THESIS FOR THE DEGREE OF DOCTOR OF ENGINEERING

New antibiotic resistance genes and their diversity

Fanny Berglund

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden 2019

New antibiotic resistance genes and their diversity
Fanny Berglund
Göteborg 2019
ISBN 978-91-7905-145-7

Doktorshavhandlingar vid Chalmers tekniska högskola
Ny serie nr 4612
ISSN 0346-718X

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

Typeset with LaTeX
Printed by Chalmers Reproservice, Göteborg, Sweden 2019

New antibiotic resistance genes and their diversity

Fanny Berglund

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

Abstract

Antibiotic resistance is increasing worldwide and is considered a severe threat to public health. Often, antibiotic resistance is caused by antibiotic resistance genes, of which many are hypothesized to have been transferred into human pathogens from environmental bacteria. It is, therefore, of great importance to explore bacterial communities to identify new antibiotic resistance genes before they reach clinical settings. The six papers presented in this thesis aim to identify new antibiotic resistance genes in large genomic and metagenomic datasets and to place them in an evolutionary context. In Paper I, a new method for the identification and reconstruction of new antibiotic resistance genes directly from fragmented metagenomic data was developed and was shown to outperform other methods significantly. In Papers II and III, novel genes of the clinically important class metallo-β-lactamases were identified. By analyzing metagenomes and bacterial genomes, 96 novel putative metallo-β-lactamase genes were predicted. In Paper IV, the diversity and phylogeny of the metallo-β-lactamases were further investigated. The results showed that the genes mainly clustered based on the taxonomy of the host species and that many of the mobile metallo-β-lactamases potentially were mobilized from species of the phylum Proteobacteria. In Paper V, the aim was to identify new genes providing resistance to the antibiotic class tetracyclines. A total of 195 gene families were predicted, of which 164 were new putative tetracycline resistance genes. Finally, in Paper VI, we searched for and predicted 20 novel putative quinolone resistance (*qnr*) genes from a large amount of metagenomic data. Throughout the thesis, a total of 54 novel genes have been functionally verified in *Escherichia coli*, of which 37 expressed the predicted phenotype. The results of this thesis provide deeper insights into the diversity and evolutionary history of three major classes of antibiotic resistance genes. It also provides new methodologies for efficient and reliable identification of new resistance genes in genomic and metagenomic data.

Keywords: antibiotic resistance, metagenomics, big data, β-lactamases, carbapenemases, tetracycline resistance, hidden Markov model.

List of papers

The thesis includes the following papers.

I. **Berglund, F.**, Österlund, T., Boulund, F., Marathe, N., Larsson, D.G.J., Kristiansson, E. (2019). Identification and reconstruction of novel antibiotic resistance genes from metagenomes. *Microbiome.* 7(1)

II. **Berglund, F.**, Marathe, N., Österlund, T., Bengtsson-Palme, J., Kotsakis, S., Flach, C.F., Larsson, D.G.J., Kristiansson, E. (2017). Identification of 76 novel metallo-β-lactamases through large-scale screening of genomic and metagenomic data. *Microbiome.* 5(1).

III. Marathe, N., **Berglund, F.**, Razavi, M., Pal, C., Dröge, J., Samant, S., Kristiansson, E., Larsson, D.G.J (2019). Sewage effluent from an Indian hospital harbors novel carbapenemases and integron-borne antibiotic resistance genes. *Microbiome.* 5(1)

IV. **Berglund, F.**, Johnning, A., Larsson, D.G.J., Kristiansson, E. (2019). An updated phylogeny of the metallo-β-lactamases. *Manuscript.*

V. **Berglund, F.**, Böhm, M.E., Martinsson, A., Ebmeyer, S., Österlund, T., Johnning, A., Larsson, D.G.J., Kristiansson, E. (2017). Comprehensive screening of genomic and metagenomic data reveals a large diversity of tetracycline resistance genes. *Manuscript.*

VI. Boulund, F., **Berglund, F.**, Flach, C.F., Bengtsson-Palme, J., Marathe, N., Larsson, D.G.J., Kristiansson, E. (2017). Computational discovery and functional validation of novel fluoroquinolone resistance genes in public metagenomic data sets. *BMC Genomics.* 18(1)

Publications not included in this thesis:

- Marathe, M., Janzon, A., Kotsakis, S., Flach, C.F., Razavi, M., **Berglund, F.**, Kristiansson, E., Larsson, D.G.J. (2018). Functional metagenomics reveals a novel carbapenem-hydrolyzing mobile beta-lactamase from Indian river sediments contaminated with antibiotic production waste. *Environ Int.* 112
- Gustavsson, K., **Berglund, F.**, Jonsson, P.R., Mehlig, B. (2016). Preferential Sampling and Small-Scale Clustering of Gyrotactic Microswimmers in Turbulence. *Physical review letters.* 116(10)

Author contributions

I. Participated in study design, developed and implemented the method, created and optimized the models, designed and executed the performance comparison against competing methods, wrote the online documentation, performed the data analysis, analyzed the results, drafted and edited the manuscript.

II. Participated in study design, collected the data for and created the model, collected the genomic and metagenomic data, developed and implemented the analysis pipeline, performed the data analysis, performed the clustering, created the phylogenetic tree, analyzed the results, drafted and edited the manuscript.

III. Performed bioinformatical analysis to identify metallo-β-lactamases, performed post-processing and annotated the gene candidates, created the phylogenetic trees, wrote the section describing the identification of new metallo-β-lactamases, and edited the manuscript.

IV. Participated in study design, selected and collected the genomic and metagenomic data, performed the data analysis, created the phylogenetic trees, performed the phylum analysis, analyzed the binding sites, analyzed the results, drafted and edited the manuscript.

V. Participated in study design, selected and collected the genomic and metagenomic data, created and optimized the models, performed the data analysis, created the phylogenetic trees, performed the phylum analysis, analyzed the results, drafted and edited the manuscript.

VI. Participated in study design, collected the data, implemented the analysis pipeline, performed the analysis of metagenomic data, performed the clustering, and edited the manuscript.

Acknowledgements

First and foremost I want to thank Erik Kristiansson. I couldn't have wished for a better supervisor and I am so grateful for your constant support. Your ability to see solutions everywhere, your constructive feedback and your enthusiasm have guided me through these five years and I have learned a lot from you, thank you! Next, I want to thank my co-supervisors Tobias Österlund, Joakim Larsson and Anna Johnning whom all have played important roles during my time as a Ph.D. student. Thank you Tobias for your excellent guidance during my first years, you were always there when I needed someone to bounce ideas with. Thank you Joakim for always giving good feedback, advice and support. Thank you Anna for your eye for details, your good ideas and all your text related feedback, it meant a lot. My thanks also goes to my examiners Olle Nerman and Marija Cvijovic.

I want to thank everyone I have worked with at Sahlgrenska for very good collaborations and all your valuable input on the many biological components of the work in this thesis, Nachiket Marathe, Marlies Böhm, Stefan Ebmeyer, Carl-Fredrik Flach, Mohammad Razavi and Johan Bengtsson-Palme. Thank you Fredrik Boulund for our nice collaboration with the research, and for sharing your knowledge about new things, both bioinformatical tools and good computer games.

One of the main reasons why I so much have enjoyed my years as a Ph.D. student is the both current and former members of Erik Kristiansson's research group: Viktor Jonsson, Mariana Pereira, Johannes Dröge, Anna Rehammar and Mikael Gutavsson. I am very lucky to have been supported by such amazing people.

Thank you Lotta Fernström and Marie Kühn for your efficient and friendly support with everything related to administration.

I also want to thank all my other colleagues at the department, especially Sandra, Olle E, Malin, Jonatan K, Jonatan N, Ivar, Claes, Juan, Marina, Johannes, Hossein, Oskar, and Felix. The lunch and fika breaks would not have been the same without you. I would like to thank my friends outside of the department for being so understanding when I have been extra stressed and have gone into "social hibernation", and for always being there when I have returned.

Till min familj: ni är bäst. Speciellt tack till farmor för alla räknestunder, stöd och uppmuntran.

Contents

1 Background

Antibiotics are substances that can kill or inhibit the growth of bacteria. Since their discovery in the early 20th century, they have saved millions of people from life-threatening bacterial infections and have facilitated major improvements in medicine and surgery. An antibiotic can either be a natural product produced by specific microorganisms, synthetically constructed, or a combination thereof. Over the past 60 years, most antibiotics have resulted from natural production by bacteria or fungi, with a few exceptions of synthetically constructed antibiotics (Walsh, 2003). There are several classes of antibiotics, but the majority of them were discovered between 1940 and 1962, and since then, only two new classes have been introduced (Coates et al., 2011; Tacconelli et al., 2018).

Antibiotic resistance is the ability of bacteria to withstand the effects of antibiotics, and consequently, medicines previously used to treat infections become ineffective. Bacteria can withstand antibiotics through different resistance mechanisms, which can be classified as intrinsic, adaptive or acquired. Bacteria with intrinsic resistance have some inherited characteristics that cause them to be unaffected by antibiotics, and this is a feature of all members of the species. This characteristic does not change over time; an example is the outer membrane of many Gram-negative bacteria, which is impermeable to certain types of antibiotics (Carlos, 2015). The second variant of resistance is adaptive resistance or stress response, where sudden changes in the environment of a bacterial community trigger defense systems that can limit the permeability of the outer membrane or overexpress efflux pumps, which leads to a decreased accumulation of antibiotics and thus a decreased susceptibility (Carlos, 2015). Finally, bacteria can also acquire resistance via mutations and horizontal gene transfer of resistance genes. Through mobile genetic elements such as plasmids and transposons, these resistance genes can move between bacterial cells and species (Stokes and Gillings, 2011), and once obtained, the acquired resistance can be inherited by daughter cells of the bacteria. Horizontal gene transfer has played a vital role in the evolution of prokaryotes and can lead to fast changes in

the genetic compositions of these organisms. Furthermore, if there is a positive selection for the acquired resistance genes then the probability for the bacteria to keep the new genetic elements will increase, as will the probability for the spread to a larger community (Thomas and Nielsen, 2005).

Among the first antibiotics used commercially was penicillin in the 1940s, but almost immediately after its release, a bacteria-produced enzyme with the ability to hydrolyze penicillin was discovered. (Abraham and Chain, 1940). New types of antibiotics were developed in response to increasing resistance, but soon after a new drug was on the market, a new type of resistance was discovered. Today, resistance to almost every developed antibiotic has been recognized (Ventola, 2015), estimated to result in 700 000 deaths every year and threatening our ability to perform key medical procedures (O'Neill et al., 2016). Over time, antibiotic-resistant bacteria have undergone development from resistance to single classes of antibiotics to being multi-drug resistant, often through horizontal gene transfer. In clinical settings, some of the most problematic resistant bacteria are the extended-spectrum β-lactamase (ESBL)-positive Enterobacteriaceae, while other common types are the Methicillin-resistant *Staphylococcus aureus* (MRSA) and the vancomycin-resistant enterococci (VRE) (Cantas et al., 2013).

As a last-resort treatment for a patient infected with a multiresistant bacteria, the β-lactam antibiotics carbapenems are often used. However, recent studies have shown that resistance to this antibiotic is emerging worldwide. The resistance to carbapenems is to a large extent due to the expression of carbapenemases, a class of enzymes where the majority have broad-spectrum substrate profiles (Papp-Wallace et al., 2011). An especially worrisome carbapenemase is the acquired NDM-1, which was first discovered in 2009 (Yong et al., 2009). In only a few years, the encoding gene had rapidly spread through pathogens and has now been found in several geographical locations and isolated in numerous bacterial species (Walsh et al., 2011), showing how fast new forms of resistance genes can become a serious problem.

1.1 Antibiotic resistance and the environment

Antibiotics have been naturally produced by environmental microbial communities since ancient times, with some estimates pointing back to two billion years ago (Hall and Barlow, 2004). As a result of evolution, antibiotic resistance has likely been around for a similarly long time (D'Costa et al., 2011). This has given the microbes ample time to develop a large and diverse set of resistance

genes, a resistome, with most of them still undiscovered.

It is clear that the increase in antibiotic usage in human and veterinary settings, together with the greater movement of people and animals, has led to the increased prevalence and spread of antibiotic-resistant bacteria (Martínez, 2008; Cantas et al., 2013). Although some human pathogens are intrinsically resistant to certain antibiotics, many of the pathogens are not originally carriers of antibiotic resistance genes (ARGs). Instead, many of the ARGs encountered in clinical settings are hypothesized to have originated from the environment (Walsh, 2013). Resistance genes similar and identical to those in pathogenic bacteria have been discovered in various environmental communities (Riesenfeld et al., 2004a; Canton, 2009; Boulund et al., 2012), including pristine environments such as glaciers (Segawa et al., 2013) and 30 000-year-old permafrost samples (D'Costa et al., 2011). However, the exact origin of most of these genes is unknown. Furthermore, environmental and commensal bacterial communities have been shown to harbor a vast diversity of ARGs, among which many have not been identified in clinical settings (Sommer et al., 2009; Forsberg et al., 2012; Wichmann et al., 2014). It has further been shown that selection pressures, such as antibiotic exposure, may enrich the abundance and diversity of resistance genetic elements in these communities (Gillings and Stokes, 2012). It is therefore likely that the environment has and will continue to act as a reservoir for ARGs that can be spread to pathogenic bacteria (Allen et al., 2010).

1.2 DNA sequencing

All known living organisms depend on their genetic material to function and grow. Genetic information is encoded in two forms of nucleic acids, where deoxyribonucleic acid (DNA) acts as repositories and ribonucleic acid (RNA) acts as transmitters. DNA molecules generally contain four types of nucleotides, namely, adenine (A), thymine (T), guanine (G) and cytosine (C), and RNA has the same nucleotides except that T is replaced with uracil (U). Depending on how these nucleotides are organized, they determine an organism's functions and traits. In each cell in every organism, a copy of the complete set of genetic information is stored, which is called the genome. The size of genomes is highly variable, with the genomes of the smallest viruses having a few thousand bases compared to the more than three billion bases of the human genome and even more in some species. Most genomes consist of a large number of genes, which are parts of the genome that contain information about the synthesis of RNA and proteins (Mathews et al., 2000).

In the early 1950s, when it was discovered that proteins are constructed by amino acids in what appeared to be defined orders (Sanger, 1960), the search for the blueprint for protein creation began. Although the existence of DNA was proposed shortly thereafter (Watson et al., 1953), it was not until 1965 that the first nucleic acid molecule in the form of tRNA was sequenced (Holley et al., 1965). The first DNA sequenced was from the bacteriophage λ, which was completed in 1971 and consisted of 12 bases (Wu and Taylor, 1971), and the first whole genome sequenced was that of another bacteriophage, ϕX with 5375 bases, in 1978. The length of the sequences continued to increase with the first human chromosome sequenced in 1999 (Dunham et al., 1999), and in 2001, the first draft versions of the human genome's 3.2 billion bases were published (Venter et al., 2001; Lander et al., 2001). The advancements in the sequencing techniques had until then been achieved mainly with the Sanger method, also referred to as "first-generation" sequencing, but over the past decade, several new sequencing techniques have emerged. A large breakthrough was the development of the high-throughput second-generation or "next-generation" sequencing (NGS) techniques. The second-generation sequencing provided the ability to produce millions of sequence reads in parallel, speeding up the process and substantially reducing the cost of each read. Although the sequence reads produced with second-generation techniques are typically not as long and not as accurate as those produced with Sanger technology, the great throughput of reads makes the coverage of each DNA fragment high and, therefore, the final sequences more accurate. Recently, next-generation sequencing technologies have been further developed with the aim of producing even longer reads while maintaining the low price per read achieved with the second-generation techniques (van Dijk et al., 2014). The rapid development of sequencing techniques has created an enormous amount of sequence data, and today, there are approximately 200 000 sequenced bacterial genomes that are publicly available (NCBI, 2019), compared to only 300 sequenced bacterial genomes a little more than a decade ago (Land et al., 2015).

