomes, as in e.g.

Vibrio spp. and *Xanthomonas* spp. (Mazel, 2006). With the introduction of the human use of antibiotics, mobile integrons, i.e. integrons located on mobile elements, have facilitated the spread of antibiotic resistance genes (Partridge et al., 2009). Their spread, together with their associated genes, is facilitated by the mobility provided by the elements that they are located on. Specifically, class 1 integrons are a good example of what integrons are capable of in terms of adaptation. Their association with a Tn402-like transposon, a genomic element that is capable of changing its position in the genome, enabled the spread of antibiotic resistance genes already carried in integrons when the use of antibiotics by humans was introduced (Gillings, 2014). In contrast to mobile integrons, chromosomal integrons are only inherited vertically. These integrons are generally not mobile on their own, and they can carry more than 200 gene cassettes (Cambray et al., 2010; Mazel, 2006). These large and diverse sets of genes together with the integron capability of incorporating and expressing potentially any gene suggest that integrons can play a major role in the adaptation and evolution of many forms of bacteria (Boucher et al., 2007; Holmes et al., 2003). Indeed, in some species, e.g. *Vibrio cholerae* and *Vibrio fisheri*, the presence of integrons in the genome is known to predate speciation (Mazel, 2006).

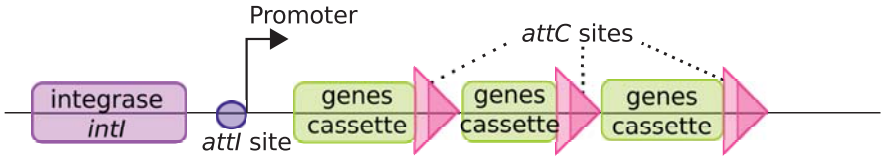

Figure 1.1: Integrons are elements of the bacterial genome that facilitate the horizontal transfer of genes. All integrons share a common structure: an *intI* gene that codes for an integrase, which is the enzyme responsible for mediating the gene incorporation; a promoter that regulates the expression of the incorporated genes; a recombination site *attI* used during the transfer process; and a number of gene cassettes, each of them with one inserted gene and one *attC* site, another recombination site used during the transfer process.

Each integron-mediated gene is typically accompanied by an *attC* site. These sites can be used as markers for genes that are mobilised by integrons. *AttC* sites are relatively short (approximately 55-141 nucleotides) with a characteristic pattern that makes them recognisable. Specifically, *attC* sites are imperfect reverse palindromes (Figure 1.2a), whose bottom DNA strand folds into a hairpin-like secondary structure during the transfer (Figure 1.2b) (Cambray et al., 2010). Secondary structures are supported by hydrogen bonds formed between the base pairs $\{AT, GC, GT\}$, which are said to be complementary in this case. The *attC* site secondary structure contains two pairs of complementary motifs called R″/R′ (marked with pink in Figure 1.2) and

L”/L’ (purple) that are recognised by the integrase (Stokes et al., 1997; Hall et al., 1991). These pairs of motifs are separated by two short spacers (green), and the two L-motifs are separated by a central loop (blue). Despite these well-described motifs, only six nucleotides, three in each R-motif, are perfectly conserved (bold). Therefore, a model that describes these sites needs to take the variability in nucleotide composition and in length into account. See the introduction of Paper **I** (Pereira et al., 2016) for a more detailed description of the different parts of an *attC* site.

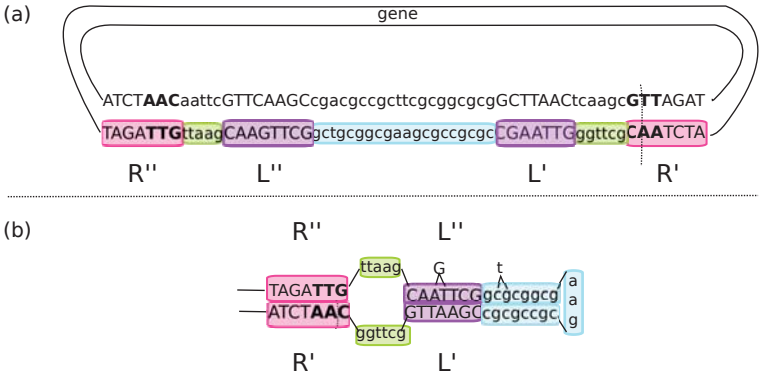

Figure 1.2: Example of an *attC* site. (a) View of a gene cassette in its circular form (outside the genome). The motifs recognised by the integrase, R”/R’(pink) and R”/R’(purple), are indicated in the bottom strand. Boldfaced nucleotides are perfectly conserved. Two spacers (green) and one central loop (blue) separate these motifs. **(b)** Fold of the bottom strand of an *attC* site. The hairpin-like secondary structure of the fold is required for the gene transfer. The recombination position is indicated by the horizontal line in R’.

Since integron-mediated genes may confer adaptive advantages to their hosts, identifying such genes can ultimately improve our understanding of bacterial evolution. The first two papers in this thesis address integrons and their gene cassettes. Specifically, Paper **I** presents a generalised hidden Markov model (gHMM) to describe *attC* sites, and Paper **II** constructs and uses a computational pipeline based on the model presented in Paper **I** and other tools to screen sequence data to search for gene cassettes.

1.3 Metagenomics

A metagenome is the collection of the DNA from all microorganisms in a community. Metagenomics is the study of this collective genome. In shotgun

metagenomics, this is performed by directly sequencing the DNA present in a sample. This approach enables studying a collective genome without the need for cultivating individual isolates or any prior knowledge or assumption regarding the sample content. Shotgun metagenomics is thus an important tool in microbiology for understanding the functional and genetic composition within a certain community or in comparison to other communities. In contrast to other genomics studies, in which one bacterium species is isolated, cultivated in the lab, and sequenced, metagenomics is performed on a sample from a microorganism community from the environment or the gut of a human or other animal. Consequently, a mixture of DNA sequences from the organisms found in that community is obtained. When studying microorganisms in a community, it is generally of interest to understand its biodiversity either through a species analysis, where the goal is to determine what species are present in the community, or through a gene analysis, where the goal is to determine which functions are performed by the genes found in the community (Wooley et al., 2010). In such studies, metagenomics is crucial since the sequencing of individual species requires cultivation in the lab, which, as stated previously, can only be performed using standard procedures for approximately 1-2% of bacterial species (Amann et al., 1995; Hugenholtz et al., 1998; Rinke et al., 2013). Metagenomics, however, enables the study of a community without the need for cultivation in the lab, which provides an unbiased and holistic overview at the genomic level of how bacteria behave *in natura*, who they live with and what they do.

After one or more microbial communities have had their DNA sequenced, computational and statistical analyses are conducted to characterise these communities. This characterisation can be descriptive if each community is described independently or comparative if the aim is to understand differences between groups of communities. In descriptive metagenomics, characterising the content and biodiversity of the community is often of interest. For example, bacterioplankton from brackish environments had their genomes first observed through the reconstruction of metagenomic data (Hugerth et al., 2015). Additionally, the metagenomes in the human gut (Yatsunenko et al., 2012) and oceanic waters (Sunagawa et al., 2015) were described for different locations across the globe. Alternatively, a more specific characteristic may be of interest. For example, genes that confer resistance against antibiotics were described in polluted lakes, highlighting the potential existence of resistance genes in the environment that might have not yet been observed in clinics (Bengtsson-Palme et al., 2014; Boulund et al., 2012). Conversely, in comparative metagenomics, two or more environments have their DNA contents compared to highlight potential differences. For example, we can consider the comparison between a polluted environment and a pristine environment, where we aim to understand which genes confer an adaptive advantage un-

der different environmental conditions (Kristiansson et al., 2011; Mason et al., 2014), or we can consider the comparison between the bacteria found in the human guts of healthy and sick individuals, where we may want to determine whether a functional genetic class has been lost and can potentially be replaced to help the treatment (Qin et al., 2012). In this thesis, Papers **I** and **II** address descriptive metagenomics, whereas Papers **III** and **IV** address comparative metagenomics.

1.4 Challenges in the interpretation of metagenomic sequencing data

In shotgun metagenomics, a sample of DNA from a microorganism community is sequenced without any prior knowledge of the sample content. Although this approach provides the flexibility to describe the community without bias, the lack of knowledge poses several challenges for data interpretation. In general terms, DNA is sequenced in short reads, assembled into longer contigs and then mapped to some reference such that species or functions found in the community can be described or quantified. Modern high-throughput DNA sequencing technologies, also known as next-generation sequencing (NGS), are able to produce large volumes of data in a short time and at a low cost (Heather and Chain, 2016). To achieve this, however, the resulting sequenced DNA is delivered in randomly generated, short reads of up to 300 nucleotides, which is shorter than the length of most genes. These short reads then need to be assembled, i.e. to be combined into longer stretches of DNA called contigs. In many genomic studies, it is typically known *a priori* which species was sequenced; thus, the assembly is typically guided by a reference genome of the same species, if available, or by a close relative. If no reference is available, as is the case in metagenomics, an assembly without a reference genome can be performed, the so-called *de novo* assembly. In these cases, the assembly is generally guaranteed by sequencing the DNA with high coverage, i.e. with a high average number of times each position in the genome is sequenced, or, as the name suggests, covered by a sequence read. Then, because the reads are randomly generated, some of them are likely to overlap, which makes the assembly possible. Note that repetitive regions are commonly found in genomes, which means that one read may overlap with two or more reads that belong to distant parts of the genome but that have the same repetitive region. Thus, these repeats make the assembly a non-trivial problem for both guided and *de novo* assembly. For *de novo* assembly in particular, because of the lack of a reference, the large amount of short reads also increases the likelihood of spurious overlaps, and all reads have to be directly or indirectly compared to each

other, thus making the process more complex (Paszkiewicz and Studholme, 2010). In metagenomics, the situation is further complicated by the lack of knowledge of sample composition; neither which species are present in a sample nor their relative abundances are known. This lack of *a priori* knowledge makes the assembly of the short random reads back into a long DNA stretch quite difficult. Generally, the data is assembled into contigs that are often short, occasionally only a few thousand nucleotides long. Consequently, the subsequent data analyses often required by descriptive metagenomics, such as annotation, gene count and functional prediction, become more challenging than in traditional genomics studies and normally require specialised tools that can robustly handle fragmented data (Escobar-Zepeda et al., 2015).

Another challenge in metagenomic analysis is correctly estimating the relative abundances of genes. When we want to quantify the relative abundances of elements, such as genes or gene families, for which reference databases are available, the sequenced reads can be directly mapped into the database without the need to assemble the data. Then, the number of matches to each entry in the database can be quantified by simply counting the number of matched reads. This is often the case in comparative metagenomics, where the differences in these abundances are detected between two or more communities. The next step is to apply a statistical test to detect which genes have a significant difference in abundance between the investigated experimental conditions. The difficulties in such an analysis are associated with the nature of the data, namely, i) the data is high-dimensional, and there are typically several thousands of genes in the communities (Galperin et al., 2015); ii) the data is discrete, which prevents the application of standard statistical tests that assume normality; and iii) the data has a high variability, which can, if not taken into account, result in low statistical power. A large part of the variability is, however, systematic, i.e. exhibits patterns that can, at least partially, be detected and removed. In this thesis, we define *systematic variability* as a source of noise that affects all the genes in the sample in a similar, systematic manner. A large source of systematic variability associated with NGS is due to the library size, i.e. the number of total sequenced reads varies for each sample. Thus, for the study of relative gene abundances, such variability must be taken into account. One approach to correct systematic variability is to normalise the data. Normalisation of the data aims to make the gene abundances in all samples comparable by adjusting the scale at which the counts are compared. Normalisation is an essential step for ensuring that any subsequent analysis, such as the detection of differently abundant genes, is reliable.

In short, metagenomic data, both as DNA sequence and gene abundance, comes in large volumes to be processed and with high variability to be modelled. The sequence data is highly fragmented, whereas the gene abundance data is

discrete and under-sampled due to their high dimensionality. Taken together, these characteristics of the data requires a careful choice of statistical models to provide correct biological interpretations and conclusions.

2 Aims

The overall aim of this thesis is to provide a means to further improve the interpretation of metagenomic data. A better interpretation can be achieved by extracting more information from the data or by improving the quality of the actual data. In both cases, modelling can be applied to describe the phenomena of interest. In this thesis, Papers **I** and **II** are dedicated to the development and application of a model to extract more information from metagenomic sequence data, whereas Papers **III** and **IV** discuss and evaluate model-based methods to reduce variability in metagenomic gene abundance data. Specifically, the aims of this thesis are as follows:

- To develop a model to identify integron-associated *attC* sites in DNA sequence data (Paper **I**).
- To implement a method for detecting *attC* sites and their associated integron-mediated genes in metagenomic DNA sequence data (Paper **II**).
- To characterise the biological functions of integron-mediated genes (Paper **II**).
- To provide a conceptual overview of computational and statistical considerations in comparative metagenomic studies (Papers **III** and **IV**).
- To develop a framework for the evaluation of normalisation methods that preserves the variability structure of metagenomic data (Paper **IV**).
- To compare the statistical performance of commonly used normalisation methods in metagenomic gene abundance data (Papers **III** and **IV**).

3 Statistical modelling and analyses

Statistical models and analyses investigate the variability of a natural phenomenon and its data, in order to make predictions that ultimately expand our knowledge about the phenomenon under investigation. This thesis addresses the application of such models and analyses to metagenomes, with the aim of obtaining a better understanding of bacterial communities. Figure 3.1 presents an overview of how the different techniques discussed in this thesis support the study of bacteria using metagenomics.

3.1 Modelling of sequence data

In descriptive metagenomics, it is often of interest to describe the DNA being studied in terms of the genes or other genomic elements that they contain. Such a description requires metagenomic data to be at least partially assembled since these elements can be considerably longer than the short reads generated by modern sequencing technology. After the data is assembled or partially assembled, the goal is to annotate the data. For instance, a major task is to describe where the genes are located in the sequence through a process known as gene finding (Axelson-Fisk, 2015). Additionally, other functional genomic elements, such as promoters, CpG islands, and recombination and binding sites, may be of interest. Genes or other genetic elements have specific motifs that characterise them, but they commonly have different variants. These variations occur in terms of nucleotide composition and quantity. For instance, one genome has thousands of genes that share similar patterns organised into smaller motifs, which determine where they start and stop and how the nucleotides are organised inside the gene, but each gene has a different function and is unique in which order and how many nucleotides that it carries. In addition, genes with the same function found in different species present small variations from each other. Other elements are also seldom perfectly conserved

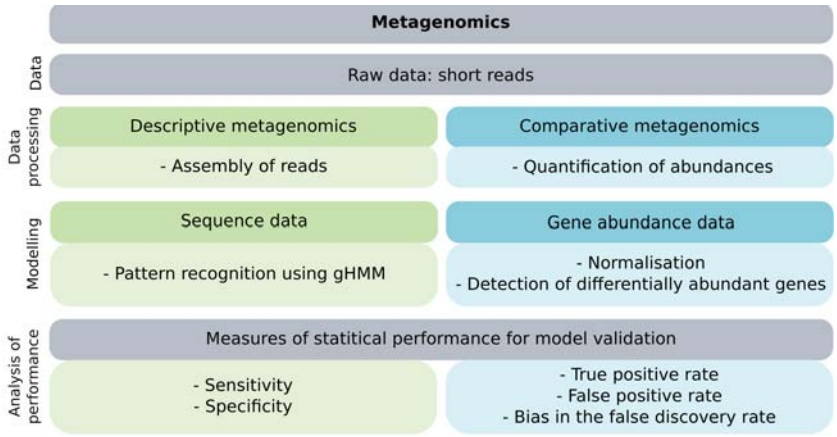

Figure 3.1: Overview of the statistical models and analyses studied in this thesis in connection with their application to metagenomic data. Metagenomic DNA sequence data is characterised by short reads. To describe the genomic elements associated with horizontal gene transfer in bacterial communities (green), the data needs to be assembled such that model(s) for pattern recognition can be applied. Conversely, the comparison of communities to determine which genes are community-specific (blue) requires the reads to be mapped into a reference database such that gene abundances can be quantified. Model-based methods can then be applied to remove systematic variability and detect differentially abundant genes. In both cases, measures of statistical performance can be used to validate the models, taking their ability to control false positives and to detect true positives into account.

across species. Probabilistic models, rather than straightforward pattern matching, are therefore often the preferred approach for identifying these patterns in DNA sequences. This is even more important for metagenomic data, where the species being analysed is not known *a priori*, which increases the variability to be addressed.

3.1.1 Generalised hidden Markov models

A hidden Markov model (HMM) is a probabilistic model that describes an observed phenomenon as a sequential series of events and assigns labels to these events. The sequence of labels is the hidden part of the model. In the case of searching for genomic elements in DNA sequences, the observed phenomenon is the sequence itself, i.e. a sequence of the nucleotides $\{A, C, G, T\}$s, and the labels to be assigned indicate whether a nucleotide is a part of the element of interest. Formally, a HMM is composed of two interrelated probabilistic

processes: one hidden process, which is Markovian, and one observed process, which is not necessarily Markovian, generated by the hidden process. Then, let the hidden process $Y = \{Y_1, \ldots, Y_T\}$ be a stochastic process that jumps between a set of states $S = \{s_1, \ldots, s_N\}$ such that for each step t, $Y_t = j$, where $j \in S$. For DNA sequences, each state represents one characteristic motif of the functional genomic element of interest. The process Y is then said to be Markovian if it obeys the Markov property, i.e. if given the present state, the past and future are independent, such that

$$P(Y_t|Y_{t-1}, Y_{t-2}, \ldots, Y_1) = P(Y_t|Y_{t-1}).$$

Note that the stochastic process that follows the above is said to be a first-order Markov process since the state in t depends only on one previous state. This dependency can be extended to depend on k previous states; thus, we can have a k-th order Markov process.

The Markov process progresses by jumping between states at each step t (note that a jump can also mean remaining at the current state). The jumps are defined by transition probabilities that the hidden process goes from state i to j, i.e. $a_{ij} = P(Y_t = j|Y_{t-1} = i)$, $i, j \in S$. The observed process $X = \{X_1, \ldots, X_T\}$ is generated by the hidden process such that at step t with $Y_t = j$, an X_t is defined by the emission probability $b_j(X_t|X_1^{t-1}) = P(X_t|X_1^{t-1}, Y_t = j)$. Thus, the emission probability defines that state j emits the observation X_t given, if no restriction is imposed, all the previous emissions. Note that imposed restrictions can be adjusted to reflect the nucleotide composition and their dependencies in the genomic element of interest. For instance, we can restrict X_t to be dependent up to k previous emissions such that the emissions become a Markov process of order k. Then, the HMM defines the joint probability $P(X, Y)$ of a hidden sequence Y and an observed sequence X, given by

$$P(X, Y) = \prod_{t=1}^{T} a_{Y_{t-1}Y_t} b_{Y_t}(X_t|X_{t-k}^{t-1}),$$

where a_{Y_0,Y_1} is defined as the initial probability $\pi_j = P(Y_1 = i)$, i.e. the probability that the sequence starts at state i.