The massive amount of genomic data has provided us with unprecedented opportunities to study all life forms on earth. The applications are numerous, ranging from evolutionary studies to disease prevention. With even better and faster sequencing techniques, together with new algorithms to process the data, the number of applications will most likely be endless.

1.3 Metagenomics

It is estimated that there are over 10^{30} individual bacterial cells (Whitman et al., 1998) and between 10^7 and 10^9 bacterial species on earth (Curtis et al., 2002; Dykhuizen, 1998). Historically, the study of microorganisms has been focused on culturing of single species, but the vast majority of all microorganisms cannot be cultured with standard techniques (Hugenholtz et al., 1998). The uncultivable microorganisms represent many diverse organisms living in several unique communities, and they are often distantly related to the cultivable ones (Riesenfeld et al., 2004b). It is therefore important to develop culture-independent methods to understand the ecological role of these communities, their genetic diversity and population structure. To address this challenge, a technique that is now called *functional metagenomics* was developed (Handelsman et al., 1998). In this technique, DNA is taken directly from environmental communities rather than from a single organism and fragmented into pieces. The DNA fragments are then inserted into cultivable bacteria that are grown under certain conditions, such as under antibiotic exposure, to assess the community for specific genes. The collective genome of all organisms present in a community is named the *metagenome*.

However, functional metagenomics has some drawbacks; the procedure is time-consuming, and not every gene can be expressed in a cultivable host. Furthermore, when searching for a specific type of gene, many fragments must be created to ensure that the complete gene is captured by the inserted fragments. However, with cheaper and faster sequencing techniques, the sequencing of uncultured microorganisms directly from their community became possible (Heather and Chain, 2016). The technique, often called *shotgun metagenomics*, is the direct isolation and sequencing of genomic DNA from a bacterial community through high-throughput DNA sequencing (Wooley et al., 2010). However, the many benefits of metagenomics are accompanied by some complications. Compared to single-organism sequencing where the whole genome of a specific species is sequenced, metagenomic data often contain DNA from a large number of species, and due to the high diversity generally present in and in-between microbial communities, metagenomes are often undersampled. Furthermore, the data are usually highly fragmented, with fragments occasionally as short as 75 bases, depending on which sequencing technique is used. The result is a dataset that is difficult to reconstruct through sequence assembly, and it is often not possible to reconstruct the complete genomes of the organisms present in the sampled community. The analysis of metagenomic data is therefore challenging and requires specifically designed methods and algorithms (Quince et al., 2017).

Over the past years, several large projects have been conducted within the metagenomic area, including the study of the gut microbiomes of 124 European individuals, which resulted in the first human gut gene catalog (Qin et al., 2010). Then, the Human Microbiome Project was presented (Consortium et al., 2012), where scientists discovered, among others, that the microbial communities between body sites of healthy individuals remarkably differed. In the Tara Ocean project, 7.2 trillion bases of metagenomic data have been sequenced, leading to a new picture of the diversity within the ocean with more than 40 million nonredundant genes discovered (Armbrust and Palumbi, 2015).

Historically, studies of antibiotic resistance have been conducted in clinical settings with pathogenic bacteria. However, to elucidate the actual diversity and abundance of resistance genes, it is necessary to focus investigations on the environmental and commensal bacterial communities (Bengtsson-Palme and Larsson, 2015; Berendonk et al., 2015). Because the majority of bacteria inhabiting these communities are uncultivable under standard laboratory conditions, the development of metagenomics has provided an unprecedented opportunity to explore these environments under unbiased conditions. Metagenomics offers the means to investigate whole bacterial communities for many resistance genes in parallel, leading to estimates of the present resistome and the possibility to compare the resistomes between different environments (Bengtsson-Palme et al., 2017b). Indeed, in recent years, several studies have focused on quantifying resistance genes in environments such as sewage treatment plants (Yang et al., 2013, 2014), pharmaceutical-polluted environments (Kristiansson et al., 2011; Bengtsson-Palme et al., 2014) and the human gut (Forslund et al., 2013). Recently, this approach has also been used as a way to monitor and compare the trends of antibiotic resistance at both local and global scales (Hendriksen et al., 2019; Huijbers et al., 2019). Furthermore, metagenomics makes it possible to screen large amounts of data for new ARGs. The identification of previously uncharacterized genes can help elucidate the evolutionary origin and history of the ARGs that are presently encountered in clinical settings. Therefore, analyzing and investigating the environmental resistomes, where in some cases the majority of the ARGs are expected to be new, is crucial for better understanding the selection pressures and dissemination routes that facilitate the mobilization and spread of ARGs (Larsson et al., 2018).

2 Aims

All papers included in this thesis are motivated by the overall aim to extend the knowledge of the unknown resistome through identifying novel antibiotic resistance genes in DNA sequence data. Paper **I** is dedicated to the development and assessment of a method with the ability to identify previously unknown antibiotic resistance genes from fragmented metagenomic data. In Papers **I-VI**, the method is applied to identify antibiotic resistance genes of different classes to further elucidate their diversity and abundance. To trace events of horizontal gene transfer and evolutionary origin, the identified genes are placed into an evolutionary context by comparing them to previously known genes (Papers **II**, **IV** and **V**). More specifically, the aims of this thesis are to

1. develop and evaluate methods for the identification of novel antibiotic resistance genes in genomes and metagenomes (Papers **I-II**).
2. predict novel β-lactamases, with a particular focus on carbapenemases, investigate their diversity and experimentally validate their function (Papers **I-IV**).
3. predict novel tetracycline resistance genes; investigate their diversity; and experimentally validate their function (Paper **V**).
4. predict novel quinolone resistance (*qnr*) genes, investigate their diversity and validate their function (Paper **VI**).
5. investigate the evolutionary aspects of antibiotic resistance genes using phylogenies inferred from both new and previously known genes (Papers **II**, **IV-V**).

3 Identification and analysis of antibiotic resistance genes

Most of the work in this thesis is related to the identification and analysis of antibiotic resistance genes (ARGs) from sequencing data. A wide range of approaches has been developed for this purpose, and the choice depends on several factors, such as the objectives of the study, type of data and available computational power. This chapter provides a methodological background to the area of identifying known and novel ARGs, from the DNA sequencing to the final analysis of the predicted genes. Figure 3.1 presents an overview of some of the major steps in the workflow when searching for ARGs.

3.1 Generating and assembling sequence data

The emergence of cost-effective, high-throughput next-generation sequencing (NGS) technologies in the mid-2000s drastically changed the bioinformatical landscape, with an explosion of large datasets as a result. Today, several commercially available DNA sequencing platforms are available, such as Illumina (Solexa), 454, Ion Torrent and SOLiD sequencing (Liu et al., 2012), and depending on the technique used, the properties of the generated sequence reads will vary. In addition to these methods, new techniques such as SMRT (PacBio) (Rhoads and Au, 2015) and Oxford Nanopore (Clarke et al., 2009) have emerged and are occasionally referred to as third-generation sequencing. What all these techniques have in common is that they do not require DNA amplification and are capable of sequencing single molecules, which is in contrast to most of the early NGS techniques. These new techniques are capable of producing significantly longer reads, up to and beyond hundreds of kilobases (Goodwin et al., 2016). Although the sequencing platforms that produce long sequence reads are continuously improving, the higher cost and error rate makes them

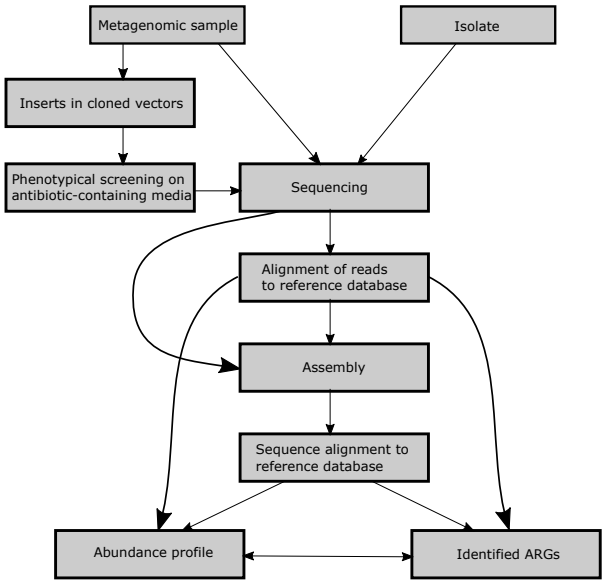

Figure 3.1: This flowchart shows possible workflows for the identification of ARGs. Isolate samples can be sequences straight away whereas metagenomic samples can be either sequenced directly or have the DNA inserted into cloned vectors that are incorporated into host bacteria which are grown on antibiotic-containing media, as a way to isolate the resistance genes. Once sequenced, the generated reads can either be subjected to a *de novo* assembly or aligned to a reference database. The aligned reads can be used to estimate the abundance of ARGs in the sample, and they can also proceed to an assembly. Assembled contigs can be aligned to reference databases to identify full-length ARGs.

inapplicable for some studies.

As a result of the short read lengths produced by many of the NGS techniques, there is often a need to assemble the short reads into longer sequences, i.e., reconstructing the, or parts of, sequenced genome. This process is called an assembly. The vast increase in short-read datasets has led to a high demand for efficient and accurate assembly algorithms. For whole-genome shotgun sequencing data, where a single organism is sequenced, the assembly algorithms have been continuously improved, and today, the main challenge with whole-genome assembly is accurate reconstruction in regions with repetitive DNA segments. Some of these issues have, however, been resolved using the new sequencing techniques with longer reads to gap the missing regions (Goodwin et al., 2016). Metagenomic data, however, consist of several different organisms,

and to obtain an accurate overview of the microbiome, there is a need to have a high sequencing depth, which leads to large datasets, often on the order of a billion reads. Here, the assembly is a highly intricate task, where, disregarding the sheer number of reads, the objective is to piece together the shuffled reads from several unknown organisms rather than assemble the genome of a single organism (Quince et al., 2017). Therefore, there are complications related to metagenomic assemblies that are not present for whole-genome assembly. For instance, several individuals from the same species could have small genetic variants, and entirely unrelated genomes may contain almost identical DNA segments, as is the case if they carry mobile genetic elements (Ayling et al., 2019).

Today, there are algorithms and software packages specifically developed to handle the complex metagenomic data and to achieve *de novo* assemblies (Wang et al., 2019), such as MetaVelvet (Namiki et al., 2012), Meta-IDBA (Peng et al., 2011), MEGAHIT (Li et al., 2015), Ray (Boisvert et al., 2010), IDBA-UD (Peng et al., 2012) and metaSPAdes (Nurk et al., 2017). However, these algorithms are computationally demanding and memory intensive, and for large datasets, these algorithms must be performed on computer clusters. Although it is possible to assemble smaller datasets on a reasonably large server, some datasets are not feasible to assemble locally due to memory, computational, and time requirements. There are different tactics to overcome this problem. One is to skip the *de novo* assembly and instead perform guided targeted assembly, where reads are mapped to available genes and then assembled gene-wise using tools such as MegaGTA (Li et al., 2017) and Xander (Wang et al., 2015). Many approaches for studying metagenomic data often circumvent the assembly altogether and perform the analysis, such as binning, mapping, or short read alignment, directly on the sequence reads. However, information such as in what organism the gene is located and the genetic context surrounding the gene, such as potential mobile elements, will then be missed. Therefore, the need for *de novo* assemblies is significant. Today, there is no perfect solution, and there is therefore no doubt that the demand for improved algorithms and software that perform metagenomic *de novo* assemblies is high and will continue to increase.

3.2 Methods for identifying resistance genes

There are many approaches that can be used for identifying resistance genes in sequence data, where the complexity ranges from an alignment of a protein sequence against a reference database to methods consisting of multiple steps and classification procedures. The most appropriate method will mainly depend on the objective of the study but also by the nature of the sequence data and

the available computer resources. However, all methods can be divided into two main categories: alignment-based methods and functional metagenomics. Alignment-based methods are, in contrast to functional metagenomics, based on sequence homology to known genes. Thus, if the study aims to identify genes that have no sequence homology to previously characterized genes, then functional metagenomics is the only available approach (see section 3.2.4).

Most of the alignment-based methods rely on sequence similarity searches against a reference database, producing scored sequence alignments. Several bioinformatics tools can perform this task, such as BLAST (Camacho et al., 2009), DIAMOND (Buchfink et al., 2015) and UBLAST (Edgar, 2010). Although sequence similarity searches against reference databases are often very useful in identifying previously known genes, these methods are less suitable for detecting distant homologs to known genes. For this purpose, methods based on profile hidden Markov models (HMMs) (Eddy, 2011), which describe the conserved regions and the variation between the genes, are better suited.

3.2.1 Databases

The performance when identifying ARGs heavily depends on how well-curated and comprehensive the database is. Today, there are several publicly available databases to choose from, and many of the databases display some overlap in content. The most extensively used databases that specialize in antibiotic resistance are **ResFinder** (Zankari et al., 2012), the Comprehensive Antibiotic Resistance Database (**CARD**, (Jia et al., 2016)), Antibiotic Resistance Gene-ANNOTation (**ARG-ANNOT**, (Gupta et al., 2014)), **Resfams** (Gibson et al., 2015), Antibiotic Resistance Genes Database (**ARDB**, (Liu and Pop, 2008)), **MEGARes** (Lakin et al., 2016), and Structured Antibiotic Resistance Genes (**SARG**, (Yin et al., 2018)). In addition, the NCBI nonredundant protein database (NCBI nr) contains all nonredundant protein sequences submitted to NCBI.

The ResFinder database has specialized in acquired antimicrobial resistance genes, is manually curated and also contains phenotypic information. In contrast to ResFinder, the CARD and ARG-ANNOT databases are not limited to acquired antimicrobial resistance genes but contain manually curated reference data on antimicrobial resistance, including both intrinsic and acquired genes and mutations involved in resistance. The MEGARes database contains manually curated sequences from ResFinder, ARG-ANNOT, CARD, and the Lahey clinic β-lactamase archive (Bush and Jacoby, 2010). MEGARes is also designed to handle large metagenomic datasets through automated searches by streamlining the annotation and structure of the database. The ARDB, focusing on all ARGs,

has not been updated since 2009 but is together with CARD included in the SARG (v.2) database, which also contains selected and curated sequences from NCBI nr. The Resfams database consists of sets of protein families confirmed for antibiotic resistance and associated profile HMMs.

3.2.2 Methods for analysis of whole genomes

For longer sequences such as whole genomes, plasmids and assembled contigs, there is generally no need to complicate the method for identifying resistance genes because the data are often manageable in size, and each segment contains much more information compared to a short sequence read. Instead, it is a question of what database and classification threshold to use. **ResFinder**, **Resfams**, **RGI** (The Resistance Gene Identifier) and **ARG-ANNOT** are four methods that use straightforward approaches for identifying resistance genes in genomes, plasmids and contigs, with the main difference being the databases and sequence alignment methods used (Table 3.1).