A HMM generates elements, i.e. segments of X where all X_t belong to the same state (or label), whose durations (or lengths) are defined by the number of steps where the process remains in the same state. These lengths are governed by the transition probabilities and can be shown to follow a geo-

metric distribution. However, a geometric distribution does not always realistically describe the observed length of genomic elements. Generalised hidden Markov models (gHMM) provide an alternative that relaxes the assumption of the HMM and allows the lengths to follow any distribution of state durations. In this way, each hidden state $Y_t = j$ has a duration (or length) d_t sampled from a distribution $f_j(d_t)$, and it emits an observed subsequence $X_{u-d_t+1}^{u} = \{X_{u-d_t+1}, ..., X_u\}$ according to the emission probability $b_j(X_{u-d_t+1}^{u}|X_1^{u-d_t}) = P(X_{u-d_t+1}, ..., X_u|X_1, X_2, ..., X_{u-d_t}, Y_t = j)$. Since each state can now emit more than one observation, the indices of the hidden and the observed processes differ. To keep track of the differing indices, partial sums $p_t = \sum_{k=1}^{t} d_k$, where $p_0 = 0$, are introduced such that the total length of X is $U = \sum_{t=1}^{T} d_t$. The joint probability of X, Y and the sequence of state durations $d = d_1, ..., d_T$ is then given by (Russell and Moore, 1985; Levinson, 1986)

$$p(X, Y, d) = \prod_{t=1}^{T} a_{Y_{t-1},Y_t} f_{Y_t}(d_t) b_{Y_t}(X_{p_{t-1}+1}^{p_t}|X_1^{p_{t-1}}), \tag{3.1}$$

which can be used for inference and parameter estimation.

Parameter estimation After the gHMM has been formulated, i.e. states have been established, a family of distributions for the state durations has been chosen and the dependencies in the emission probabilities determined, the unknown parameters need to be estimated. Parameters characterise the model such that it can be used for inference on new data. In the case of DNA sequence data, this means that the model can be used to find the genomic element of interest in a given DNA sequence. Parameter estimation requires training data; thus, the parameters are determined to maximise their likelihood given the data under the model. A gHMM determines the joint probability of the observed sequence X and the hidden sequence Y. Thus, for such a model, training data consists of a set of pairs of sequences (X, Y). For DNA sequence data, complete training data corresponds to a set of DNA sequences (X) for which the positions of genomic elements are known (Y). Such data can, for instance, be collected from literature searches or well-annotated sequences present in public databases. When complete training data (X, Y) is available, parameters can be computed using maximum likelihood estimation. In this case, the likelihood of the data is given by the model (Equation 3.1), and the parameter estimations are found to maximise the likelihood of the parameters with respect to the complete training data. In a gHMM, the set of parameters $\theta = \{a_{i,j}, b_j(\cdot), f_j(d)\}$ has their maximum likelihood estimates given by the following:

$$a_{ij} = \frac{c_{ij}}{c_i}$$

$$b_j(x) = \frac{c_i(x)}{c_i}$$

$$f_j(d) = \frac{c_i(d)}{c_i}$$

where c_i is the counts for $Y_t = i$ in the training data, c_{ij} is the counts of transitions from state i to state j, $c_i(x)$ is the counts for state i emitting an observation x, and $c_i(d)$ is the counts for state i having a duration d. Note that counts $c_i(d)$ can alternatively be used to fit a predefined family of distributions.

When the hidden sequence Y is not available, the model can be used to determine X and then estimate the parameters in an iterative manner. In such cases, the parameters can be estimated using the Baum-Welch algorithm, a variant of the EM algorithm, which alternates between an expectation (E) step and a maximisation (M) step to find the optimum parameters. The basic idea is to start with initial guesses of the parameter θ, use them with the model to determine Y and then use (X, Y) to compute the expectation of the likelihood given by the model (Equation 3.1) as a function of θ in the E-step. Then, the estimation of θ is updated such that the expectation of the likelihood is maximised in the M-step. This procedure is repeated until convergence. For further details of the Baum-Welch algorithm, see Axelson-Fisk (2015).

Parsing sequences The major application of a gHMM is to infer the sequence of hidden states, or labels, given an observed sequence. In the case of DNA sequence data, this task corresponds to labelling stretches of DNA with the genomic elements that they belong to, i.e. each nucleotide in the DNA sequence is labelled as a characteristic motif of a gene, promoter, recombination site, or any other type of genomic element that it belongs to. In computational terms, the observed sequence is the input, which is parsed to determine the hidden sequence. For a gHMM, the hidden sequence Y is determined as the sequence of states that maximises the likelihood of the observed sequence X, i.e. we want to determine a sequence of states that maximises the probability $P(Y, d|X)$. Note that this probability maximises at the same point as $P(X, Y, d)$, given by Equation 3.1. The task is then to find an efficient computational approach to maximise $P(X, Y, d)$, which can be achieved by dynamic programming. In this process, the computations are broken down into recursive relations such that for each nucleotide, a computation is performed using the computation for

previous nucleotides, which are stored for this recursive use.

Specifically, the Viterbi algorithm is the dynamic programming technique for inferring the most probable hidden sequence. For each position u in the observed sequence X, the Viterbi variable $\delta_j(u)$ for each state j is defined as the maximum joint probability of the observed sequence X_1^u up to position u, the hidden sequence Y_1^t up to position t ending at state j and the partial sum of durations d of the states in Y, which sums to u, i.e. $p_t = u$. Using the model in Equation 3.1, we can find recursive relations to $\delta_j(u)$ (see Axelson-Fisk (2015) for details), such that

$$\begin{aligned}\delta_j(u) &= \max_{t, Y_t^{t-1}, d_1^t} P(X_1^u, Y_1^{t-1}, Y_t = j, p_t = u) \\ &= \max_{i,d} \delta_j(u-d) a_{ij} f_j(d) b_j(X_{u-d+1}^u | X_1^{u-d})\end{aligned}$$

The states i and durations d that maximise $\delta_j(u)$ for each position u are stored, and the optimal hidden sequence is found by tracing back these states and durations.

A gHMM for the identification of *attC* sites is presented in Paper **I**. In the model, the DNA sequence is the observed process with $\{A, C, G, T\}$ as emissions. Each motif of an *attC* site is one possible state of generalised duration, and a non-generalised state describes the non-*attC* site stretches of DNA. In Paper **II**, the model is applied for identifying *attC* sites as markers of integron-mediated genes. For further details on HMM, gHMM, and their applications to bioinformatics, see Rabiner (1989); Durbin et al. (1998); Axelson-Fisk (2015).

3.2 Modelling of gene abundance data

In comparative metagenomics, detecting differences between communities from different environments or clinical conditions is often of interest. Such differences may be defined in terms of relative gene abundances, i.e. the relative number of times that a certain gene appears in a sample. This analysis requires the sequenced DNA reads to be mapped to reference genes. These genes are often organised as a reference database, which typically contains annotated genomes, *de novo* assembled contigs, or a catalogue of genes or profiles that represent gene families. The resulting data is the counts of reads that match each entry in the database. After the reads have been mapped or quantified,

systematic variability has to be removed or at least minimised, and then a statistical test to detect differentially abundant genes can be applied.

3.2.1 Normalisation

Systematic variability affects all samples in a similar manner and creates a recognisable pattern. Normalisation aims to identify this pattern and remove, or at least substantially reduce, this variability. In the case of gene abundances derived from sequencing data, where DNA reads are mapped into a reference database to quantify the gene abundances in a sample, one of the main sources of systematic variability is the differences in library size (Marioni et al., 2008). Library size is the number of reads generated during the DNA sequencing. This number depends on the sequence platform, and it typically varies substantially between samples. Thus, for the same community, if we have two samples, where one has twice the library size of the other, the gene abundances will be on average twice as large on the first sample only due to this technical issue and not to any relevant biological variation. Other sources of systematic variability include variations in DNA quality, sample handling, quality of the sequencing run and bioinformatical errors produced in the gene quantification step. In addition, the number of reads mapped to a gene will depend on the length of a gene (Oshlack and Wakefield, 2009) and on the average genome size of the community studied (Nayfach and Pollard, 2015). When combined, these factors will result in an unwanted between-sample variation that can reduce the statistical power. Thus, when studying gene abundance, each sample needs to be normalised to make the samples comparable.

Several methods have been proposed for normalising gene abundance data. To remove systematic variability, most methods operate by computing a set of normalisation factors, one per sample, which re-scales the samples such that the counts become comparable. The most straightforward normalisation method, which is known as *total counts*, uses the sum of all counts, i.e. the sum of the abundance of all genes in a sample, as its normalisation factor (Marioni et al., 2008; Mitra et al., 2009; White et al., 2009). Since the total abundance in a sample is dominated by highly abundant genes, the *upper quartile* method was proposed as a robust alternative. The upper quartile method sets the normalisation factor for each sample as its 75% quantile of the gene abundance distribution (Bullard et al., 2010). Even if the upper quartile method avoids the variability carried by highly abundant genes, it calculates the normalisation factor based on absolute gene abundance values. Conversely, the *trimmed mean of M-values* (TMM) uses differences between samples to derive the normalisation factor. For this purpose, the method calculates the

ratio between the counts in each sample and a reference, which is typically set to one of the studied samples. Extreme values of these ratios are trimmed, as well deviating absolute values. The normalisation factor is then robustly calculated from the ratios of gene abundances in the sample (Robinson and Oshlack, 2010). In addition, normalisation methods that do not re-scale the data has been proposed. Among them, the *rarefying* method is a commonly used approach (McMurdie and Holmes, 2013). In this method, a reference sequence depth is set, which is often the minimum sequence depth among all samples to be normalised, and then counts are randomly discarded without replacement until all samples have the reference sequence depth. Other commonly used normalisation methods are median, DESeq2, cumulative sum scaling, and quantile-quantile. Please refer to Paper **III** for details and references on these methods.

A large number of normalisation methods have been proposed. However, it is unclear which methods are the most suitable for the normalisation of gene abundance data produced by shotgun metagenomics. The choice becomes particularly difficult given that some methods have been developed for other types of data, in which they exhibit a satisfactory performance. Thus, there is a need for a comprehensive evaluation of normalisation methods for metagenomic gene abundance data. In Paper **IV**, we performed such an evaluation and showed that the final result may substantially differ depending on which normalisation method is applied to the data.