ResFinder is both a web-based and standalone method that aligns the input data using BLAST against the ResFinder database (Zankari et al., 2012). The classification is based on sequence similarity, and the best hits to acquired resistance genes are displayed together with the resistance phenotype. The web tool can take both whole genomes/contigs and fragmented data; however, the fragmented data are assembled prior to the analysis using a dedicated server such that the analysis is performed on the resulting contigs. Similar to ResFinder, **ARG-ANNOT** performs a similarity search using BLAST to identify resistance genes. Here, the analysis is performed through the tool BioEdit (Hall et al., 2011), which enables the user to create their own database and perform the search locally on a Windows computer, although the database ARG-ANNOT is provided (Gupta et al., 2014). **Resfams** is a set of curated profile HMMs, where gathering thresholds have been added to each of the profile HMMs (Gibson et al., 2015). The analysis is performed using the software HMMER (Eddy, 2011), and the specificity of the annotation will depend on how specific the HMMs used are. Although Resfams is based on HMMs, the models are not optimized to identify fragmented data; hence, the performance for short sequence reads as input is poor, as has been shown in Paper I. **RGI** is slightly more complex than ResFinder and Resfams, where the input data are first analyzed for open reading frames (ORFs), which are then used as the query sequences for a sequence similarity search against the CARD database using either BLAST or DIAMOND. Here, the user can choose from among three criteria for gene identification: Perfect, Strict and Loose. The Perfect option only searches for perfect matches to the database. The Strict option

can detect variants of previously known genes using curated similarity cutoffs, while the Loose option will display hits that are below the cutoffs to enable the identification of more distant homologs of previously known resistance genes (Jia et al., 2016).

Table 3.1: An overview of bioinformatic tools to identify resistance genes in sequence data.

	Query type	Classification method
ResFinder	Assembled	Alignment to ref. DB
Resfams	Assembled	Alignment to profile HMM
RGI	Assembled	Alignment to ref. DB
ARG-ANNOT	Assembled	Alignment to ref. DB
AmrPlusPlus	Reads	Alignment to ref. DB.
GROOT	Reads	Hierarchical local align. to variation graph
MEGAN	Reads	Alignment to ref. DB
ARIBA	Reads	Alignment to ref. DB
ARGs-OAP	Reads	Alignment to ref. DB
ARGs-OAP SARGfam	Open Reading Frames	Alignment to profile HMM
deepARG	Reads & Assembled	Alignment to ref. DB + Neural networks
fARGene	Reads & Assembled	Alignment to profile HMM

3.2.3 Methods for analysis of metagenomic reads

When the data of interest are fragmented metagenomic data, the analysis becomes more complicated, which can be addressed by several approaches. Because metagenomic data consist of fragments from several different organisms and suffer from a wide range of noise (Boulund et al., 2018), an assembly of the data requires extensive computer resources and is occasionally not even practically feasible. Furthermore, metagenomic data are often undersampled, and genes that are present in low abundance might be missed in the assembly due to low coverage, while regions containing mobile genetic elements, where many ARGs are located, are notoriously hard to assemble (Bengtsson-Palme et al., 2014; Wu et al., 2012; Ellington et al., 2017). Instead, there are several bioinformatic approaches developed to perform the analysis directly on the short reads (Table 3.1).

Among the tools that provide resistome profiles is **AmrPlusPlus**. This tool can analyze large metagenomic datasets and provides count files for each sample that contain information on what resistance genes are identified in the samples and its corresponding abundance (Grüning, 2016). The method operates on short sequence reads that are quality controlled and then aligned to the MEGARes database using the short-read mapping tool BWA (Li and Durbin, 2009). A

similar tool is **ARGs-OAP**, which takes fragmented metagenomic samples as input and offers ARG profiles for the data, including taxonomic profiles and abundance analysis. The method consists of a two-step analysis, wherein the short sequence reads in the first step are screened for 16S rRNA for taxonomic profiling and potential ARGs using UBLAST against the database SARG. The second step is once again a similarity search against the SARG database but this time using BLASTX. Here, the limiting factor for finding new ARGs is the first stage of the pipeline, where many potential ARG reads are discarded, which we also show in Paper I. The SARG database also includes sets of profile HMMs (SARGfam) that were created after phylogenetic analysis of the sequences in the database, and gathering thresholds were set using leave-one-out cross-validation. However, the specificity for the SARGfam is estimated from the analysis of randomly chosen sequences from other types of resistance genes, which do not necessarily have to be the sequences most likely to be classified as false positives. The SARGfam is furthermore only accessible online and requires predicted ORFs as input.

MEGAN is another tool used for the analysis of both the taxonomic and functional compositions of large microbial datasets (Huson et al., 2016). Although not specialized in the identification of resistance genes, this tool, to some extent, includes functionality to do so. In the approach, short sequence reads are mapped to the reference database NCBI nr using the DIAMOND aligner (Buchfink et al., 2015). Then, both taxonomic and functional profiling are performed based on NCBI taxonomy, SEED (Overbeek et al., 2013), eggNOG (Powell et al., 2011) and InterPro2GO mapping of reads (Hunter et al., 2013; Mitchell et al., 2014). A gene-centric assembly can then be performed, where all reads assigned to a given functional node can be assembled and given as output. However, because MEGAN does not offer any resistance gene database for binning, the performance for the identification of resistance genes is poor, mainly due to the lack of resolution of the resistance-related genes in the database.

The method **ARIBA** (Antimicrobial Resistance Identification By Assembly) (Hunt et al., 2017), similarly to MEGAN, takes fragmented metagenomic data as input and then uses a targeted local assembly approach to identify resistance genes. However, ARIBA has a more sophisticated classification scheme in which the short-read data are mapped to the reference sequences using minimap (Li, 2016). The reference sequences consist of any of the databases ARG-ANNOT, CARD, MEGARes or ResFinder that have been clustered by similarity before use. Then, the mapped reads are assembled separately for each cluster, and for each contig, the closest reference sequence is identified using the MUMmer sequence alignment package (Delcher et al., 2003). The reads for each cluster are mapped to the assembly using the read aligner Bowtie2 (Langmead and Salzberg, 2012), from which potential sequence variants are identified. Com-

pared to AmrPlusPlus and ARGs-OAP, ARIBA predicts full-length ARGs from fragmented metagenomic data and is therefore able to identify potential mutations and small sequence variations. However, when testing ARIBA's ability to identify genes distantly related to previously known genes, the method was not able to predict a single gene (Paper I).

GROOT (Graphing Resistance Out of meTagenomes) takes metagenomic sequence reads as input and applies a custom hierarchical local alignment using a hashing-based indexing scheme for classification (Rowe and Winn, 2018). The alignment is performed against a resistance database of choice, although GROOT provides a database that is a combination of ResFinder, ARG-ANNOT, and CARD where duplicated sequences have been removed. The chosen database is prepared by clustering the sequences based on a 90% sequence similarity to group the sequences into sets of similar genes. Each resulting set can be considered a multiple sequence alignment, which is converted into a variation graph that uses the multiple sequence alignment as the foundation. The alignment of sequence reads is then performed using an approximate nearest-neighbor search to identify the correct region of a graph followed by a hierarchical local alignment to obtain a fully aligned read. If a read has been successfully aligned to a variation graph, it is classified as antibiotic resistance derived, and gene annotations are performed based on the reference information of the graphs.

The method **deepARG** can handle both full-length and short-read sequences (Arango-Argoty et al., 2018). The classification consists of two steps, where the first step is a sequence alignment using DIAMOND against a custom-made resistance database based on sequences in the CARD and ARDB databases together with manually selected sequences from the UNIPROT database with the keyword antibiotic resistance. The results from the alignment then proceed to a custom-made deep learning model (DeepARG), which will annotate the sequences to antibiotic resistance categories. Although the method generally has a high precision and recall, it is not optimal for discovering resistance genes distantly related to the previously known, especially for short sequence reads, as shown in Paper I. This is mainly due to the first alignment step to the reference database, where many of the potential resistance genes are being discarded because of their low alignment coverage. However, note that among the methods validated in Paper I, deepARG was the one, except for fARGene, that had the highest sensitivity for identifying short reads from distantly related ARGs.

In Paper I, the method **fARGene** (fragmented Antibiotic Resistance Gene Identifier) is described. Compared to the competing methods, fARGene was specifically developed to identify *new* ARGs from fragmented metagenomic data. The method operates directly on short sequence reads that are aligned to profile HMMs that are optimized for the identification of short sequence

reads of resistance genes. The reads that are classified as belonging to the resistance gene class of interest are retrieved together with their read pair and then assembled using SPAdes meta (Nurk et al., 2017). Another classification is performed on the assembled contigs but this time using a threshold score optimized for full-length genes. The final output is predicted full-length ARGs. For full details, see Chapter 4 and Paper I.

3.2.4 Functional metagenomics

Shotgun metagenomics has allowed us to study the taxonomical and functional composition, including the resistome, of large microbial communities, including uncultivable bacteria. However, all methods described above rely on prior knowledge of the genes coding for resistance phenotypes to find them and their homologs. It is possible to successively expand the known gene catalog by predicting homologs to already know ARGS and perform functional verification of the identified candidates, but it is almost impossible to identify genes with little to no sequence identity to previously known ARGs through shotgun metagenomics. Functional metagenomics, however, is a technique that, instead of focusing on the similarity to known genes, focuses on the function of segments of DNA inserted in an external host, such as *Escherichia coli*, and by doing so can identify completely new ARGs. In a typical project that aims to identify ARGs by functional metagenomics, the method starts by extracting and preparing DNA from the microbial community of interest. Gene libraries are then constructed by allowing the DNA to be cloned into vectors (e.g. plasmids) that are incorporated into a host bacterium. Then, the bacterial clones are plated onto agar plates containing antibiotics. The antibiotic should be of a concentration such that it would kill the host bacteria that has not acquired any resistance gene. The DNA sequences of the inserts of the surviving bacteria are then determined, and ORFs can be predicted. There are, however, drawbacks to this approach. First, the amount of cloned DNA is limited. Therefore, genes that are low in abundance will often be missed. Furthermore, genes that are not being functional in the host bacterium will not be discovered, and the opposite is true when a gene that does not confer resistance in its native host is functional in the host bacterium. However, the latter will still provide information about the gene being able to confer resistance if being transferred to another species other than the native species. Another issue is whether the host bacterium is intrinsically resistant to the antibiotic being screened for. One approach to overcome this problem is to use another host. Despite these potential obstacles, functional metagenomics is still the only metagenomic approach that can identify completely new ARGs (Mullany, 2014; Riesenfeld et al., 2004a; Marathe et al., 2018).

3.3 Phylogenetic analysis

When new ARGs are identified, it is common to compare them to the previously known resistome to see how they relate to each other and hopefully understand their evolutionary origin. For this purpose, an inferred phylogeny is often used. Phylogenetic analysis can be described as our attempt to reconstruct the evolutionary history of organisms. Following the rapid increase in sequenced DNA and better computers, the ability to infer phylogeny has increased substantially, and today, phylogenies are used in almost every part of biology. The subject of molecular phylogeny is vast, and this section will only cover the basics with a short background to some of the approaches used to infer phylogeny.

Typically, phylogenetic relationships are visualized using a phylogenetic tree. The tree can be described as a branching diagram that displays the evolutionary relationships among the studied biological entities. The trees can be either unrooted or rooted, and the branching diagram of the tree is called a topology (Figure 3.2). The branches connect the nodes, which can be viewed as the start of a new lineage, while the branches represent the consistency of a genetic lineage over time. The number of possible topologies is a function of the number of sequences included in the analysis, and it increases drastically with increasing number of sequences. Because a phylogenetic tree is inferred from data, it is a challenging task, often even impossible, to find the true topology for large trees.

The methods to reconstruct a phylogenetic tree can be either distance- or character-based, where the distance-based methods utilize the calculated distances between every pair of sequences, summarized in a distance matrix, while the character-based methods compare all sequences simultaneously character by character in a multiple sequence alignment. There are many approaches that can be used to reconstruct a phylogenetic tree using the aforementioned methods. The distance matrix methods include, for example, neighbor-joining, least squares and minimal evolution, and for the character-based methods, we have maximum parsimony, maximum likelihood and Bayesian inference methods. Here, the distance matrix methods, maximum likelihood and Bayesian inference all assume an underlying evolutionary substitution model to describe the data and can therefore be described as model-based methods, whereas maximum parsimony does not (Nei and Kumar, 2000; Yang and Rannala, 2012).

Parsimony methods were one of the first methods to be used for inferring phylogenies and are based on the assumption that the reconstructed evolutionary events that lead to our given data should contain as few events as possible (Cavalli-Sforza and Edwards, 1967; Fitch, 1971). In the maximum parsimony method, this is incorporated by minimizing the number of changes on a phy-

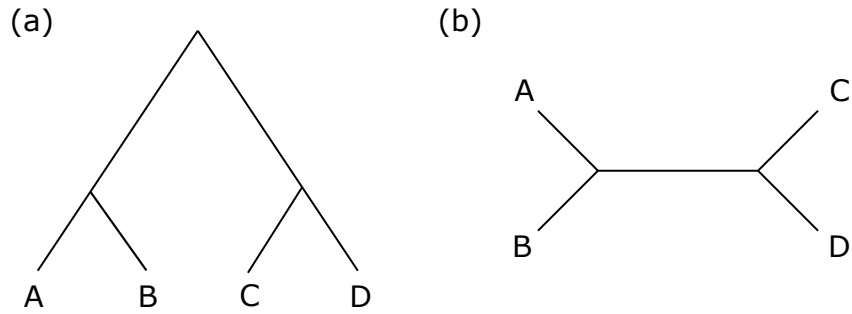

Figure 3.2: Rooted phylogenetic tree (a) and unrooted phylogenetic tree (b).

logenetic tree. The minimum number of changes required for a site is called the character length, and the tree score is the sum of all character lengths over all sites. Then, the tree that minimizes the tree score is the maximum parsimony tree. The simplicity of parsimony trees makes them easy to interpret and enables efficient computer programs to generate trees. However, simplicity is also a major drawback because parsimony does not incorporate any knowledge of the process of sequence evolution in tree reconstruction.

In distance matrix methods, the pairwise sequence distances are calculated assuming a nucleotide substitution model in the form of a Markov chain. The substitution models can assume an equal rate of substitution between any two nucleotides or different rates between the substitution of $T \leftrightarrow C$ and $A \leftrightarrow G$. Furthermore, the models can assume equal or unequal frequencies of the four nucleotides. To accommodate the variation in the local mutation rate, the distribution of rates for sites are often assumed to be gamma distributed (Yang and Rannala, 2012). When the distance matrix is calculated, there are several ways to reconstruct the topology. The least squares method attempts to minimize the calculated distances in the distance matrix (d_{ij}) to that of the expected distances in the tree ($\hat{d}_{ij}$), i.e., the branch lengths, leading to a score Q (Equation 3.1). The tree that has the smallest score will then be the least squares estimate of the true tree (Fitch and Margoliash, 1967).

$$Q = \sum_{i=1}^{s} \sum_{j=1}^{s} (\hat{d}_{ij} - d_{ij})^2 \tag{3.1}$$

In the minimal evolution method, the sum (S) of all branch length estimates is calculated for all possible topologies, and the topology that minimizes S is

chosen (Rzhetsky and Nei, 1992). Conversely, neighbor-joining also uses the minimal evolution principle, but instead of examining all possible topologies, it successively chooses a pair of taxa to join together based on the distances. Then, the distance matrix is updated with the joined taxon, replacing the two original taxa (Saitou and Nei, 1987). Distance-based methods are often computationally efficient, especially neighbor-joining, which could be desirable when analyzing large datasets. However, it is often important to choose the correct substitution model, and the distance calculation can be problematic when there are divergent sequences and many alignment gaps.

In maximum likelihood (ML) methods, for a specific substitution model, the likelihood of observing a given set of sequence data is maximized for each topology. The topology that yields the highest maximum likelihood is chosen as the final tree. It is important to note that the resulting likelihood of a tree is not the probability of the tree being the correct one (Felsenstein, 1981; Nei and Kumar, 2000). In ML tree estimation, there are two steps: first, optimization of branch length for each candidate tree, and then a search in the tree space for the maximum likelihood tree. Furthermore, for the likelihood to be computed, two assumptions must be made:

- On the given tree, evolution in different sites is mutually independent.
- Evolution in different lineages is independent.