3.2.2 Detection of differentially abundant genes

The comparison between different experimental conditions, representing different bacterial communities, involves detecting which genes are more abundant in one condition than in the other(s). For this purpose, a statistical test is used to assess the difference in the abundance of each gene to determine whether it is caused by random fluctuations ('noise') or whether it is a true biological effect ('signal') (Jonsson et al., 2016). However, metagenomic gene abundance data is affected by large variability. As previously discussed, the systematic part of this variability can be partially removed by normalisation, whereas the remaining variability typically does not present any detectable patterns and needs to be statistically modelled. Furthermore, since gene abundances are estimated from counting reads, the model needs to take into account that the data is discrete and follow a non-normal distribution. Moreover, the data generally contains a few thousand genes and at most a hundred samples, which further complicates the analysis.

Several statistical methods have been proposed for the analysis of gene abun-

dance data. These methods can be grouped into three categories depending on the approach that they use. The first category uses variance stabilising transformations to convert the counts into approximately normally distributed data (Paulson et al., 2013; Sohn et al., 2015; Law et al., 2014). The main advantage of these methods is that they work under the assumption of a normal distribution and are therefore flexible and available in a wide range of computer software. However, the transformation can rarely be performed perfectly; thus, the assumption of normality will often be violated, which may decrease their performance. The second category of methods consists of count-based models that describe the gene abundance data using discrete distributions, such as Poisson, over-dispersed Poisson or negative binomial distributions (Love et al., 2014; Kristiansson et al., 2009; Robinson et al., 2010). These models typically provide a more realistic representation of the data but are generally only available through specialised software packages. The last category consists of non-parametric methods that avoid any parametric distributional assumptions when describing the data (White et al., 2009; Segata et al., 2011; Parks et al., 2014). Rather, these methods estimate the distribution under the null hypothesis (e.g. no difference in gene abundance between two conditions) using, for instance, data permutations or asymptotics. The drawbacks of these methods are that they require larger sample sizes to achieve good performance and that they may be sensitive to ties in the data, which can be common when working with counts. In short, several models have been proposed in each of these categories, where each has its own advantages and disadvantages.

In Paper **III**, we present a detailed list of nine statistical models in these three categories along with a discussion of their performances. Briefly, we observed that methods that use normality assumptions have the lowest performance, count-based models generally presented the best performances, and non-parametric models suffered from small sample sizes. In Paper **IV**, for comparing the normalisation methods, we used a count model based on an over-dispersed Poisson distribution, which has previously been shown to have a high and stable performance (Jonsson et al., 2016).

3.3 Measures of statistical performance

When evaluating a statistical model, we need measures to quantify its performance. This includes measurements of both the ability to correctly detect the signal of interest (sensitivity) and the ability to avoid the incorrect classification of noise as signal (specificity). In other words, a good model should have both a high sensitivity and a high specificity. In the case of a model for sequence data, high sensitivity corresponds to correctly detecting the genomic structures

of interest, such as the *attC* sites in Papers **I** and **II**, and a high specificity represents avoiding labelling other DNA stretches as such. For a model of gene abundance data, a high sensitivity corresponds to correctly detecting the differentially abundant genes, whereas a high specificity ensures that non-differentially abundant genes are not incorrectly identified as differentially abundant genes.

Estimating the performance of a statistical model ideally requires a test dataset that contains both positive (signal) and negative (noise) examples. Then, if a positive example is detected by the model, it is labelled a true positive (TP); otherwise, it is a false negative (FN). Similarly, if a negative example is detected by the model, it is labelled a false positive (FP); otherwise, it is a true negative (TN) (Burset and Guigó, 1996). Sensitivity is then estimated as

$$Sensitivity = \frac{TP}{TP + FN},$$

while specificity is estimated as

$$Specificity = \frac{TN}{FP + TN}.$$

Note that in the above, TP, FN, FP and TN are the total counts of the corresponding labels predicted in the test dataset by the model.

A set of negative examples can, in some cases, be difficult to assemble since an exact definition of a counter-example to the signal may not be straightforward to determine. For example, in Papers **I** and **II**, negative examples were created by reshuffling a bacterial genome, which created a dataset that did not contain any signal of interest. Specificity was then measured as the number of false positives predicted per megabase, i.e. 10^6 nucleotides, of analysed sequence data. Another example where negative examples are not easily defined is presented in Paper **IV**. Here, the problem was further complicated since neither positive nor negative examples were known in the data. We addressed the problem by repeatedly re-sampling the data and by introducing signals using a binomial model.

Then, in Paper **IV**, sensitivity and specificity were adjusted to address the high dimensionality of the data. The genes and their associated p-values were sorted such that the genes with the most significant difference in abundance were found at the top of the resulting list. Note that the list contains both the positive and the negative examples. All p-values smaller than a pre-defined significance

level were then set as differentially abundant genes and were labelled as TP or FP according to the nature of positive or negative example, respectively. The number of TPs divided by the total number of positive examples subsequently defined the true positive rate (TPR), which describes the overall sensitivity for the entire dataset. Similarly, the number of FPs divided by the total number of negative examples defined the false positive rate (FPR), which describes the overall specificity.

For a single statistical test, the significance level α controls the probability of incorrectly rejecting the null hypothesis (i.e. generate a false positive). In the case of metagenomic gene abundance data, statistical tests are independently performed for each of the many thousands of genes present in the data. Thus, each test has a probability of producing a false positive. To interpret the data, the total number of false positives needs to be estimated. Multiple-testing correction procedures have therefore been developed and can be used to control the false positive rates when several statistical tests are performed simultaneously (Dudoit and van der Laan, 2008). One such method is the Bonferroni correction, which replaces the significance level α by $\frac{\alpha}{n}$, where n is the number of tests. This ensures that the probability of one or more false positives in the entire dataset is less than α. However, this procedure can be too conservative for the analysis of metagenomic data. Indeed, in most cases where n is very large and there is a substantial number of truly differentially abundant genes, a small fraction of false positives is often acceptable. Rather, the false discovery rate (FDR) can be estimated, which is defined as the expected proportion of false positives among the genes classified as significant. Here, the Benjamini-Hochberg algorithm can be used to control the FDR by calculating the adjusted p-value. Then, if the genes with an adjusted p-value below a certain significance level (e.g. 0.05, 0.10 or 0.20) are classified as significant, the FDR is ensured to be below this level (Benjamini and Hochberg, 1995).

The Benjamini-Hochberg algorithm is currently the most common approach to control for multiple testing in high-dimensional biological data. However, the algorithm depends on several assumptions, including independence of the hypothesis tests and specific distributions of p-values under the null hypothesis. If these assumptions are violated, the FDR estimates may be biased. It is thus important to evaluate how good the estimated FDRs are under different situations, including different datasets, normalisation methods and statistical models. In Paper **IV**, we evaluated the impact of normalisation on the ability to control the FDR when using the Benjamini-Hochberg algorithm. This evaluation was performed by comparing the FDR estimated by the algorithm with the true FDR defined as

$$trueFDR(k) = \frac{FP(k)}{k}$$

where $FP(k)$ is the number of false positives found up to position k in the ordered list of p-values.

A biased FDR can have significant effects on the interpretation of the analysis. In particular, the estimated FDR is used to determine how many of the genes are differentially abundant. Thus, an unbiased FDR is necessary to ensure that the number of false positives is controlled. However, note that a biased FDR can be either conservative, resulting in too few significant genes, or too relaxed, resulting in too many false positives.

Finally, note that the evaluation of the statistical performance of a model needs to take more than one aspect into account. Specifically, measures should reflect both the capability of the model to detect the signal of interest and its ability to avoid generating false positives. This typically means that at least one measure reflecting each aspect should be used. The selection of measures may depend on the problem. In particular, the measure that reflects the ability of the model to control false positives may require special attention because negative examples are often not available. The choice of measure is thus an important aspect in the statistical analysis of a model and should be performed carefully.

4 Summary of results

In this chapter, the background, aims and main results of this thesis are summarised and organised according to the included papers. The figures presented in this chapter are taken from the corresponding papers.

4.1 Paper I

AttC sites are recombination sites that are required for the incision and excision of integron-mediated genes during horizontal gene transfers. Once in the DNA, these sites are part of the gene cassette, generally accompanying the gene that they transferred in a one-to-one correspondence. This makes *attC* sites good markers for genes that are mobilised in this way. *AttC* sites are imperfect palindromic repeats, with a length of 55 to 141 nucleotides, that acquire a secondary structure necessary for the gene transfer. The structure is supported by the presence of two pairs of moderately conserved motifs, R"/R'and L"/L', of 7 or 8 nucleotides in length that form two helices when the site is folded. Except for these motifs, the other parts of an *attC* site have low conservation across different gene cassettes. Moreover, the other parts of the site have a variable length. The low conservation combined with the wide range of possible lengths creates a large variability between *attC* sites. Probabilistic models are therefore required to accurately describe the patterns of *attC* sites.

Paper **I**, *HattCI: Fast and accurate attC site identification using hidden Markov models*, presented HattCI, a hidden Markov model (HMM) for *attC* site identification. The model is an eight-state generalised hidden Markov model (gHMM); thus, the length of each state can be explicitly modelled using any distribution rather than the geometric distribution of lengths imposed by standard HMMs. In the model, one state represents the non-*attC* site regions of DNA, and seven states describe the different parts of an *attC* site (Figure 4.1). Of those seven states, four states have a fixed length and correspond to conserved motifs of the *attC* site, whereas the other three states have their lengths modelled by an

empirical distribution. These variable-length states correspond to two short spacers and one longer central loop that separates the conserved motifs.

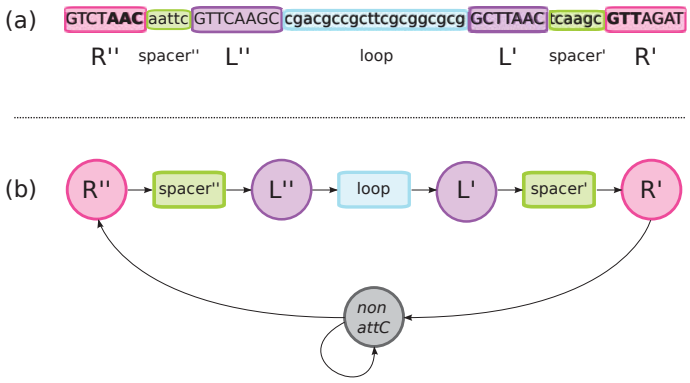

Figure 4.1: Overview of the gHMM implemented in HattCI. (a) The different parts of the *attC* sites are shown. **(b)** The HattCI model is shown with its 8 states and arrows indicating the possible state transitions. The parts of the *attC* sites are represented by 7 different states, and the remaining non-*attC* state represents sequences in between the *attC* sites. Circular states correspond to fixed-length parts, whereas rectangular states depict variable-length parts.

The model sensitivity was evaluated using a dataset with 231 manually annotated *attC* sites. Specificity was measured as a false positive rate, i.e. the number of hits found in the reshuffled *Escherichia coli* genome. The model parameters were then optimised by testing different options using two-fold cross-validation. In terms of pseudo-counts added to the counts of the partially conserved motifs, we observed that a higher number of pseudo-counts (up to 10) only increased the model performance when added to less-conserved motifs. In particular, L”/L’are less conserved than R”/R’. Thus, more pseudo-counts in L”/L’ increased the performance, whereas more pseudo-counts in R”/R’ decreased the performance. In addition, different distributions were tested to model the length of the central loop, the longest and most variable part of an *attC* site. Also, Markov models of different orders were tested to model the loop emissions. The best performance was achieved for a second-order Markov model and a smoothed empirical central loop length distribution. The final model performance was a 91.9% sensitivity and a false positive rate of 26.4 hits per megabase (Figure 4.2).

The model was applied to a test-case dataset containing a set of 35 metagenomic samples covering microbial communities found in different environments. Different amounts of *attC* sites were found in different environments. In particular, samples from polluted environments had up to 2.3 *attC* sites per

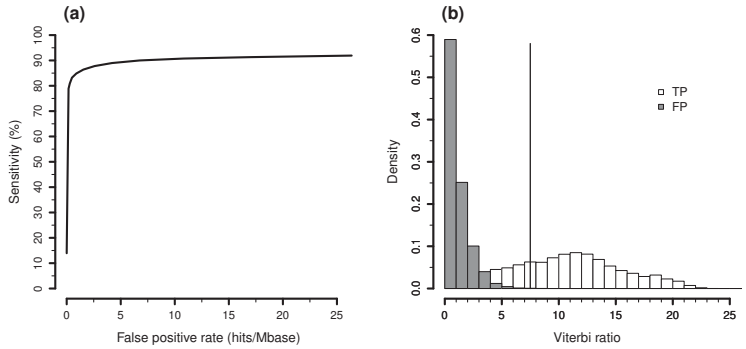

Figure 4.2: Cross-validation of the performance of the final model. (a) Receiver operating characteristic (ROC) curve for the final model based on 100 two-fold cross-validation iterations. Sensitivity in the reference dataset (y-axis) is plotted against the false positive rate estimated based on the reshuffled *Escherichia coli* K-12 genome (x-axis) for a varying significance cut-off k. **(b)** A histogram of the Viterbi scores for the true positives (TP) and false positives (FP). The Viterbi scores for the true positives (white) are substantially higher and relatively well separated from the false positives (grey). The vertical line indicates the significance cut-off of $k = 7.5$.

megabase, whereas some pristine environments had none (Figure 4.3). This result is consistent with the theory that polluted environments contain higher levels of integrons. The results demonstrated that our model can efficiently identify *attC* sites and provide a useful tool for the identification of integron-mediated genes.

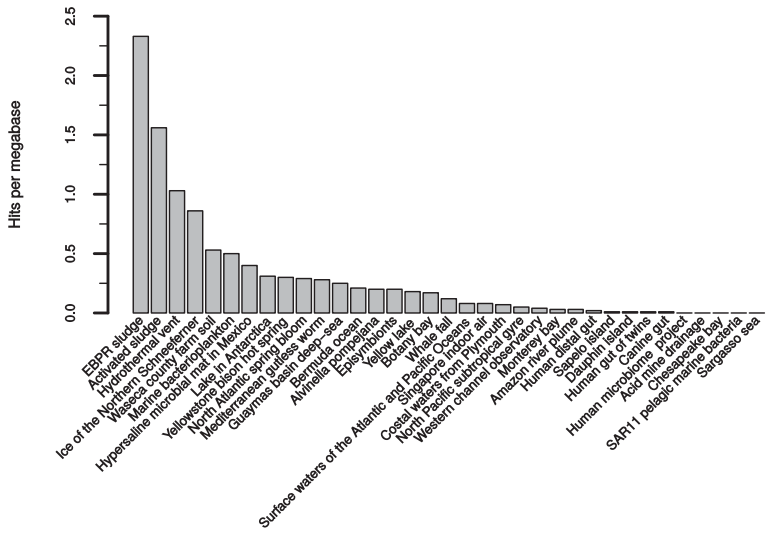

Figure 4.3: The number of *attC* sites found per megabase in metagenomic data. The number of *attC* sites found per megabase (y-axis) is shown for 35 metagenomic samples of the CAMERA project (x-axis).

4.2 Paper II

Integrons are known to carry genes with a wide range of functions that help bacteria adapt under periods of selective pressure. Studying integron-mediated genes can therefore provide a further understanding of bacterial evolution. However, more than 98% of bacterial species are not able to be cultivated in the lab using standard procedures. Therefore, an unbiased survey of integron-mediated genes requires the use of metagenomic DNA sequence data, where the need for cultivating bacteria is avoided and the collective genome of entire bacterial communities is observed instead. Integrons and their associated gene cassettes, i.e. the integron-mediated genes and their *attC* sites, have previously been studied in the clinical context, systematically in the genome of isolates or using targeted sequencing of metagenomes from specific environments. However, these studies have been limited to certain species of bacteria or to certain types of integrons.

Paper **II**, *A comprehensive survey of integron-mediated genes present in metagenomes*, presented a catalogue of integron-mediated genes found in metagenomes from diverse environments, including marine waters and the human gut. The analysis was performed by applying a computational pipeline to more than

10 terabases of data assembled in approximately 370 million contigs. The pipeline (Figure 4.4) first identified *attC* sites based on its conserved motifs using HattCI (Pereira et al., 2016). Then, their secondary structure was validated using a covariance model (CM), which was created from a set of 109 manually curated *attC* sites using Infernal (Nawrocki and Eddy, 2013). Next, false positives were removed by discarding *attC* sites that were isolated and that did not have a neighbouring *attC* site within 4,000 nucleotides. Finally, ORFs were predicted upstream of the *attC* sites using Prodigal (Hyatt et al., 2012).

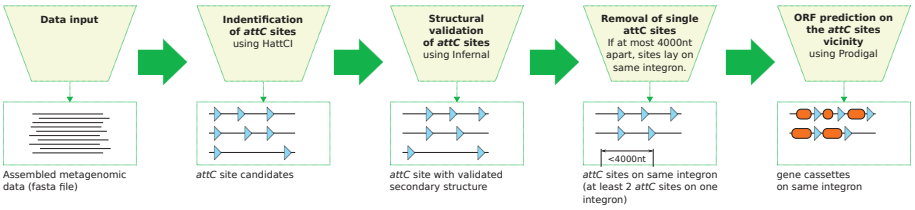

Figure 4.4: Flowchart for the identification of integron-mediated genes in metagenomic data. The computational pipeline starts by detecting *attC* sites and then identifies the associated downstream ORF. First, data is input as a fasta file containing assembled metagenomic DNA sequences. Next, HattCI, which implements a gHMM, is used to detect the *attC* sites present in the input sequences. Subsequently, the secondary structure of the detected *attC* sites is evaluated by the covariance model implemented in Infernal, which runs the search in its most sensitive mode. Identified *attC* sites on the same strand are considered to be part of the same integron when they are at max 4,000 nucleotides (nt) apart. Note that integrons with only one *attC* site are removed from the analysis. Finally, the ORFs are predicted upstream of the *attC* sites.

Consequently, we found 16,148 gene cassettes, consisting of 11,585 unique *attC* sites and 13,397 unique ORFs. A catalogue of integron-mediated genes was created using the predicted unique ORFs. The length of the genes in the catalogue was short, with a median of 402 nucleotides and standard deviation of 308 nucleotides (in comparison, the median length of chromosomal bacterial genes was 831 with a standard deviation of 735 nucleotides). The G/C-content of the genes in the catalogue varied substantially and was between 0.20 and 0.74 with a median of 0.50 and standard deviation of 0.09, which is a wider range than what was typically encountered within a single bacterial genome, where the G/C-content standard deviation was between 0.04 and 0.05.

The genes in the catalogue were functionally annotated using three general databases, Cluster of Orthologous Genes (2003-2014 COG) (Galperin et al., 2015), TIGRFAM 15.0 (Haft et al., 2003) and PFAM 29.0 (Finn et al., 2015), and two specialised databases, one containing antibiotic resistance genes (ResFinder) (Zankari et al., 2012) and one containing biocide and metal resistance

genes (BacMet) (Pal et al., 2014). In total, 5,183 (39%) of the genes had a significant match against at least one of the five databases (Figure 4.5a). In particular, 2,111 (16%) genes were matched to COG profiles with a known biological function (Figure 4.5b). The most common COG functional classes were defence mechanisms (23%), followed by transcription (15%) and mobility (12%). Specifically, the catalogue contained a wide range of genes associated with toxin-antitoxin systems, which are a system of two types of genes: one encoding a toxin that can destroy the host cell and one encoding an antitoxin that inhibits the toxin. Examples of identified genes associated with defence mechanisms were formaldehyde-activing genes, which allow bacteria to grow in environments were formaldehyde is present. Moreover, many genes in the catalogue contained a helix-turn-helix domain, which is a DNA binding domain often associated with transcription regulation but also part of the antitoxin vapI in the XRE-family, or found in prophage proteins, which are associated with genetic mobility. The gene catalogue also contained a wide range of transposases, which are enzymes that catalyses the movement of stretches of DNA. Their presence in gene cassettes may be explained by the fact that integrons are not mobile on themselves but their mobilisation is advantageous. Another common family of genes was endonucleases, which are enzymes that are capable of cleaving DNA, a property that can be used as defence mechanism by invalidating the genetic material of an incoming gene, for example, or as part of mobile mechanisms. In addition, acetyltransferases and methylases were commonly found among the catalogue, which are predicted to be associated with transcription.