Therefore, the overall likelihood for the tree is a product of the probabilities for each site, where the probabilities depend on the substitution model used. The likelihood of any particular site is computed by an average over the unobserved character states at the ancestral nodes (Felsenstein, 2004).

Bayesian inference is a general statistical inference methodology that is closely related to likelihood methods. In maximum likelihood, the model parameters are considered to be fixed constants, whereas in Bayesian inference, they are considered to be random variables with probability distributions. The parameters are assigned prior distributions that, when combined with the observed data, generate a posterior distribution from which the inference is based. Bayesian inference is based on Bayes's theorem, which states that given a hypothesis T, which in this case is a possible tree, and some observed data D, the probability of the hypothesis given the data is

$$\frac{P(T)|P(D|H)}{P(D)}. \tag{3.2}$$

Since the normalization constant P(D) generally involves high-dimensional

integrals and summation over all possible trees, the posterior probabilities of trees are often impossible to calculate exactly. However, the posterior distribution can be numerically estimated using Markov chain Monte Carlo methods, without the need for determining the denominator P(D) explicitly (Larget and Simon, 1999).

Although the Bayesian and ML methods share many properties, there is an ongoing controversy regarding which one to use when drawing inferences from data. Both methods have strengths and weaknesses. First, both methods utilize explicit models, which can be chosen to best suit the data, which is a significant advantage compared to maximum parsimony. Moreover, both methods require heavy computation, which requires fast computers and smart algorithms. The main difference is that the Bayesian methods use of prior distributions, which allows for the incorporation of prior knowledge about the system. The complication with priors is that such information is rarely available; generally, the prior is set by the software, and an incorrectly chosen prior can have an unexpected effect on the posterior (Yang and Rannala, 2012; Felsenstein, 2004).

In the papers presented in this thesis, phylogenies have been used to place the identified genes into an evolutionary context. In the majority of the papers, an approximate ML approach has been used (Price et al., 2010) due to the large number of sequences and the aforementioned computational requirements for ML trees. In Paper IV, however, we selected a smaller subset of all predicted genes to infer phylogenetic trees using analytical ML. The reason for choosing the ML over the other approaches was mainly because it uses complex substitution models to describe the biology of the data and that there are parallelized software available to perform the analysis (Stamatakis, 2014).

3.4 Functional verification

When new ARGs are predicted, it is important to perform a functional verification to verify that the gene confers resistance. If the isolate in which the gene was found is accessible, the bacterium of interest can be grown on antibiotic-containing media, and from there, a minimum inhibitory concentration (MIC) can be obtained. However, this will not provide information about whether the resistance is caused by the predicted gene or by some other acquired or pre-existing mechanism. One solution is to construct the predicted gene synthetically and then insert a vector that contains the synthetic gene into a host bacterium that is susceptible towards the antibiotic of interest. Then, the MIC can be determined for the cloned bacterium. In this case, a positive

result shows that the gene is functional in the tested host, although it might not be expressed in its original host. Conversely, a negative result does not necessarily mean that the predicted gene is not a resistance gene. Instead, the lack of function could be due to host incompatibility such that the gene could be functional in its original- and other potential hosts. Here, there are some approaches, such as codon optimization, that could be incorporated, but then it is not the predicted gene that is tested but a synthetic modification of it, and one could argue that because it is not naturally present in the environment, such verification is pointless. The second issue is to determine a reasonable threshold MIC value for considering a tested gene an ARG (Martínez et al., 2015). Here, the purpose of the study will affect what is to be considered a resistance gene. If the sole purpose is to show that the gene is functional, then a higher MIC value than the same strain without the gene would be considered sufficient. However, when the gene is expected to be functional but expressed at a low level in the host bacterium, an increased MIC value can be difficult to detect. Then, more sensitive methods are required, and in Paper V, a method that investigates the growth behavior in the presence of antibiotics was used. Although the MIC results for two predicted ARGs were inconclusive, the test showed a significant increase in the growth rate of the host carrying the predicted genes compared to the control in the presence of low concentrations of antibiotics.

4 Summary of results

This chapter provides a background and the main results of the six papers included in this thesis.

4.1 Paper I

Antibiotic resistance can be an intrinsic trait of bacteria but may also be acquired via mutations in existing chromosomal DNA or through the horizontal transfer of genes (Blair et al., 2015). Environmental bacterial communities harbor a large diversity of antibiotic resistance genes (ARGs) that can be mobilized and spread to pathogenic bacteria either directly or via human- and animal-related bacteria. Indeed, many of the current clinically relevant ARGs are believed to have an environmental origin (Forsberg et al., 2012; Bengtsson-Palme et al., 2017a). Furthermore, it has been shown that areas subjected to intense selection pressure, such as pollution from antibiotics, can further increase the abundance of ARGs (Gillings and Stokes, 2012; Bengtsson-Palme et al., 2014). It is therefore essential to investigate the environmental and commensal resistomes, including the previously uncharacterized ARGs, to understand the evolutionary process behind their evolution and mobilization. Furthermore, increased knowledge of the available resistome will facilitate surveillance and enable confinement actions before new ARGs reach clinical settings.

Shotgun metagenomics provides a holistic approach to study the environmental resistome via the sequencing of random fragments from a whole microbial community. However, the data are often highly fragmented and suffer from a considerable amount of noise. Assembly of the data is therefore often a challenging task that requires substantial computational resources. Furthermore, metagenomic data are often undersampled, which makes genes that are low in abundance hard to assemble, as are genes located on mobile genetic regions.

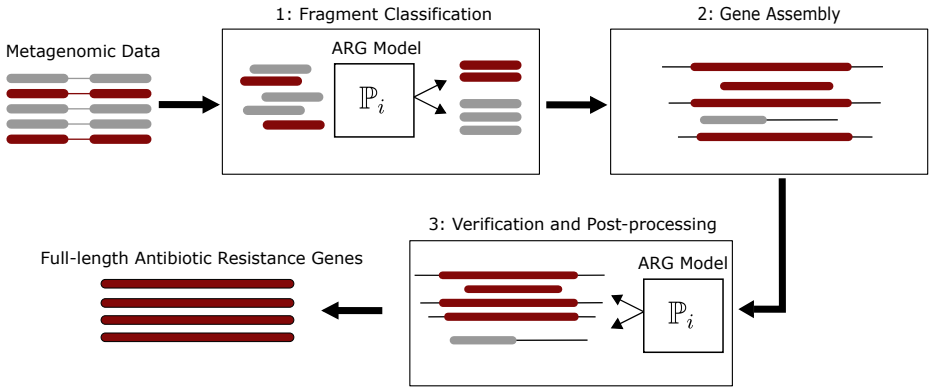

Figure 4.1: fARGene takes metagenomic paired-end data as input and analyzes the data with a probabilistic gene model that classifies the reads as coming from a resistance gene or not (Panel 1). The paired-end sequences of the positively classified reads are extracted, quality controlled and then assembled into full-length genes (Panel 2). The reconstructed gene sequences are once again classified by the ARG model (Panel 3). The output consists of nucleotide and amino acid sequences of the reconstructed full-length ARGs. The method can also be directly applied to whole genomes and metagenomic contigs, and then the classification, extraction and assembly of reads are not performed.

In Paper I, *Identification and reconstruction of novel antibiotic resistance genes from metagenomes*, we developed and implemented the fARGene method, which is specifically designed to identify and reconstruct novel ARGs from fragmented metagenomic data without the need for *de novo* assembly. This method consists of three main steps, and a schematic overview is presented in Figure 4.1. The method starts by analyzing short sequence reads using a probabilistic gene model in the form of a profile hidden Markov model (HMM) that has been optimized to identify short reads of the ARG class of interest (panel 1 in Figure 4.1). The reads classified as belonging to the modeled ARG class are then retrieved, together with their read-pair, even if only one of the reads in the pair was classified as an ARG. This first classification step significantly reduces the amount of data and enables the method to be computationally efficient. The retrieved reads are then assembled using a paired-end assembler, enabling less conserved regions at the end of the ARGs to be included in the assembly (panel 2, Figure 4.1). Then, the assembled contigs are once again subjected to the model, but this time, the threshold score used is optimized for full-length genes (panel 3, Figure 4.1). The contigs that are classified as belonging to the ARG class of interest are retrieved and scanned for open reading frames (ORFs). The ORFs are once again subjected to the full-length model, where the ORFs that passed this classification are retrieved if they are longer than a specific length

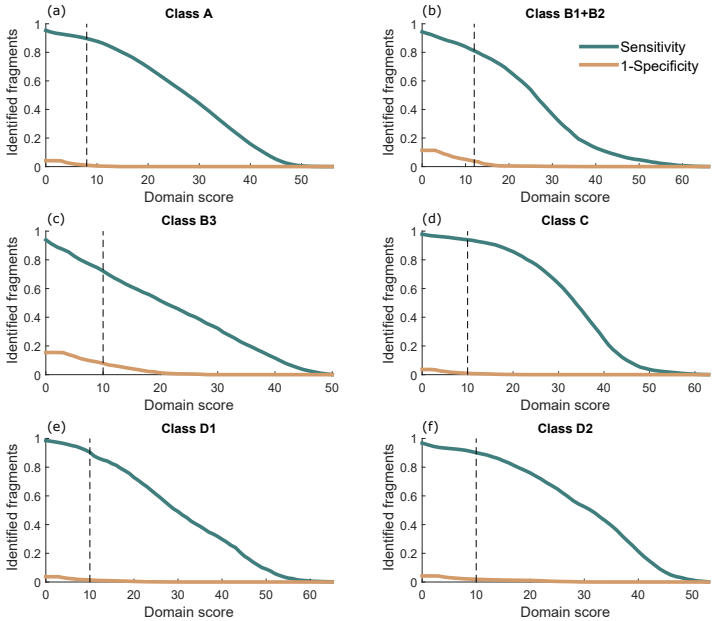

Figure 4.2: Results from the estimation of sensitivity and false positive rate (1-specificity) for the four β-lactamase classes. Each figure shows the performance of correctly classifying short reads as ARGs. The green curve represents the sensitivity, i.e., the fraction of correctly identified ARG reads, while the orange curve shows the false positive rate, i.e., the fraction of reads from genes without resistance phenotype incorrectly classified as an ARG. The dashed line corresponds to the set threshold score for each model.

threshold.

The method depends on optimized profile HMMs, and fARGene includes functionality to create and optimize customized models of the ARG class of interest. The optimization is based on leave-one-out cross-validation and is aimed at achieving sensitivity that is as high as possible while keeping the false positive rate low. For short-read data, this is achieved by excluding one gene at a time from the dataset of ARGs of interest. The excluded gene is randomly fragmented into short reads that are analyzed by a model created from the remaining ARGs in the dataset, and this process is then repeated for each gene. When estimating the false positive rate, it is important to carefully choose a negative dataset, preferably one with genes that have high sequence homology to the ARG class of interest but lacks the studied phenotype. The genes in the

chosen negative dataset are then fragmented into short reads and analyzed by the model created from the ARG class of interest. The sensitivity and false positive rate are then estimated based on the proportion of correctly classified reads, from which an optimal threshold score can be decided. The sensitivity and false positive rate for full-length genes are estimated similarly as for the short reads but by allowing the models to analyze full-length genes instead of short reads. Here, the threshold score should preferably be set such that there is a complete separation between the ARGs and the false positives.

As a case study, models were created to represent all genes within the four classes A, B, C and D of the clinically important β-lactamases. To capture the large sequence diversity within each class correctly, class B and D β-lactamases were separated into two models each. For full-length genes, the models were able to completely separate the ARGs from the negative dataset, while the threshold score for short-read data was set to keep the sensitivity as high as positive while keeping the false positive rate below 10% (Figure 4.2). The models were then used to analyze almost five billion reads from five metagenomic datasets, from which 221 ARGs were reconstructed, of which 58 were previously unknown (<70% sequence similarity to any reported gene in NCBI GenBank). The most commonly found gene class among the five datasets was class A with 149 reconstructed genes, followed by class B with 46, class D with 18 and class C with

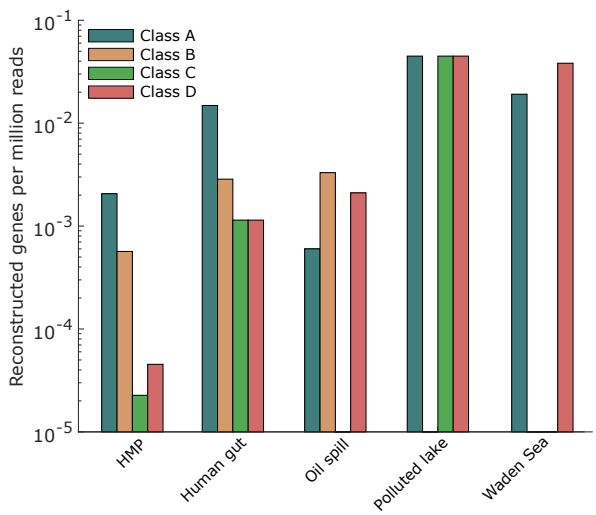

Figure 4.3: Number of reconstructed full-length genes per million reads (y-axis) for the four β-lactamase classes A, B, C and D.

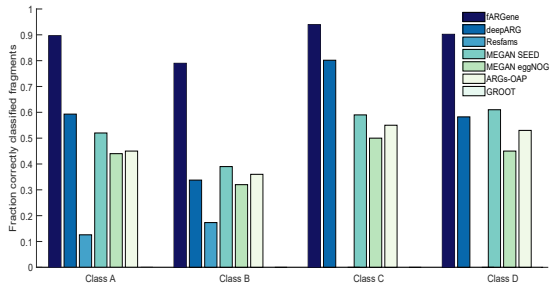

Figure 4.4: The ability to correctly classify short sequence reads for fARGene and five competing methods. The performance of fARGene was consistently higher than all compared methods.

8 reconstructed genes. The metagenome from an antibiotic-polluted Indian lake was the environment in which the highest relative abundance of reconstructed full-length genes for classes A, C and D was found, while full-length genes from class B had the highest relative abundance in an oil-contaminated deep-sea metagenome (Figure 4.3).

fARGene's ability to identify novel ARGs in fragmented metagenomic data was compared against five competing methods: deepARG (Arango-Argoty et al., 2018), Resfams (Gibson et al., 2015), ARGs-OAP (Yin et al., 2018), GROOT (Rowe and Winn, 2018) and MEGAN (Huson et al., 2016). fARGene had a consistently higher sensitivity than all compared methods, with an average sensitivity of 0.87 compared to 0.55 for deepARG and 0.52 for MEGAN, which were the second and third best, respectively (Figure 4.4). In addition, fARGene's ability to identify and reconstruct novel full-length ARGs was compared against the method ARIBA (Hunt et al., 2017), where fARGene was able to correctly reconstruct all 168 tested ARGs compared to none for ARIBA. Hence, fARGene offers both superior performance and the ability to reconstruct novel ARGs from fragmented metagenomic data and therefore provides the means to study the ARG composition of bacterial communities holistically. We also conclude that fARGene enables exploration of the resistome at an unprecedented scale.

4.2 Papers II and III

Carbapenems are important broad-spectrum antibiotics that are often used as a last-resort treatment for patients infected with multiresistant bacteria. Carbapenem resistance is often caused by carbapenemases, enzymes that can

often hydrolyze almost all known β-lactams in addition to carbapenems. Over the past few years, resistance towards carbapenems has rapidly increased in many regions of the world (Papp-Wallace et al., 2011), and carbapenem resistance genes, which up until some years ago had never been seen, are now detected in pathogens worldwide. The majority of mobile carbapenem resistance genes have been identified in clinical settings, but it is hypothesized that these genes originated from environmental bacteria. Many acquired and clinically relevant carbapenem resistance genes, such as VIM, IMP and NDM, belong to the class metallo-β-lactamases. This class is further divided based on molecular structure into the three subclasses B1, B2 and B3. In Papers II and III, the aim was to identify previously uncharacterized metallo-β-lactamases to obtain a more detailed picture of the origin and diversity of this important resistance gene class.

In Paper II, *Identification of 76 novel B1 metallo-β-lactamases through large-scale screening of genomic and metagenomic data*, an early version of the method fARGene, described in Paper I, was used to analyze more than five terabases of metagenomic data from both human and environmental bacterial communities, as well as bacterial genomes and plasmids available in the NCBI GenBank database. The probabilistic model used was optimized to find new metallo-β-lactamases of subclass B1 and was constructed using the 20 subclass B1 genes previously verified as of that date. In total, 76 novel subclass B1 genes were identified, and when clustered together with all previously reported subclass B1 genes using an amino acid sequence similarity cutoff of 70%, they formed 59 novel gene families. A phylogenetic tree was created based on one representative sequence from each of the 59 novel gene families together with the previously reported subclass B1 genes (Figure 4.5). Analysis of the resulting tree indicated that the genes could be organized into five groups, mainly defined by the taxonomy of the host species. We further noticed that all except one previously identified mobile subclass B1 gene clustered together with chromosomally encoded genes of the phylum Proteobacteria, indicating that many of the acquired subclass B1 genes have mobilized from species within this phylum.

Of the identified novel genes, 21 were selected for experimental verification, in which the genes were synthesized and inserted into an *Escherichia coli* host. A CarbaNP test (Nordmann et al., 2012) was conducted, and the results of this test showed that 18 of the tested genes had carbapenemase activity (Table 4.1). Among the verified genes, there was one, named SPS-1, that had a previously unseen binding site atypical for subclass B1. The results from this paper significantly extended the number of identified subclass B1 metallo-β-lactamases and provided a more detailed picture of their diversity and evolutionary history.

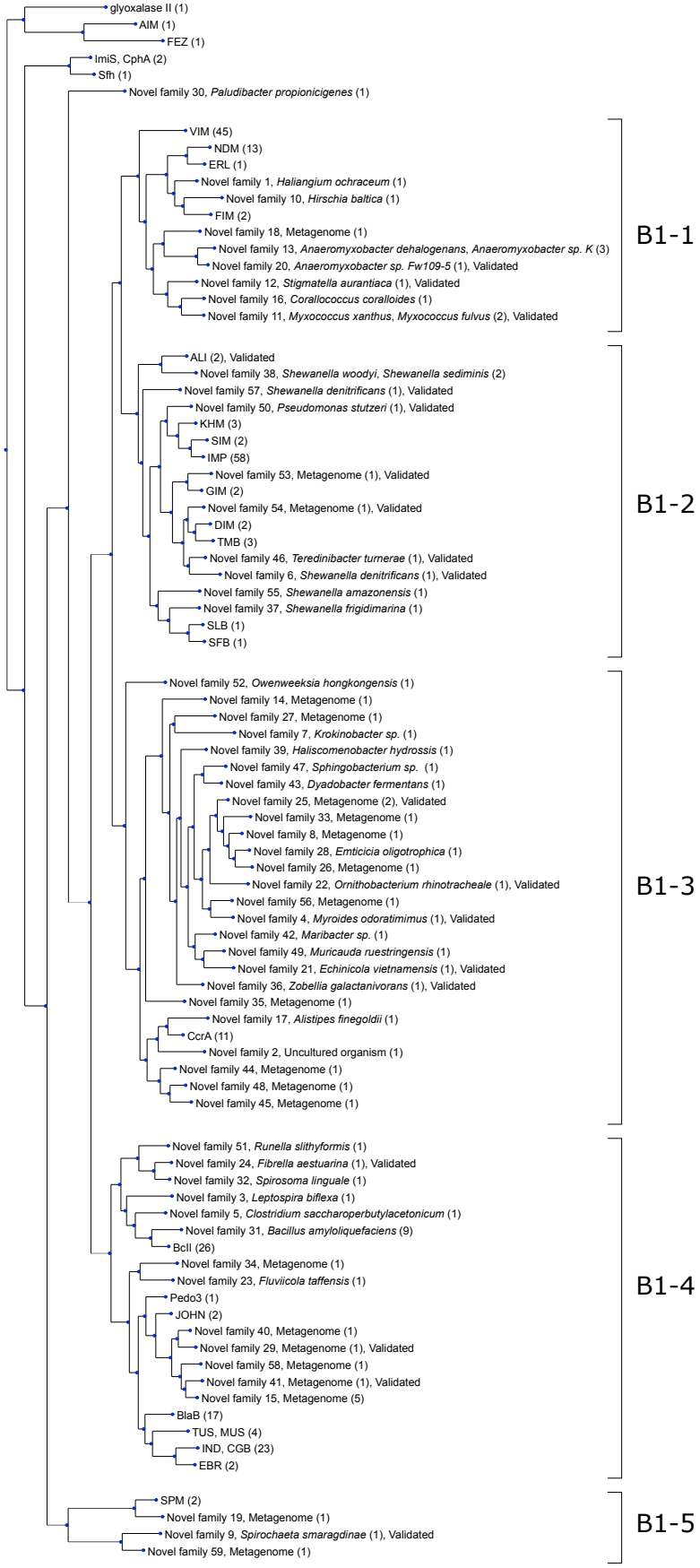
glyoxalase II (1)
AIM (1)
FEZ (1)
ImiS, CphA (2)
Sfh (1)
Novel family 30, *Paludibacter propionicigenes* (1)
VIM (45)
NDM (13)
ERL (1)
Novel family 1, *Haliangium ochraceum* (1)
Novel family 10, *Hirschia baltica* (1)
FIM (2)
Novel family 18, Metagenome (1)
Novel family 13, *Anaeromyxobacter dehalogenans*, *Anaeromyxobacter sp. K* (3)
Novel family 20, *Anaeromyxobacter sp. Fw109-5* (1), Validated
Novel family 12, *Stigmatella aurantiaca* (1), Validated
Novel family 16, *Corallococcus coralloides* (1)
Novel family 11, *Myxococcus xanthus*, *Myxococcus fulvus* (2), Validated
B1-1
ALI (2), Validated
Novel family 38, *Shewanella woodyi*, *Shewanella sediminis* (2)
Novel family 57, *Shewanella denitrificans* (1), Validated
Novel family 50, *Pseudomonas stutzeri* (1), Validated
KHM (3)
SIM (2)
IMP (58)
Novel family 53, Metagenome (1), Validated
GIM (2)
Novel family 54, Metagenome (1), Validated
DIM (2)
TMB (3)
Novel family 46, *Teredinibacter turnerae* (1), Validated
Novel family 6, *Shewanella denitrificans* (1), Validated
Novel family 55, *Shewanella amazonensis* (1)
Novel family 37, *Shewanella frigidimarina* (1)
SLB (1)
SFB (1)
B1-2
Novel family 52, *Owenweeksia hongkongensis* (1)
Novel family 14, Metagenome (1)
Novel family 27, Metagenome (1)
Novel family 7, *Krokinobacter sp.* (1)
Novel family 39, *Haliscomenobacter hydrossis* (1)
Novel family 47, *Sphingobacterium sp.* (1)
Novel family 43, *Dyadobacter fermentans* (1)
Novel family 25, Metagenome (2), Validated
Novel family 33, Metagenome (1)
Novel family 8, Metagenome (1)
Novel family 28, *Emticicia oligotrophica* (1)
Novel family 26, Metagenome (1)
Novel family 22, *Ornithobacterium rhinotracheale* (1), Validated
Novel family 56, Metagenome (1)
Novel family 4, *Myroides odoratimimus* (1), Validated
Novel family 42, *Maribacter sp.* (1)
Novel family 49, *Muricauda ruestringensis* (1)
Novel family 21, *Echinicola vietnamensis* (1), Validated
Novel family 36, *Zobellia galactanivorans* (1), Validated
Novel family 35, Metagenome (1)
Novel family 17, *Alistipes finegoldii* (1)
CcrA (11)
Novel family 2, Uncultured organism (1)
Novel family 44, Metagenome (1)
Novel family 48, Metagenome (1)
Novel family 45, Metagenome (1)
B1-3
Novel family 51, *Runella slithyformis* (1)
Novel family 24, *Fibrella aestuarina* (1), Validated
Novel family 32, *Spirosoma linguale* (1)
Novel family 3, *Leptospira biflexa* (1)
Novel family 5, *Clostridium saccharoperbutylacetonicum* (1)
Novel family 31, *Bacillus amyloliquefaciens* (9)
BcII (26)
Novel family 34, Metagenome (1)
Novel family 23, *Fluviicola taffensis* (1)
Pedo3 (1)
JOHN (2)
Novel family 40, Metagenome (1)
Novel family 29, Metagenome (1), Validated
Novel family 58, Metagenome (1)
Novel family 41, Metagenome (1), Validated
Novel family 15, Metagenome (5)
BlaB (17)
TUS, MUS (4)
IND, CGB (23)
EBR (2)
B1-4
SPM (2)
Novel family 19, Metagenome (1)
Novel family 9, *Spirochaeta smaragdinae* (1), Validated
Novel family 59, Metagenome (1)
B1-5

Figure 4.5: A phylogenetic tree describing the evolutionary relationship between the B1 metallo-β-lactamases predicted in this paper. The tree was created from representative sequences for the gene families generated through clustering predicted and previously characterized subclass B1 genes at a 70% amino acid sequence identity cutoff. Previously characterized subclass B1 genes are annotated with the protein name. Novel gene families predicted in this paper are annotated with the corresponding family number and the source (metagenome or species name of the host). The numbers in parentheses show how many unique genes there are in each family.

Table 4.1: Summary of the 22 experimentally tested subclass B1 genes.

Source data set	Gene ID	Predicted family	Proposed gene name	Assembled sequence length (aa)	Positive Carba NP test
RefSeq Plasmid	G04	4	MYO-1	266	Yes
RefSeq Bacteria	G06	6	SHD-1	265	Yes
RefSeq Bacteria	G09	9	SPS-1	263	Yes
RefSeq Bacteria	G12	11	MYX-1	262	Yes
RefSeq Bacteria	G13	12	STA-1	262	Yes
RefSeq Bacteria	G24	17		259	No
RefSeq Bacteria	G27	20	ANA-1	258	Yes
RefSeq Bacteria	G28	21	ECV-1	258	Yes
RefSeq Bacteria	G29	22	ORR-1	256	Yes
RefSeq Bacteria	G31	24	FIA-1	254	Yes
WWTP	G33	25		252	Yes
WWTP	G37	29		251	Yes
RefSeq Bacteria	G52	36	ZOG-1	247	Yes
Patancheru-well	G58	41		245	Yes
RefSeq Bacteria	G63	46	TTU-1	244	Yes
Pune-river	G65	48		242	No
RefSeq Bacteria	G67	50	PST-1	243	Yes[1]
RefSeq Bacteria	G69	52		242	No
Patancheru-well	G70	53		242	Yes
Oil spill	G71	54		240	Yes
RefSeq Bacteria	G74	57	SHN-1	238	Yes
RefSeq Bacteria	G77	ALI	ALI-2	246	Yes

[1]Strain tested

In Paper III, *Sewage effluent from an Indian hospital harbors novel carbapenemases and integron-borne antibiotic resistance genes*, we searched for novel metallo-β-lactamases in metagenomic data from sewage effluent from an Indian hospital. The hospital wastewater was assumed to contain fecal material from many individuals being treated with antibiotics. Therefore, we hypothesized that the likelihood of finding ARGs, including novel metallo-β-lactamases, was

Table 4.2: A list of the predicted metallo-β-lactamases and their closest BLAST hit against NCBI nr.

Gene name	MBL subclass	Length aa	Closest homolog in NCBI protein database	%identity aa	Accession number
DIM	B1	202	Subclass B1 metallo-beta-lactamase DIM-1 [Pseudomonas stutzeri]	100.0	WP_063860203.1
NDM	B1	248	Metallobetalactamase NDM-1 [Klebsiella pneumoniae]	100.0	AGC54622.1
IMP	B1	240	Beta-lactamase IMP-1 precursor [Pseudomonas aeruginosa]	100.0	CRX26419.1
VIM	B1	248	Subclass B1 metallo-beta-lactamase VIM-2 [Pseudomonadales]	100.0	WP_003108247.1
IMP	B1	240	Subclass B1 metallo-beta-lactamase IMP-15 [Pseudomonas aeruginosa]	100.0	WP_063860575.1
1N26	B1	241	Beta-lactamase [uncultured bacterium]	74.79	ALG03680.1
1N27	B1	242	Subclass B1 metallo-beta-lactamase [bacterium 336/3]	71.97	WP_054042800.1
2N30*	B2	252	ChpA family subclass B2 metallo-beta-lactamase [Aeromonas lacus]	51.48	WP_033113784.1
1N32	B1	241	Hypothetical protein A3D31_07435 [Fluviicola sp. RIFCSPHIGHO2_02_FULL_43_260]	55.04	OGS79779.1
1N4	B1	247	Subclass B1 metallo-beta-lactamase [bacterium 336/3]	73.55	WP_054042800.1
1N59	B1	244	Hypothetical protein A2041_05420 [Bacteroidetes bacterium GWA2_31_9b]	52.7	OFX20903.1
1N6	B1	246	Subclass B1 metallo-beta-lactamase [bacterium 336/3]	59.5	WP_054042800.1
1N8	B1	234	Subclass B1 metallo-beta-lactamase [Flectobacillus major]	70.67	WP_044171073.1
1N9	B1	240	Beta-lactamase [uncultured bacterium]	66.53	ALG03680.1
1N7	B1	273	Hypothetical protein Gferi_08260 [Geosporobacter ferrireducens]	48.0	AOT69571.1
POM-1	B3	223	B3 beta-lactamase [Pseudomonas otitidis]	100.0	ADC79563.1
L-1	B3	210	LWB2289.1 metallo-beta-lactamase L1 family protein [Acinetobacter sp. WC-743]	99.52	WP_009585815.1
3N14	B3	283	Subclass B3 metallo-beta-lactamase [Phenylobacterium sp. Root700]	69.58	WP_056733210.1
3N32*	B3	281	BJP_beta_lactamase [uncultured bacterium]	61.59	AIA10847.1
3N33	B3	289	BJP_beta_lactamase [uncultured bacterium]	61.4	AIA10847.1
3N40*	B3	290	Subclass B3 metallo-beta-lactamase [Phenylobacterium sp. Root700]	69.96	WP_056733210.1
3N51*	B3	299	Subclass B3 metallo-beta-lactamase [Novosphingobium sp. PP1Y]C	70.57	WP_013834039.1
3N55*	B3	284	Subclass B3 metallo-beta-lactamase [Phenylobacterium sp. Root700]K	70.82	WP_056733210.1
3N61*	B3	297	Subclass B3 metallo-beta-lactamase [Croceicoccus marinus]	57.65	WP_066847047.1
3N73#	B3	295	BJP_beta_lactamase [uncultured bacterium]	62.98	AIA10847.1
3N8	B3	297	Subclass B3 metallo-beta-lactamase [Sphingomonadaceae]E	55.85	WP_008831296.1
3N1*	B3	300	Subclass B3 metallo-beta-lactamase [Novosphingobium sp. Leaf2]	53.0	WP_056771586.1

*These genes were positive in the CarbaNP test
#This gene was negative in the CarbaNP test

increased.