Moreover, 38 resistance genes were present in our data, and of these genes, nine were novel and had not previously been reported in resistance gene databases. Our results suggest that integrons in environmental bacterial communities maintain resistance genes that have not yet taken the step into human pathogens, which further highlights these communities as a source of antibiotic resistance genes.

The wide range of functions found on the integron-mediated genes confirms that they are a reservoir of genes that can be harvested in times of selective pressure. Moreover, 8,214 genes (61% of the total number in the catalogue) did not match any of the three databases with known functional gene classes or domains or the two databases with antibiotic and metal resistance. This large amount of unknown genes indicates that many of the functions carried by integrons are yet to be discovered. It thus emphasises that much about bacteria is still not known and that additional studies are needed to unravel their biology.

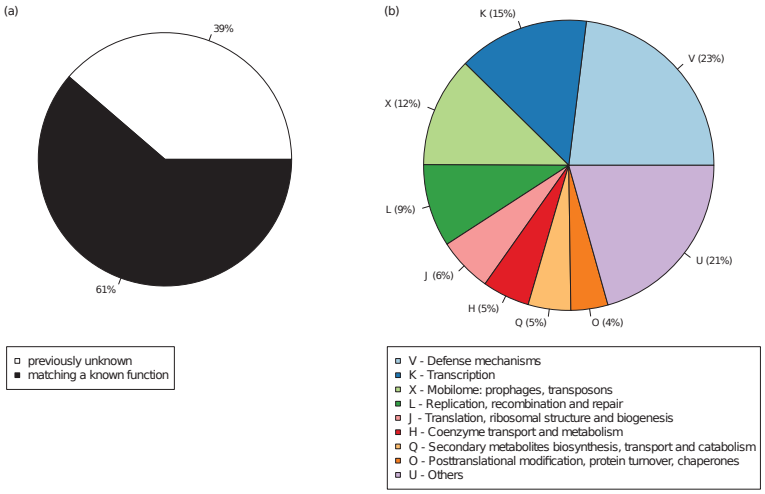

Figure 4.5: Biological functions of integron-mediated genes found in metagenomes. **(a)** 39% of the 13,397 integron-mediated genes found matched at least one entry in the COG, TIGRFAM, PFAM, ResFinder or BacMet databases. **(b)** Functional annotation of the integron-mediated genes using COG functional categories. Of the 13,397 integron-mediated genes in our catalogue, 2,111 genes matched a COG with a known function. Percentages on the plot are given in relation to this number. The eight most common functional classes are shown, and the remaining grouped under "U – Others".

4.3 Papers III & IV

Comparative metagenomics provides a powerful method for studying and highlighting the differences between microbial communities. These differences are often based on the total gene content, which reflects the biochemical potential of the communities. For the comparison, each gene in each community has to be quantified, i.e. counts of sequenced reads matching each gene have to be computed. The quantification poses computational challenges due to the large volumes of data and due to the large size of the gene databases since every sequenced read has to be compared to every entry in the reference database. Once quantified, a statistical test is applied to detect significant differences in the gene abundances between communities. The nature of the data presents further challenges to the statistical analysis, namely, the data is discrete and have high variability. This variability is largely due to systematic errors that can be addressed using normalisation methods. Several normalisation methods have been proposed to similar biological count data; however, their performance has not previously been studied for shotgun metagenomic data.

Paper **III**, *Computational and statistical considerations in the analysis of metagenomic data*, provides a conceptual overview of the challenges involved in the processing and statistical comparison of metagenomic data. This paper includes both the computational aspects related to addressing large metagenomic sequence datasets and the statistical issues related to the high variability of the data. In particular, in this paper, we showed that different methods for normalising gene abundances may result in different interpretations (Figure 4.6). We also concluded that while a systematic evaluation of normalisation methods had previously been performed for RNA-seq and species abundances estimated by amplicon metagenomic sequencing, no such comparison had been performed for shotgun metagenomic data. The lack of systematic comparisons of normalisation methods for gene abundance was the motivation for Paper **IV**, *Comparison of normalization methods for the analysis of metagenomic gene abundance data*. In this paper, nine commonly used normalisation methods were compared, focusing on the performance under different characteristics of the metagenomic data. See the legend of Figure 4.7 for a list of the methods used. Further details of the methods can be found in Papers **III** and **IV**.

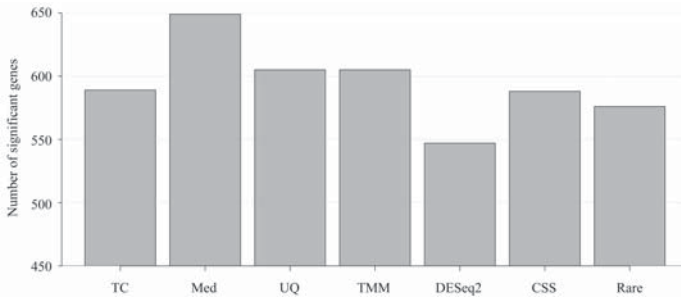

Figure 4.6: The choice of normalisation methods affects the total number of significantly differentially abundant genes identified in shotgun metagenomic data. Seven normalisation methods (total counts (TC), median (Med), upper quartile (UQ), TMM, DESeq2Norm, CSS and rarefying (Rare)) were applied to gene abundances from the human gut metagenomes of 71 diabetic and 74 healthy individuals. The genes were annotated using the TIGRFAM database, and the differences in gene abundance between the two conditions were assessed using an over-dispersed Poisson linear model. The figure shows the number of differentially abundant genes found using a false discovery rate (FDR) of less than 0.05. Note that a high number of significant genes does not necessarily indicate a better performance since the choice of normalisation method may affect the numbers of both true and false positives.

For the comparison, we implemented a framework to evaluate the methods that preserved the variability of the data. This was achieved using re-sampled metagenomic data, where differentially abundant genes were added in a controlled manner, rather than fully simulated data. The evaluation was performed

using three comprehensive datasets from metagenomic studies of human gut and marine samples (Qin et al., 2012; Yatsunenko et al., 2012; Sunagawa et al., 2015). For each dataset, a subset of samples from the same condition was re-sampled and divided into two groups. Effects were added by down-sampling DNA reads using a binomial model such that for a chosen gene, its counts were reduced in all samples of one of the two groups. The two groups then represented two artificial conditions that corresponded, for instance, to two different environments or different gut situations. The re-sampling with added effects provided a way to create several different scenarios, in which group size, proportion of genes with effect (i.e. the number of genes with effects), and effect size (i.e. probability used in the binomial model representing how large the average effect is) could be varied. Moreover, we varied the distribution of effects between groups (i.e. the proportion of effects added to each group). The combination of proportion and distribution of effects created balanced or unbalanced scenarios. Namely, we created 'balanced', where effects were added to 10% of the genes, equally distributed between the two groups; 'lightly unbalanced', where one group received 75% of the effects and the other received 25%; 'unbalanced', where all effects were added to only one group; and 'heavily unbalanced', where 20% of the effects were added to only one group. Performance was measured in terms of true positive rate, false positive rate and bias of false discovery rate.

In terms of true positive rate (TPR), i.e. the ability of the method to detect the effects that we introduced into the data, the performance of all methods monotonically decreased for more unbalanced effects (Figure 4.7). For the balanced case, all methods in all datasets had a TPR of at least 0.60, except for rarefying in the dataset with the lowest sequence depth. For some methods, such as quantile-quantile, upper quartile and median, the decrease in TPR between the balanced and heavily unbalanced cases was approximately 0.30, whereas for the TMM and DESeq2Norm methods, the decrease was approximately 0.20.

When we analysed the false positive rate (FPR), i.e. the proportion of incorrect differentially abundant genes detected among all genes with no effect, the performances of several of the methods were unsatisfactory, particularly in the heavily unbiased case (Figure 4.8). Most methods had a low number of FPRs for the balanced and lightly unbalanced cases. The exception was rarefying, which for Human gut II had a FPR of 0.011 already in the balanced case, whereas the other methods had an FPR of no more than 0.0022. In the unbalanced case, all methods showed an increased FPR. The increase was particularly large for quantile-quantile, upper quartile and rarefying with a FPR of up to 0.050. For the heavily unbalanced case, all methods exhibited an even higher FPR. The levels were particularly high for RCSS, quantile-quantile, upper quartile, median, total count and rarefying, with a FPR above 0.20 for upper-quartile in

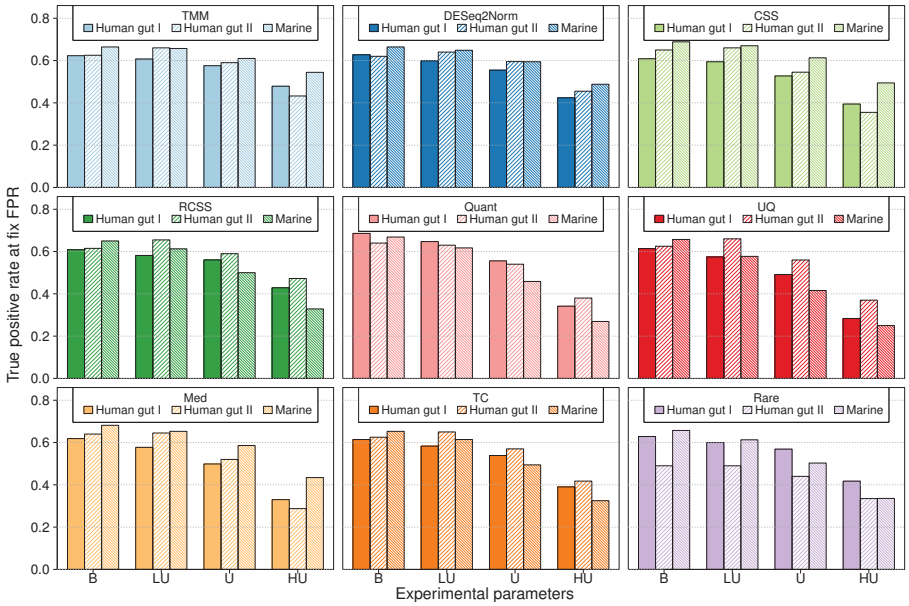

Figure 4.7: True positive rate at a fixed false positive rate of 0.01 (y-axis) for different distributions of effects between groups (x-axis): balanced ('B') with 10% of effects divided equally between the two groups, lightly unbalanced ('LU') with effects added 75%-25% in each group, unbalanced ('U') with all effects added to only one group, and heavily unbalanced ('HU') with 20% of effects added to only one group. The results were based on re-sampled data consisting of two groups with 10 samples in each and an average fold-change of 3. Three metagenomic datasets were used: Human gut I, Human gut II and Marine. The following methods are included in the figure: trimmed mean of M-values (TMM), DESeq2Norm, cumulative sum scaling (CSS), reversed cumulative sum scaling (RCSS), quantile-quantile (Quant), upper quartile (UQ), median (Med), total count (TC) and rarefying (Rare).

the marine dataset. This result indicates that the number of false positives in some cases surpassed the number of differentially abundant genes. Conversely, TMM, DESeq2Norm and CSS presented an overall stable performance, with a FPR of no more than 0.041.