The two probabilistic models used to search for the metallo-β-lactamases were optimized for subclass B3 and subclass B1/B2, respectively. In addition to the 20 genes included to construct the B1 model used in Paper II, the 18 experimentally validated genes reported in the paper were included, which significantly improved the ability to detect new subclass B1 genes. In total, 14 unique full-length genes representing subclass B1, one unique full-length gene representing subclass B2 and 12 unique full-length genes representing subclass B3 were predicted. Of the 27 unique full-length predicted genes, seven were previously characterized: NDM-1, IMP-1, IMP-15, VIM-2, DIM-1, POM-1 and L1. A list of the predicted genes and their closest BLAST hits against NCBI nr is shown in Table 4.2. The gene predicted as subclass B2, together

with seven of the predicted subclass B3 genes, was selected for functional verification. The genes were synthesized and transformed into an *E. coli* host, where their ability to hydrolyze carbapenems was assessed using the CarbaNP test. Six of the seven tested subclass B3 genes and the tested subclass B2 were positive for the test (Table 4.2). In addition, one of the predicted genes had the same atypical zinc-binding site as the subclass B1 gene SPS-1 predicted and validated in Paper II. The results of this paper showed that the diversity of metallo-β-lactamases in the sewage effluent from the Indian hospital is vast, with many of the clinically relevant mobile metallo-β-lactamases coexisting with several previously unknown ones. Because the effluent was expected to contain pathogens, the results show that this is an environment in which the recruitment of novel ARGs by pathogens may occur.

4.3 Paper IV

The number of identified metallo-β-lactamases increases every year, and today, there are more than twice as many characterized genes compared to that only a decade ago (Widmann et al., 2012; Somboro et al., 2018). Furthermore, this class of ARGs includes many mobile clinically relevant genes whose evolutionary origin and mobilization history are largely unknown. Previous studies, including the ones presented in Papers I-III, have indicated that the number of characterized metallo-β-lactamases only represents a fraction of the total diversity. Thus, as the amount of sequence data increases, the number of identified genes will most likely continue to increase, which calls for a more fine-tuned phylogeny such that the new genes can be directly placed into an evolutionary context.

In Paper IV, *An updated phylogeny of the metallo-β-lactamases*, we aimed to expand the phylogeny of the metallo-β-lactamases to provide a more detailed picture of their evolutionary history. Using the fARGene method described in Paper I, we analyzed more than 16 terabases of genomic and metagenomic data. The microbiomes analyzed were from pristine and polluted environments; from pig, poultry, and fish farms; and from the human microbiome. A total of 2290 unique metallo-β-lactamases were predicted, and when clustered using a 70% amino acid sequence similarity, these formed 817 gene families, of which 744 did not contain any previously known gene. Representative sequences for each cluster were used to create phylogenetic trees using an approximative maximum likelihood approach (Price et al., 2010), one each for subclass B1/B2 and subclass B3. Because of the problems related to inferring phylogenies from a large number of sequences, all previously known genes within each subclass together with a smaller set of chosen predicted sequences were used to create another set of phylogenetic trees using maximum likelihood (Stamatakis,

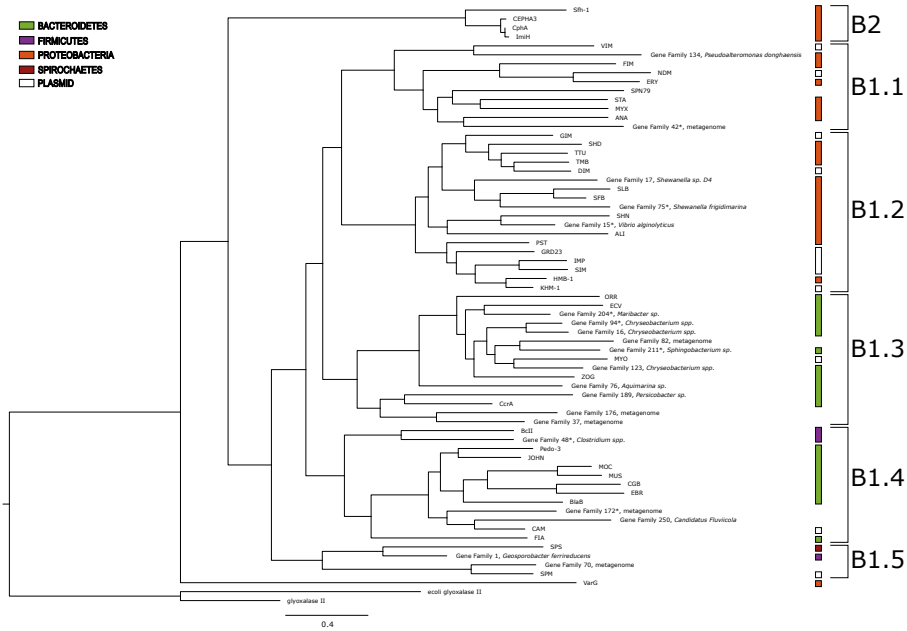

Figure 4.6: A phylogenetic tree describing all previously known subclass B1/B2 genes together with selected predicted gene families, with a member of the metallo-β-lactamase superfamily, glyoxalase II, as an outgroup. The phylum of the host species is represented by the colored bars on the far right. Gene families that contain genes predicted in previous studies, but have not yet been functionally validated, are marked with an asterisk.

2014). For subclass B1/B2, 318 gene families were predicted. Among these, 276 did not contain any previously known subclass B1/B2 gene. The phylogenetic analysis revealed that subclass B1 could be further divided into five phylogenetic groups, named B1.1-B1.5 here (Figure 4.6), as previously suggested in Paper II. The phylogenetic groups primarily correlated to the taxonomy of the host species, where groups B1.1-B1.2 almost exclusively consisted of genes located in Proteobacteria, group B1.3 was dominated by Bacteroidetes, and group B1.4 was a mix of Bacteroidetes and Firmicutes. Group B1.5 consisted of genes with the subclass B1 atypical binding site discovered in Paper II, together with the structural B1/B2 hybrid SPM-1 and its predicted homologs.

For subclass B3, 499 gene families were predicted, of which 468 were novel. The phylogenetic analysis showed that subclass B3 could be further divided into four phylogenetic groups, named B3.1-B3.4 here (Figure 4.7). Similar to subclass B1, the phylogenetic groups could to a large extent be explained by the phylum of the

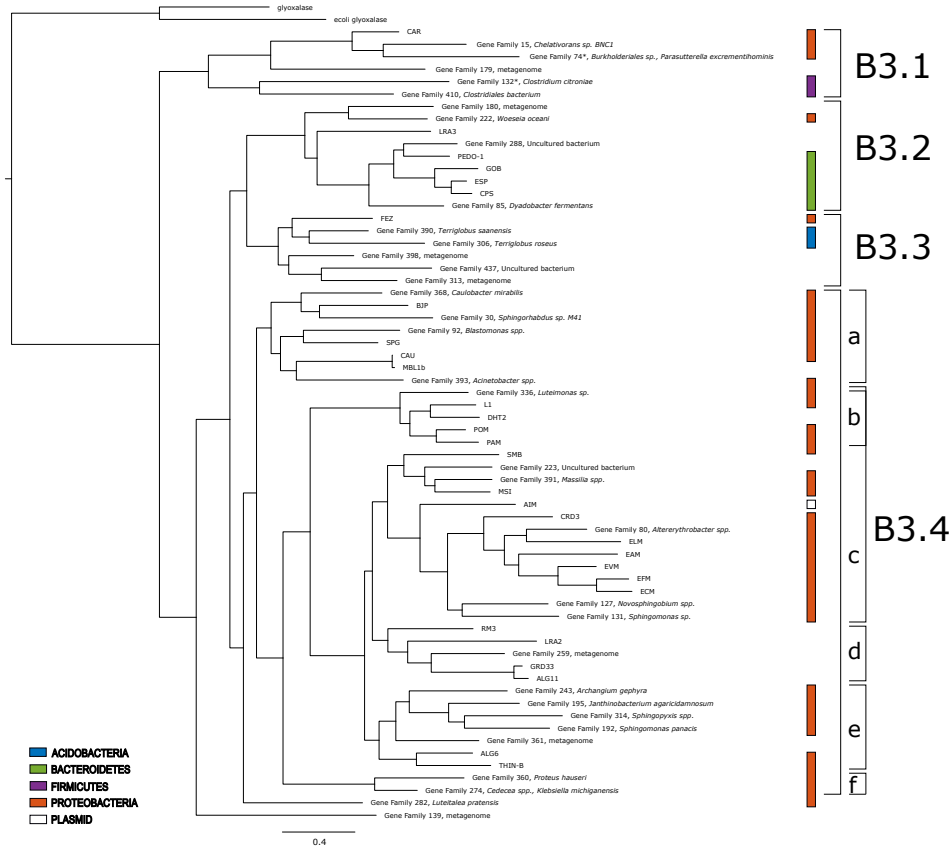

Figure 4.7: A phylogenetic tree describing all previously known subclass B3 genes together with selected predicted gene families, with a member of the metallo-β-lactamase superfamily, glyoxalase II, as an outgroup. The phylum of the host species is represented by the colored bars on the far right. Gene families that contain genes predicted in previous studies, but have not yet been functionally validated, are marked with an asterisk.

host species, where group B3.1 consisted of species of the phylum Proteobacteria (55%) and Firmicutes (26%), group B3.2 was dominated by Bacteroidetes (78%) but also contained Proteobacteria and Acidobacteria (11% each), and group B3.3 consisted mainly of Acidobacteria (55%) but also contained genes in Proteobacteria and Verrucomicrobia (33% and 8%, respectively). Group B3.4 almost exclusively contained genes in Proteobacteria species (96%).

All but two previously characterized acquired subclass B1 genes were located

in groups B1.1 and B1.2 (Figure 4.6), while the two previously known mobile subclass B3 genes, together with the three predicted genes located on plasmids all were located in group B3.4 (Figure 4.7). These groups all have in common that they almost exclusively consist of genes located in Proteobacterial species. Thus, it is plausible that today's clinically relevant metallo-β-lactamases have mobilized from a species within this phylum. When analyzing the zinc-binding sites of all previously known genes together with the predicted genes, a total of eight variants were discovered. Among these, one variant in subclass B1 (present in three gene families) and two in subclass B3 (present in eight and 17 gene families, respectively) were novel.

In conclusion, the results in Paper IV offer a more detailed view of the evolutionary history of the metallo-β-lactamases and show that the diversity of these genes is larger than what was previously known. The expanded nomenclature includes information about the phylogenetic groups, providing information about the taxonomy, possible origin and potential for mobilization.

4.4 Paper V

Tetracyclines are broad-spectrum antibiotics that are used to treat a wide range of bacterial infections. The efficiency together with the relatively few side effects has made tetracyclines one of the most used classes of antibiotics (Thaker et al., 2010). Tetracyclines are also used in subtherapeutic doses to promote growth of animals such as swine and poultry (Hao et al., 2014). The large consumption of these antibiotics has resulted in increases in a large variety of resistant bacteria, with new forms continuously emerging. Tetracycline resistance is based on three main mechanisms: active export of tetracycline via efflux pumps, ribosomal protection and enzymatic degradation. Over the past few decades, the number of confirmed tetracycline resistance genes has dramatically increased, mainly because of horizontal gene transfer between bacteria (Thaker et al., 2010). Although many of today's characterized tetracycline resistance genes are hypothesized to have originated from the environment, their exact evolutionary origin remains unclear (Chopra and Roberts, 2001; Forsberg et al., 2012).

In Paper V, *Comprehensive screening of genomic and metagenomic data reveals a large diversity of tetracycline resistance genes*, the aim was to expand the number of characterized tetracycline resistance genes to provide a more clear picture of their diversity, evolutionary history and origin. More than 12 terabases of genomic and metagenomic data were analyzed using the fARGene method (Paper I), with three models optimized to identify tetracycline resistance genes of the three mechanisms: efflux pumps, ribosomal protection and enzymatic

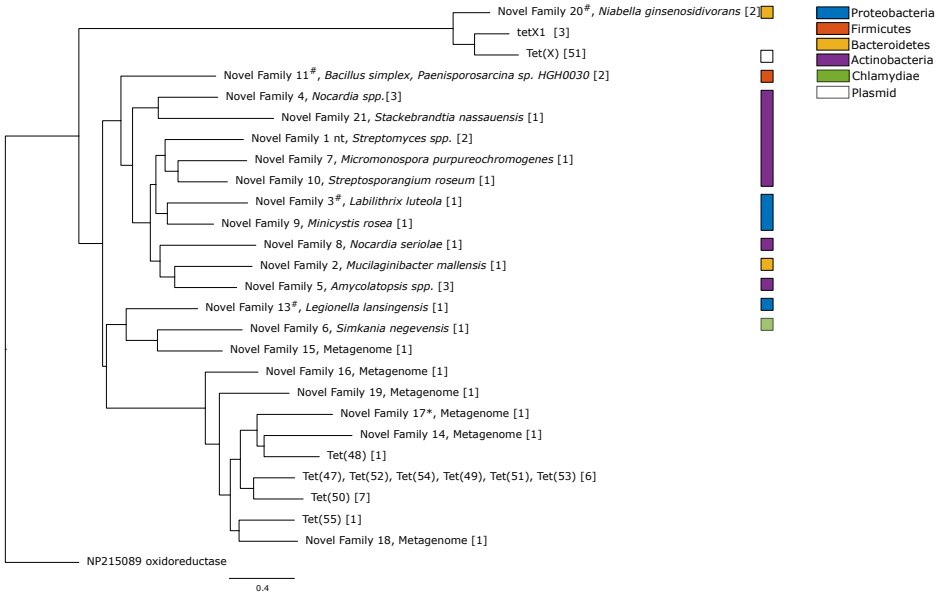

Figure 4.8: Phylogenetic tree of the predicted and previously known enzymatic tetracycline resistance genes. The number of unique genes in each group is presented within square brackets. The gene shown to be functional in *E. coli* is represented with an asterisk and the gene families that did not function in *E. coli* are represented with a hash sign.

degradation. A total of 1354 tetracycline resistance genes were predicted that were clustered into 195 gene families using a 70% sequence similarity, of which 164 did not contain any previously known gene. The ribosomal protection genes were the most diverse, followed by efflux genes and enzymatic degradation genes, with 116, 53 and 26 predicted gene families, respectively. Furthermore, phylum analysis of the identified species carrying tetracycline resistance genes showed that ribosomal protection genes were overrepresented in Firmicutes, Bacteroidetes and Actinobacteria while underrepresented in Proteobacteria; efflux pump genes were overrepresented in Proteobacteria while underrepresented in Firmicutes and Actinobacteria; and enzymatic degradation genes were overrepresented in Bacteroidetes while underrepresented in Proteobacteria.

Seventeen of the predicted novel genes were selected for functional verification in *E. coli*, of which five, six and six were enzymatic degradation, ribosomal protection and efflux pump genes, respectively. Of the tested genes, seven induced a resistance phenotype when expressed in *E. coli*, of which one was an enzymatic degradation gene and three were ribosomal protection and efflux

pump genes, respectively.

Phylogenetic trees were created for the three main mechanisms of tetracycline resistance using representative sequences from each gene family. The enzymatic degradation genes, which were rare, were divided into two groups, where the previously known mobile gene *tet*(X) (Whittle et al., 2001) was located in one group together with two homologs, while the remaining previously known enzymatic degradation genes were located in a separate group (Figure 4.8). The tree describing the predicted ribosomal protection genes was divided into two groups (Figure 4.9). The majority of the previously known mobile ribosomal protection genes were closest related to genes located in the phylum Firmicutes, except for the gene *otr*(A), which was surrounded by genes located in species of the phylum Actinobacteria. In the Actinobacteria group, there was a clade with genes located in Proteobacterial species, among which some were located on plasmids, indicating mobilization events from Actinobacterial species to Proteobacterial species. The phylogenetic tree describing the efflux pump genes is shown in Figure 4.10, and here, the previously known genes were distributed over the entire tree. However, the previously known efflux pump genes located on plasmids clustered together and were mainly surrounded by genes located in Proteobacterial species.

The results from this paper increase the knowledge of the potential tetracycline resistome and further describe the reservoir of resistance genes harbored by environmental and commensal bacteria. The phylogenetic analysis indicated that the mobile tetracycline resistance genes were mobilized from different phyla, where ribosomal protection genes originate from Firmicutes and Actinobacteria, while efflux pump genes originate from Proteobacteria. We further conclude that enzymatic degradation genes are scarce and that their evolutionary origin remains unknown.

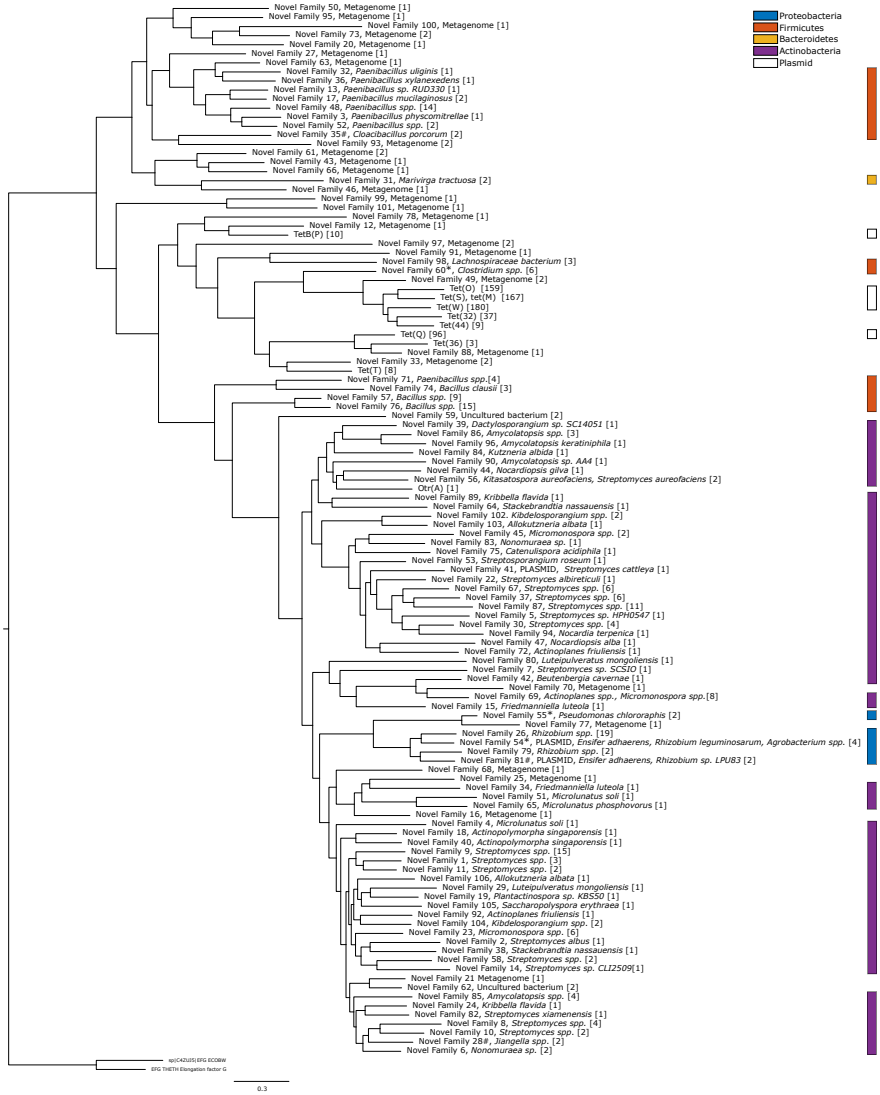

Figure 4.9: Phylogenetic tree of the predicted and previously known ribosomal protection genes. The number of unique genes in each family is presented within square brackets. The gene families functional in *E. coli* are represented with an asterisk and the gene families that did not function in *E. coli* are represented with a hash sign.

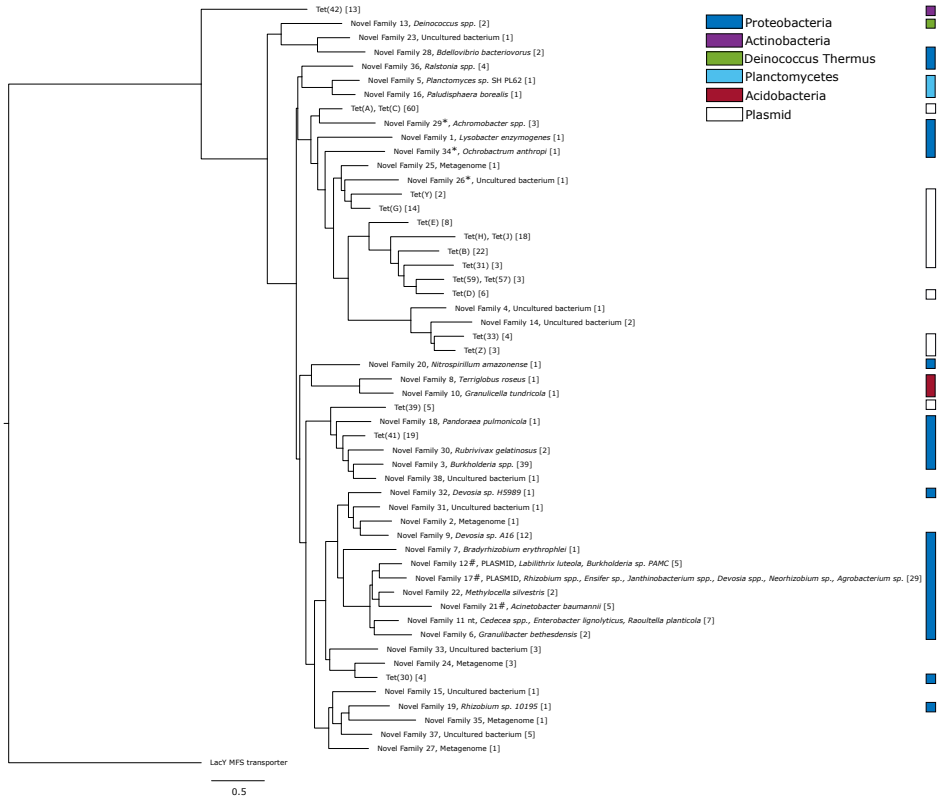

Figure 4.10: A phylogenetic tree describing the previously known efflux genes of MFS group 1 together with the predicted genes. The number of unique genes in each group is presented within square brackets. The gene families functional in *E. coli* are represented with an asterisk and the gene families that did not function in *E. coli* are represented with a hash sign.

4.5 Paper VI

The fully synthetic antibiotic class quinolones began clinical use in 1962, and later in the 1980s, the addition of fluorine to the quinolone yielded the fluoroquinolones. Because the antibiotic was not derived from natural components, it was considered unlikely that fluoroquinolone resistance would emerge. Nevertheless, in 1998, the first plasmid-mediated quinolone resistance (*qnr*) gene was discovered. Since then, the *qnr* genes have rapidly spread and are now distributed globally and are present in many bacterial genera. As with most other ARGs, the *qnr* genes are assumed to have originated from the environment (Sánchez et al., 2008). Although several families of *qnr* genes have been discovered over the past decade, their true abundance and diversity are still unknown.

The aim of paper VI, *Computational discovery and functional validation of novel fluoroquinolone resistance genes in public metagenomic data sets*, was to screen a high number of genomic and metagenomic datasets to discover previously unknown *qnr* genes and to estimate their abundances in environmental and commensal bacterial communities. A total of almost 13 terabases of genomic and metagenomic data were analyzed using a computational pipeline based on an HMM. A total of 362 843 *qnr* gene fragments were identified, and 611

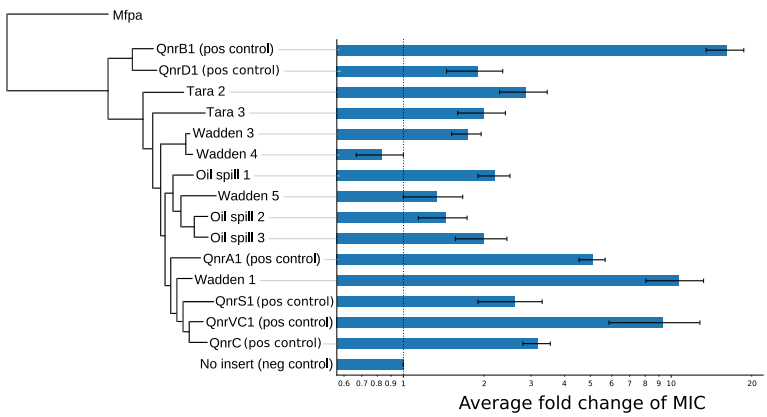

Figure 4.11: A phylogenetic tree describing the relationships between the previously known plasmid-mediated *qnr* genes and the predicted genes validated in Paper VI. The bar plot shows the fold change of ciprofloxacin minimum inhibitory concentration for all tested genes with three separate measurements each.

qnr genes were predicted. Among the 611 predicted *qnr* genes, all previously described plasmid-mediated *qnr* gene families were identified. Fifty-two of the predicted genes were reconstructed from the metagenomic data, of which 20 were not present or annotated as fluoroquinolone resistance genes in GenBank and therefore considered putatively novel *qnr* genes. Nine of these were selected for experimental verification, of which six showed an increased MIC to ciprofloxacin when expressed in *E. coli* (Figure 4.11). This paper provided the most exhaustive search for *qnr* genes in both genomic and metagenomic data ever conducted and contributed to the extended knowledge about their relative abundance in various bacterial communities.

5 Conclusions and discussion

In this thesis, bioinformatical methods and probabilistic models have been developed and used to investigate the unknown resistome of, in particular, genes providing resistance towards β-lactams, tetracyclines and fluoroquinolones. The methods have mainly been applied to fragmented metagenomic data from a diverse set of microbial communities, such as human- and animal-related microbiomes, as well as pristine and polluted environments. In addition, NCBI GenBank's repository of genomes, plasmids and nonredundant sequences has continually been explored. In total, more than 50 terabases of sequence data have been analyzed, resulting in the prediction of 4484 unique putative ARGs. The predicted genes were organized into approximately 1400 gene families, and among these, there were approximately 1000 gene families that did not contain any previously known ARG. In other words, approximately 2/3 of all predicted gene families in this thesis were novel, indicating that the current knowledge of ARGs only reflects the tip of the iceberg of the total resistome. Of the novel predicted genes, 54 have been functionally validated, and among these, 37 were shown to be functional when expressed in an *Escherichia coli* host.

We identified ARGs in almost every studied environment, supporting the hypothesis that antibiotic resistance determinants are present everywhere (Allen et al., 2010; D'Costa et al., 2011). In addition, the diversity of ARGs was shown to be vast, particularly in the environmental microbiomes, in which many new ARGs were predicted. However, the relative abundance of ARGs was generally higher in environments with different types of pollution, such as two rivers in India (Papers II, IV-VI), sewage effluent from a hospital in India (Paper III), aquaculture systems and fecal samples from poultry and pigs, where antibiotics are expected to have been extensively used (Paper IV). Nevertheless, the number of predicted new ARGs in these environments still, in many cases, greatly exceeded the number of identified previously known ARGs. In addition, many of the studied environments contained a large number of clinically relevant previously known ARGs; in particular, the sewage effluent from

an Indian hospital contained the majority of the currently clinically relevant carbapenemases NDM, VIM, IMP, KPC and OXA-48. Several new resistance genes coding for carbapenem resistance coexisted with these genes, suggesting that even if mobile ARGs are widely distributed within a microbial community, there still appears to be room for other ARGs with similar resistance profiles. However, the genes identified in Paper III were reconstructed from metagenomes; thus, an analysis of the genetic context and potential mobility was not possible. Nevertheless, even though they may be intrinsic, there is still a risk that they can be mobilized and transferred into pathogens (Stokes and Gillings, 2011; Ebmeyer et al., 2019a,b).

Phylum analysis of the predicted ARGs together with the previously known revealed that some phyla were overrepresented carriers of certain types of ARGs. Of the metallo-β-lactamases, subclass B1 genes were overrepresented in species of the phylum Bacteroidetes, while subclass B3 genes were overrepresented in species of the phylum Proteobacteria and underrepresented in Bacteroidetes species (Papers II and IV). For the tetracycline resistance genes, the ribosomal protection genes were overrepresented among Firmicutes, Bacteroidetes and Actinobacteria while underrepresented among Proteobacteria. The efflux pumps were overrepresented among Proteobacteria while underrepresented in Firmicutes and Actinobacteria (Paper V). The predicted enzymatic tetracycline resistance genes were overrepresented among Bacteroidetes; however, all but one of the identified Bacteroidetes species carrying enzymatic genes were homologs to the mobile *tet*(X) (Whittle et al., 2001). It therefore appears that some types of ARGs are more widely distributed among certain phyla compared to others. Although it is difficult to know for certain what causes this discrepancy, it could be a consequence of that some types of ARGs might be less compatible with bacteria from specific phyla, or that the barriers for horizontal gene transfer have hindered the genes from traveling vast taxonomical distances, or a combination thereof.

Phylogenetic analysis of the predicted and previously known genes further showed that many ARGs clustered based on the phylum of the host species. Interestingly, almost all of the previously known mobile subclass B1 genes, including the clinically relevant NDM, VIM and IMP, were located in parts of the phylogenetic tree that were dominated by predicted ARGs located in Proteobacterial species. The same pattern could be observed for subclass B3, where the two previously known mobile genes together with the predicted genes in Paper IV located on plasmids were most closely related to predicted ARGs located in Proteobacterial species. This result suggests that the acquired metallo-β-lactamases that we see today in clinical settings have mobilized from a Proteobacterial species. Furthermore, note that among subclass B1, there are 13 confirmed acquired genes to date, many being widespread and clinically relevant,

while among subclass B3, there are only two mobile genes (Somboro et al., 2018; Boyd et al., 2019; Wachino et al., 2011), not counting the three genes located on plasmids predicted in Paper IV. However, their resistance spectra are very similar (Bebrone, 2007) and, based on the number and in which environments they are predicted, they appear to be approximately equally diverse.

For the tetracycline resistance, the phylogenetic analysis in Paper V revealed that the ribosomal protection genes could be divided into two groups. The first group contained all previously known ribosomal protection genes except for the *otr*(A) gene. In this group, the predicted ARGs located in genomes were for all but one located in species of the phylum Firmicutes. Thus, it is plausible that these genes have mobilized from a species from the Firmicutes phylum. However, the second group containing *otr*(A) consisted almost exclusively of species within the phylum Actinobacteria. Within this group, there was a small clade of genes located in Proteobacterial species, some of which were located on plasmids. Hence, it appears that these genes have mobilized from Actinobacteria. Furthermore, the previously known mobile efflux genes, which are most often found in Proteobacterial species, were also surrounded by predicted genes mainly located in Proteobacterial species, indicating that this is also the phylum from which they have mobilized. Taken together, phylogenetic analysis can provide clues about the mobilization process and evolutionary origin of clinically relevant ARGs. However, we did not find any chromosomally located ARGs that were closely related to the previously known mobile genes. This result suggests that their origin might be from species that have not yet been sequenced or that the mobilization events occurred so far back in time that the original gene sequences are not conserved. Although there is evidence of relatively recent mobilization events (Ebmeyer et al., 2019a,b), the identified cases are few, and it is far from evident that this would be the case for all ARGs. Thus, knowledge about this process is rather scarce, and we do not fully understand the mobilization process, nor do we know the evolutionary origin of the majority of today's clinically relevant ARGs (Forsberg et al., 2012). More identified ARGs are therefore required to infer better phylogenies to increase the knowledge of this process and to better the evolutionary history at a higher detail.