The performance in terms of bias in the false discovery rate (FDR) estimation presented similar trends in terms of the best and worse methods as that observed in terms of TPR and FPR (Figure 4.9). For all methods, the FDR provided by the Benjamini-Hochberg algorithm was fixed to 0.05, and the corresponding true FDR was calculated for each method and dataset. For the balanced case, all methods were conservative and showed a true FDR that was smaller than

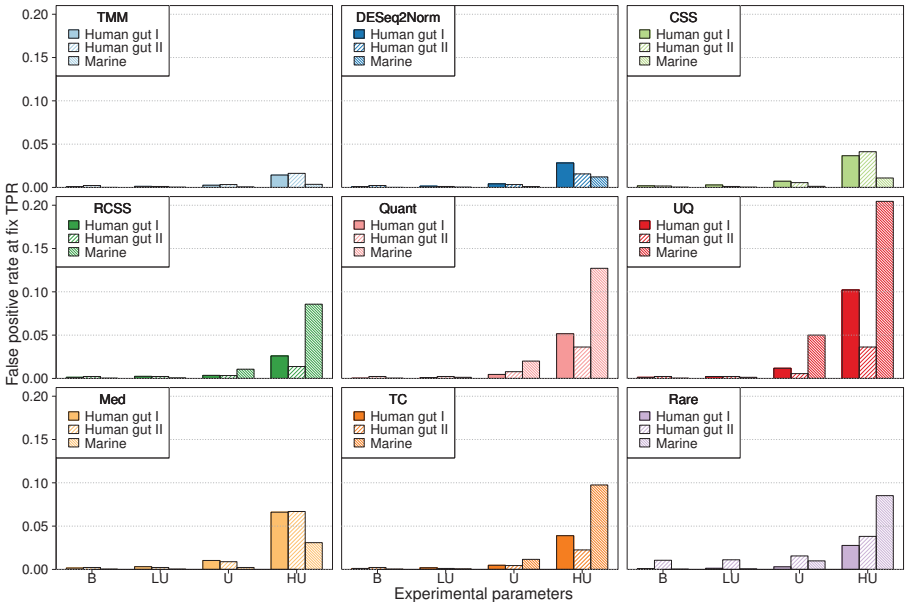

Figure 4.8: False positive rate at a fixed true positive rate of 0.5 (y-axis) for different distributions of effects between groups (x-axis): balanced ('B') with 10% of effects divided equally between the two groups, lightly unbalanced ('LU') with effects added 75%-25% in each group, unbalanced ('U') with all effects added to only one group, and heavily unbalanced ('HU') with 20% of effects added to only one group. The results were based on re-sampled data consisting of two groups with 10 samples in each and an average fold-change of 3. Three metagenomic datasets were used: Human gut I, Human gut II and Marine. The following methods are included in the figure: trimmed mean of M-values (TMM), DESeq2Norm, cumulative sum scaling (CSS), reversed cumulative sum scaling (RCSS), quantile-quantile (Quant), upper quartile (UQ), median (Med), total count (TC) and rarefying (Rare).

the estimated FDR. This changed, however, when the effects were added in an unbalanced manner. For the lightly unbalanced case, quantile-quantile and upper quartile already showed a true FDR higher than the estimated FDR. In the unbalanced case, five of the nine methods were unable to control the FDR in at least one dataset. For the heavily unbalanced cases, none of the methods were able to control the FDR in any of the datasets. Regardless, TMM, DESeq2Norm and CSS had a less biased true FDR than the other methods. In particular, the true FDR of TMM was close to 0.10 in all three datasets. Conversely, RCSS, quantile-quantile, upper quartile, total count and rarefying resulted in unacceptably high FDRs (close to or above 50%) in at least one dataset.

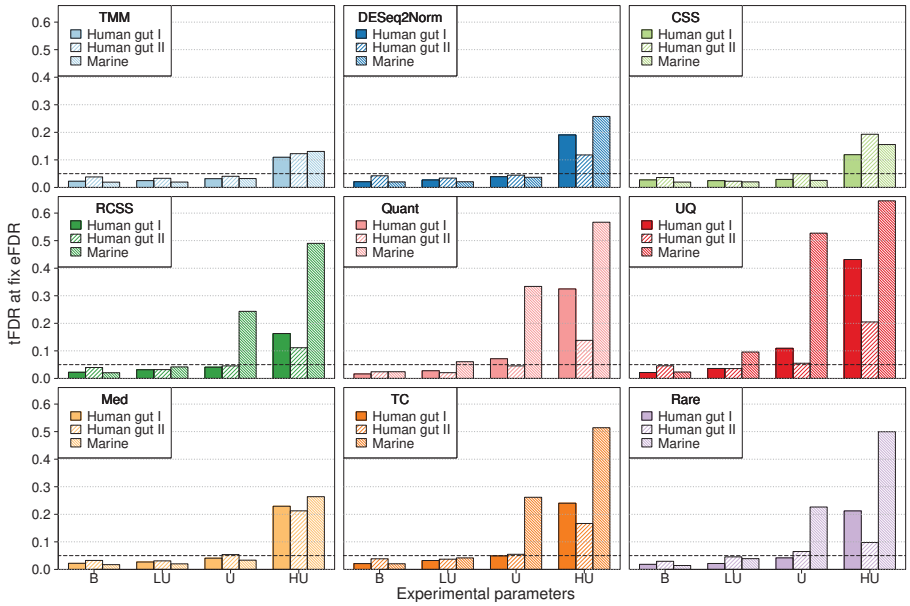

Figure 4.9: True false discovery rate at an estimated false discovery rate of 0.05 (y-axis) for different distributions of effects between groups (x-axis): balanced ('B') with 10% of effects divided equally between the two groups, lightly unbalanced ('LU') with effects added 75%-25% in each group, unbalanced ('U') with all effects added to only one group, and heavily unbalanced ('HU') with 20% of effects added to only one group. The results were based on re-sampled data consisting of two groups with 10 samples in each and an average fold-change of 3. Three metagenomic datasets were used: Human gut I, Human gut II and Marine. The following methods are included in the figure trimmed mean of M-values (TMM), DESeq2Norm, cumulative sum scaling (CSS), reversed cumulative sum scaling (RCSS), quantile-quantile (Quant), upper quartile (UQ), median (Med), total count (TC) and rarefying (Rare).

In conclusion, our observations confirm that the performance of the normalisation method depends on the characteristics of the data being analysed. In particular, all methods performed better for larger group sizes and larger fold-changes. More importantly, when there is an unbalanced distribution of effects, i.e. when the differentially abundant genes are all present in one condition, removal of systematic variability is more challenging for all studied methods. However, TMM and DESeq2Norm generally have a more robust performance and are therefore recommended methods.

5 Conclusions

In this thesis, statistical modelling was used to analyse metagenomic data. Here, we examined both sequence and count data from the DNA of bacterial communities from diverse environments. The studied models aimed to extract information from sequence data, to correct systematic variability of count data, or to detect differences in count data from different environments. In all these cases, the adequate statistical description of variability in the data was crucial for achieving a better understanding of the phenomena of interest.

Specifically, when dealing with the extraction of information from sequence data, Paper **I** showed that probabilistic models are efficient at detecting weak signals, such as the one from *attC* sites. By addressing the variability in both length and nucleotide composition of the sites, the model achieved a good sensitivity and specificity and simultaneously exhibited a high computational performance. In Paper **II**, the model was implemented in a computational tool that was in turn used to detect several thousands of horizontally transferred genes in bacterial communities. These genes had their biological functions characterised by comparison to databases of known genes or gene domains. The vast majority of the integron-mediated genes found in the environment had not been previously described. Among the integron-mediated genes with a known function, most were associated with gene mobility or defence mechanisms. In addition, potential novel antibiotic resistance genes might have been identified. In this way, Paper **II** substantially expands the knowledge about integron-mediated genes, which may lead to further insights regarding the evolution of bacterial genomes and how they can adapt to selective pressure.

The statistical modelling and analysis of metagenomic gene abundance presented in this thesis are challenging due to the discrete nature, high variability and high dimensionality of the data. Paper **III** provides a conceptual overview of the analyses involved in the study of comparative metagenomics, highlighting the computational and statistical challenges of the field. In particular, the removal of systematic variability using normalisation techniques appeared to depend on the characteristics of the data and had not previously been investigated for use in metagenomic gene abundance data. This led us to Paper **IV**, where we methodically compared the statistical performance of nine commonly

used normalisation methods. For this purpose, we developed an evaluation framework that preserved the structure of the data. This framework was based on the re-sampling of data, where two different conditions were simulated by splitting the re-sample into two groups and down-sampling a selected number of genes in one of the two groups. The framework is as general as possible and is recommended to be used for studying the performance of different types of statistical models that address the same or other types of high-throughput data. The study concluded that the characteristics of the data did affect the performance of the normalisation. Nevertheless, methods that more efficiently used the information present in the data performed consistently better. Importantly, the choice of normalisation method needs attention since using a method that cannot properly reduce the systematic variability of the data can lead to a large number of false positives in the detection of differently abundant genes. This failure of reducing the variability may ultimately lead to incorrect biological conclusions regarding the differences between the studied environments. Our study can therefore be used as guidance for selecting a suitable method for the normalisation of metagenomic gene abundance data.

Overall, this thesis highlights the importance of statistical modelling in the study of metagenomic data. The statistical approach is particularly relevant for addressing the high variability and high dimensionality typically present in these types of data. Indeed, these characteristics make rigorous statistical analysis necessary to ensure a high performance while avoiding the generation of excessively many false positives. Furthermore, since less than 2% of bacterial species can be cultivated in the laboratory, it is essential to continue studying bacteria via metagenomics, which allows for the analysis of collective genomes without the need for cultivation. In this way, it can be ensured that we continuously learn about their microscopic world, and statistical techniques, such as the ones presented in this thesis, are fundamental to further improving the understanding of the data generated within the field.

Bibliography

Amann, R. I., Ludwig, W., and Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews*, 59(1):143–169.

Axelson-Fisk, M. (2015). *Comparative Gene Finding: Models, Algorithms and Implementation*. Springer, 2nd edition.

Bengtsson-Palme, J., Boulund, F., Fick, J., Kristiansson, E., and Joakim Larsson, D. G. (2014). Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. *Frontiers in Microbiology*, 5((12)):1–14.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):298–300.

Boucher, Y., Labbate, M., Koenig, J. E., and Stokes, H. W. (2007). Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends in Microbiology*, 15(7):301–309.

Boulund, F., Johnning, A., Pereira, M. B., Larsson, D. G. J., and Kristiansson, E. (2012). A novel method to discover fluoroquinolone antibiotic resistance (qnr) genes in fragmented nucleotide sequences. *BMC genomics*, 13:695.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11:94.

Burset, M. and Guigó, R. (1996). Evaluation of gene structure prediction programs. *Genomics*, 34(3):353–367.

Cambray, G., Guerout, A.-M., and Mazel, D. (2010). Integrons. *Annual Review of Genetics*, 44(1):141–166.

Collis, C. M. and Hall, R. M. (1992). Gene cassettes from the insert region of integrons are excised as covalently closed circles. *Molecular Microbiology*, 6(19):2875–2885.

Cury, J., Jove, T., Touchon, M., Neron, B., and Rocha, E. P. C. (2016). Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Research*, 44(10):4539–4550.

Dudoit, S. and van der Laan, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer Series in Statistics. Springer-Verlag New York, New York, 1st edition.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. J. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge Univ. Press, Cambridge.

Escobar-Zepeda, A., De Leon, A. V. P., and Sanchez-Flores, A. (2015). The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in Genetics*, 6:348.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2015). The Pfam protein families database: towards a more sustainable future. *Nucleic acids research*, 44(D1):D279–D285.

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2015). Expanded Microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, 43(D1):D261–D269.

Gillings, M. R. (2014). Integrons: past, present, and future. *Microbiology and molecular biology reviews : MMBR*, 78(2):257–77.

Glud, R. N., Wenzhöfer, F., Middelboe, M., Oguri, K., Turnewitsch, R., Canfield, D. E., and Kitazato, H. (2013). High rates of microbial carbon turnover in sediments in the deepest oceanic trench on Earth. *Nature Geoscience*, 6(4):284–288.

Gribaldo, S., Poole, A. M., Daubin, V., Forterre, P., and Brochier-Armanet, C. (2010). The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nature Reviews Microbiology*, 8(10):743–752.

Haft, D. H., Selengut, J. D., and White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Research*, 31(1):371–373.

Hall, R. M., Brookes, D. E., and Stokes, H. W. (1991). Sitespecific insertion of genes into integrons: role of the 59base element and determination of the recombination crossover point. *Molecular Microbiology*, 5(8):1941–1959.

Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107:1–8.

Holmes, A. J., Gillings, M. R., Nield, B. S., Mabbutt, B. C., Nevalainen, K. M. H., and Stokes, H. W. (2003). The gene cassette metagenome is a basic resource for bacterial genome evolution. *Environmental Microbiology*, 5(5):383–394.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., and Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, 1(5):16048.

Hugenholtz, P., Goebel, B. M., and Pace, N. R. (1998). Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. *Journal of Bacteriology*, 180(18):4765–4774.

Hugerth, L. W., Larsson, J., Alneberg, J., Lindh, M. V., Legrand, C., Pinhassi, J., and Andersson, A. F. (2015). Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biology*, 16:279.

Hyatt, D., Locascio, P. F., Hauser, L. J., and Uberbacher, E. C. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, 28(17):2223–2230.

Jonsson, V., Österlund, T., Nerman, O., and Kristiansson, E. (2016). Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC Genomics*, 17:78.

Kristiansson, E., Fick, J., Janzon, A., Grabic, R., Rutgersson, C., So, H., and Larsson, D. G. J. (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*, 25(20):2737–2738.

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

Levinson, S. (1986). Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech & Language*, 1(1):29–45.

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):550.

MacDonald, D., Demarre, G., Bouvier, M., Mazel, D., and Gopaul, D. N. (2006). Structural basis for broad DNA-specificity in integron recombination. *Nature*, 440(7088):1157–1162.

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

Mason, O. U., Scott, N. M., Gonzalez, A., Robbins-Pianka, A., Bælum, J., Kimbrel, J., Bouskill, N. J., Prestat, E., Borglin, S., Joyner, D. C., Fortney, J. L., Jurelevicius, D., Stringfellow, W. T., Alvarez-Cohen, L., Hazen, T. C., Knight, R., Gilbert, J. a., and Jansson, J. K. (2014). Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. *The ISME journal*, 8(7):1464–75.

Mazel, D. (2006). Integrons: agents of bacterial evolution. *Nature Reviews Microbiology*, 4(8):608–620.

McMurdie, P. J. and Holmes, S. (2013). Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE*, 8(4):e61217.

Mitra, S., Klar, B., and Huson, D. H. (2009). Visual and statistical comparison of metagenomes. *Bioinformatics*, 25(15):1849–1855.

Nawrocki, E. P. and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935.

Nayfach, S. and Pollard, K. S. (2015). Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biology*, 16:51.

Oshlack, A. and Wakefield, M. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology direct*, 4:14.

Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E., and Larsson, D. G. J. (2014). BacMet: antibacterial biocide and metal resistance genes database. *Nucleic Acids Research*, 42(D1):D737–D743.

Parks, D. H., Tyson, G. W., Hugenholtz, P., and Beiko, R. G. (2014). STAMP: Statistical analysis of taxonomic and functional profiles. *Bioinformatics*, 30(21):3123–3124.

Partridge, S. R., Tsafnat, G., Coiera, E., and Iredell, J. R. (2009). Gene cassettes and cassette arrays in mobile resistance integrons: Review article. *FEMS Microbiology Reviews*, 33(4):757–784.

Paszkiewicz, K. and Studholme, D. J. (2010). De novo assembly of short sequence reads. *Briefings in Bioinformatics*, 11(5):457–472.

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.

Pereira, M. B., Wallroth, M., Kristiansson, E., and Axelson-Fisk, M. (2016). HattCI: Fast and Accurate attC site Identification Using Hidden Markov Models. *Journal of Computational Biology*, 23(11):891–902.

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., Yang, H., Wang, J., Ehrlich, S. D., Nielsen, R., Pedersen, O., and Kristiansen, K. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, pages 257–286.

Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y., and Blankenship, R. E. (2002). Whole-Genome Analysis of Photosynthetic Prokaryotes. *Science*, 298:1616–1619.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W.-T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., Rubin, E. M., Hugenholtz, P., and Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–437.

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

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, 11:R25.

Russell, M. and Moore, R. (1985). Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition. *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, 10(1):5–8.

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

Sohn, M. B., Du, R., and An, L. (2015). A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics*, 31(14):2269–2275.

Solden, L., Lloyd, K., and Wrighton, K. (2016). The bright side of microbial dark matter: Lessons learned from the uncultivated majority. *Current Opinion in Microbiology*, 31:217–226.

Stokes, H. W. and Hall, R. M. (1989). A novel family of potentially mobile DNA elements encoding sitespecific geneintegration functions: integrons. *Molecular Microbiology*, 3(12):1669–1683.

Stokes, H. W., O'Gorman, D. B., Recchia, G. D., Parsekhian, M., and Hall, R. M. (1997). Structure and function of 59-base element recombination sites associated with mobile gene cassettes. *Molecular microbiology*, 26(4):731–745.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., D'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B. T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M. B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S. G., Bork, P., Boss, E., Bowler, C., Follows, M., Karp-Boss, L., Krzic, U., Reynaud, E. G., Sardet, C., Sieracki, M., and Velayoudon, D. (2015). Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359.

Swithers, K. S., Soucy, S. M., and Gogarten, J. P. (2012). The role of reticulate evolution in creating innovation and complexity. *International journal of evolutionary biology*, 2012:418964.

Von Wintersdorff, C. J. H., Penders, J., Van Niekerk, J. M., Mills, N. D., Majumder, S., Van Alphen, L. B., Savelkoul, P. H. M., and Wolffs, P. F. G. (2016). Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Frontiers in Microbiology*, 7:173.

White, J. R., Nagarajan, N., and Pop, M. (2009). Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLoS Computational Biology*, 5(4):e1000352.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579.

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

Xiong, J., Xiong, J., Bauer, C. E., and Bauer, C. E. (2002). Complex evolution of photosynthesis. *Annual Review of Plant Biology*, 53:503–21.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., Knight, R., and Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., and Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, 67(11):2640–2644.