In Papers II and IV, we also analyzed the variations of zinc-binding sites of the metallo-β-lactamases. In Paper II, we identified a new gene, SPS-1, that had an atypical and previously unseen variant of the zinc-binding sites for the subclass B1 genes. The gene was confirmed to hydrolyze carbapenems, and further studies from another research group showed that SPS-1 had a more narrow resistance spectrum compared to the previously known subclass B1 (Cheng et al., 2018). In Paper IV, we identified additional subclass B1 genes that had the same atypical binding site as SPS-1 but also predicted three genes with a third variant of the subclass B1 zinc-binding site. For the subclass B3 genes,

we predicted two previously unseen variants of zinc-binding sites present in a total of 35 genes, in addition to the two previously known variants. However, the structure of the predicted genes needs to be resolved, and they have to be functionally verified to be able to draw any conclusions on how these changes in zinc-binding sites affect the biochemical properties. Nevertheless, the results show that the variation among the metallo-β-lactamase binding sites is far more diverse than what was previously known, and further research about their functional diversity is therefore needed.

Although the results presented in this thesis have shown that the resistome is much larger and more diverse than what has been observed in the clinical setting, it is important to note that most of these genes do not pose an immediate threat to public health (Martínez et al., 2015). For an ARG to become relevant in the clinical setting, it needs to be located in a pathogenic or at least an opportunistic bacterium. Furthermore, the ARG needs to be functionally compatible with the host and properly expressed to be efficient. However, if the ARG is located on a mobile genetic element, it has the potential to become, if not already, problematic because it has a higher potential to spread between strains and species. An ARG located on a mobile genetic element in a pathogenic bacteria where it is also efficient would therefore constitute a much larger risk to public health. Several of the ARGs predicted in this thesis that were functional in *E. coli* were already located in opportunistic bacteria. This included, for example, the tetracycline efflux pump gene located on *Ochrobactrum anthropi* (Aguilera-Arreola et al., 2018) (Paper V) and the subclass B1 metallo-β-lactamase (named MYO-1) located in *Myroides odoratimimus* (Hu et al., 2016) (Paper II). The latter was furthermore located on a plasmid and thus already has a theoretical possibility to transfer to other species. Moreover, three ribosomal protection genes and two efflux pumps predicted to provide tetracycline resistance, identified in Paper V, were located on plasmids in soil-related bacteria. Here, one of the plasmid-located ribosomal protection genes was shown to be functional in *E. coli*. Additionally, three plasmid-located genes encoding subclass B3 metallo-β-lactamases were predicted in Paper IV. However, an ARG located on a plasmid does not necessarily imply that the ARG can be horizontally transferred between bacteria because there are many barriers reducing the possibility of genes moving between distantly related species. However, an ARG located on a broad-range plasmid makes it more likely to occur (Thomas and Nielsen, 2005). In addition, the acquisition of an ARG is often associated with a fitness cost, often leading to a reduction in the growth rate for the bacteria (Andersson and Levin, 1999). Therefore, if the ARG does not have any other beneficial function, the bacteria will not benefit from carrying the gene in the absence of antibiotics (Andersson and Hughes, 2010). Furthermore, an ARG can only be transferred horizontally if the carrier and the donor are in contact; hence, an ARG discovered in bacteria living in an environment far from any human-linked environment will be less

likely to find its way to a human pathogen (Baquero et al., 2009; Martínez, 2012). However, with the increase in human population and inefficient wastewater treatment plants, more human and commensal bacteria are disseminated in the environment, and the risk of such a transfer event thus increases (Martínez, 2008).

In environments where the abundance and diversity of ARGs are large, the probability of an ARG being mobilized will be higher. If the environment is also subjected to selection pressure, such as antibiotic pollution, and the fitness cost is not too high or negligible due to selection pressure, then the likelihood for the recipient bacteria to maintain the ARGs is further increased (Andersson and Hughes, 2010; Gullberg et al., 2014; Bengtsson-Palme and Larsson, 2015; Bengtsson-Palme et al., 2017a). Consequently, there will be environments in which the circumstances are beneficial for recruiting novel ARGs. Many of the metagenomes in which we find new ARGs comes from such environments, for instance, sewage effluent from an Indian hospital (Paper III), a river in India that flows through the city Pune and is subjected to untreated wastewater (Marathe et al., 2017) (Papers II, IV, V, VI), antibiotic-polluted lake and river (Papers I, V), fecal samples from animal farms and water from fish farms (Papers IV, V). Indeed, many of these environments showed a high diversity and relative abundance of ARGs. Furthermore, all of these environments are expected to harbor human- and animal-related bacteria; therefore, potential dissemination to human pathogens cannot be excluded. However, some of the identified genes are likely already located in human-related bacteria that stem from, for instance, fecal contamination (Pune river and hospital effluent). Nevertheless, since mobilization events probably occur continuously, the high diversity and abundance of ARGs in these environments increase the risk of an ARG being mobilized and recruited, and due to the diversity, eventually one that is efficient enough to be fixated by the bacterial community.

The amount of available sequencing data will undoubtedly continue to increase. With this comes the requirement for efficient and reliable methods to analyze the massive amounts of information. The method presented in Paper I was explicitly developed to predict new ARGs from large amounts of fragmented metagenomic data and provides an efficient way to identify resistance genes without the problematic *de novo* assembly. The method was able to identify all previously known ARGs included in the study and outperformed the compared competing methods in the identification of short-read sequences from new ARGs with a known resistance phenotype. The method was developed such that it could be applied to any class of ARGs and can therefore be used to, in theory, characterize the remaining unknown resistome of today's known resistance classes. However, the method based on probabilistic gene models in the form of profile HMMs requires prior knowledge of the gene class of interest. Therefore,

in contrast to functional metagenomics, it is not capable of detecting new forms of resistance but rather only homologs to the previously characterized. It is therefore essential to continue the search for new ARGs; as more genes are being characterized and added to databases, the search area will expand. Furthermore, it is crucial that the genes reported as ARGs to the databases are phenotypically tested; otherwise, there will be a downward spiral where false positives might be included in model creations, enabling even more false positives to be reported. Throughout this thesis we have therefore given formal names only to the ARGs that have been phenotypically confirmed. With well-curated and comprehensive databases, the future for the identification of ARGs using methods such as the one presented in this thesis looks bright, and we will most likely have a much more comprehensive picture of the available resistome in the coming years.

In conclusion, this thesis has contributed to the delineation of the present resistome and has significantly increased the number of characterized resistance genes. Analysis of the predicted and characterized genes has provided insights about the diversity, their evolutionary history and potential origin. Better knowledge of these genes is crucial to hinder the ongoing flow of genes providing resistance to both new and currently available antibiotics, from harmless bacteria to human pathogens. Increased knowledge about the resistome and the dispersion routes can enable management strategies to be implemented to limit the presence of and spread from environments where the potential for mobilization and selection of ARGs are elevated. Moreover, information about the structural and functional diversity of the resistome can be taken into account when developing new antibiotics. Furthermore, whole-genome sequencing is being progressively implemented as a diagnostic tool; therefore, with more comprehensive knowledge about the resistome and reliable and efficient bioinformatical methods, novel ARGs can be detected at an early stage, and actions can be taken before they become a serious issue for human health.

Bibliography

Abraham, E. P. and Chain, E. (1940). An enzyme from bacteria able to destroy penicillin. *Nature*, 146(3713):837.

Aguilera-Arreola, M. G., Ostria-Hernández, M. L., Albarrán-Fernández, E., Juárez-Enriquez, S. R., Majalca-Martínez, C., Rico-Verdín, B., Ruiz, E. A., del Socorro Ruiz-Palma, M., Morales-García, M. R., and Contreras-Rodríguez, A. (2018). Correct identification of ochrobactrum anthropi from blood culture using 16rrna sequencing: A first case report in an immunocompromised patient in mexico. *Frontiers in medicine*, 5.

Allen, H. K., Donato, J., Wang, H. H., Cloud-Hansen, K. A., Davies, J., and Handelsman, J. (2010). Call of the wild: antibiotic resistance genes in natural environments. *Nature Reviews Microbiology*, 8(4):251–259.

Andersson, D. I. and Hughes, D. (2010). Antibiotic resistance and its cost: is it possible to reverse resistance? *Nature Reviews Microbiology*, 8(4):260.

Andersson, D. I. and Levin, B. R. (1999). The biological cost of antibiotic resistance. *Current opinion in microbiology*, 2(5):489–493.

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., and Zhang, L. (2018). Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6(1):23.

Armbrust, E. V. and Palumbi, S. R. (2015). Uncovering hidden worlds of ocean biodiversity. *Science*, 348(6237):865–867.

Ayling, M., Clark, M. D., and Leggett, R. M. (2019). New approaches for metagenome assembly with short reads. *Briefings in bioinformatics*.

Baquero, F., Alvarez-Ortega, C., and Martinez, J. (2009). Ecology and evolution of antibiotic resistance. *Environmental Microbiology Reports*, 1(6):469–476.

Bebrone, C. (2007). Metallo-β-lactamases (classification, activity, genetic organization, structure, zinc coordination) and their superfamily. *Biochemical pharmacology*, 74(12):1686–1701.

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

Bengtsson-Palme, J., Kristiansson, E., and Larsson, D. J. (2017a). Environmental factors influencing the development and spread of antibiotic resistance. *FEMS microbiology reviews*, 42(1):fux053.

Bengtsson-Palme, J. and Larsson, D. J. (2015). Antibiotic resistance genes in the environment: prioritizing risks. *Nature Reviews Microbiology*, 13(6):396–396.

Bengtsson-Palme, J., Larsson, D. J., and Kristiansson, E. (2017b). Using metagenomics to investigate human and environmental resistomes. *Journal of Antimicrobial Chemotherapy*, 72(10):2690–2703.

Berendonk, T. U., Manaia, C. M., Merlin, C., Fatta-Kassinos, D., Cytryn, E., Walsh, F., Bürgmann, H., Sørum, H., Norström, M., Pons, M.-N., et al. (2015). Tackling antibiotic resistance: the environmental framework. *Nature Reviews Microbiology*, 13(5):310.

Blair, J. M., Webber, M. A., Baylay, A. J., Ogbolu, D. O., and Piddock, L. J. (2015). Molecular mechanisms of antibiotic resistance. *Nature reviews microbiology*, 13(1):42.

Boisvert, S., Laviolette, F., and Corbeil, J. (2010). Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of computational biology*, 17(11):1519–1533.

Boulund, F., Johnning, A., Pereira, M. B., Larsson, D. J., and Kristiansson, E. (2012). A novel method to discover fluoroquinolone antibiotic resistance (qnr) genes in fragmented nucleotide sequences. *BMC genomics*, 13(1):695.

Boulund, F., Pereira, M. B., Jonsson, V., and Kristiansson, E. (2018). Computational and statistical considerations in the analysis of metagenomic data. In *Metagenomics*, pages 81–102. Elsevier.

Boyd, D. A., Lisboa, L. F., Rennie, R., Zhanel, G. G., Dingle, T. C., and Mulvey, M. R. (2019). Identification of a novel metallo-β-lactamase, cam-1, in clinical pseudomonas aeruginosa isolates from canada. *Journal of Antimicrobial Chemotherapy*, 74(6):1563–1567.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59.

Bush, K. and Jacoby, G. A. (2010). Updated functional classification of β-lactamases. *Antimicrobial agents and chemotherapy*, 54(3):969–976.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). Blast+: architecture and applications. *BMC bioinformatics*, 10(1):421.

Cantas, L., Shah, S. Q., Cavaco, L., Manaia, C., Walsh, F., Popowska, M., Garelick, H., Bürgmann, H., and Sørum, H. (2013). A brief multi-disciplinary review on antimicrobial resistance in medicine and its linkage to the global environmental microbiota. *Frontiers in Microbiology*, 4:96.

Canton, R. (2009). Antibiotic resistance genes from the environment: a perspective through newly identified antibiotic resistance mechanisms in the clinical setting. *Clinical Microbiology and Infection*, 15(s1):20–25.

Carlos, F. A.-C. (2015). *Antibiotics and Antibiotic Resistance in the Environment*. CRC Press.

Cavalli-Sforza, L. L. and Edwards, A. W. (1967). Phylogenetic analysis: models and estimation procedures. *Evolution*, 21(3):550–570.

Cheng, Z., VanPelt, J., Bergstrom, A., Bethel, C., Katko, A., Miller, C., Mason, K., Cumming, E., Zhang, H., Kimble, R. L., et al. (2018). A noncanonical metal center drives the activity of the sediminispirochaeta smaragdinae metallo-β-lactamase sps-1. *Biochemistry*, 57(35):5218–5229.

Chopra, I. and Roberts, M. (2001). Tetracycline antibiotics: mode of action, applications, molecular biology, and epidemiology of bacterial resistance. *Microbiol. Mol. Biol. Rev.*, 65(2):232–260.

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore dna sequencing. *Nature nanotechnology*, 4(4):265.

Coates, A. R., Halls, G., and Hu, Y. (2011). Novel classes of antibiotics or more of the same? *British journal of pharmacology*, 163(1):184–194.

Consortium, H. M. P. et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214.

Curtis, T. P., Sloan, W. T., and Scannell, J. W. (2002). Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences*, 99(16):10494–10499.

D'Costa, V. M., King, C. E., Kalan, L., Morar, M., Sung, W. W., Schwarz, C., Froese, D., Zazula, G., Calmels, F., Debruyne, R., et al. (2011). Antibiotic resistance is ancient. *Nature*, 477(7365):457–461.

Delcher, A. L., Salzberg, S. L., and Phillippy, A. M. (2003). Using mummer to identify similar regions in large sequence sets. *Current protocols in bioinformatics*, (1):10–3.

Dunham, I., Hunt, A., Collins, J., Bruskiewich, R., Beare, D., Clamp, M., Smink, L., Ainscough, R., Almeida, J., Babbage, A., et al. (1999). The dna sequence of human chromosome 22. *Nature*, 402(6761):489–495.

Dykhuizen, D. E. (1998). Santa rosalia revisited: why are there so many species of bacteria? *Antonie van Leeuwenhoek*, 73(1):25–33.

Ebmeyer, S., Kristiansson, E., and Larsson, D. J. (2019a). Cmy-1/mox-family ampc β-lactamases mox-1, mox-2 and mox-9 were mobilized independently from three aeromonas species. *Journal of Antimicrobial Chemotherapy*, 74(5):1202–1206.

Ebmeyer, S., Kristiansson, E., and Larsson, D. J. (2019b). Per extended-spectrum β-lactamases originate from pararheinheimera spp. *International journal of antimicrobial agents*, 53(2):158–164.

Eddy, S. R. (2011). Accelerated profile hmm searches. *PLoS computational biology*, 7(10):e1002195.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461.

Ellington, M., Ekelund, O., Aarestrup, F. M., Canton, R., Doumith, M., Giske, C., Grundman, H., Hasman, H., Holden, M., Hopkins, K. L., et al. (2017). The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the eucast subcommittee. *Clinical microbiology and infection*, 23(1):2–22.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376.

Felsenstein, J. (2004). *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20(4):406–416.

Fitch, W. M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, 155(3760):279–284.

Forsberg, K. J., Reyes, A., Wang, B., Selleck, E. M., Sommer, M. O., and Dantas, G. (2012). The shared antibiotic resistome of soil bacteria and human pathogens. *Science*, 337(6098):1107–1111.

Forslund, K., Sunagawa, S., Kultima, J. R., Mende, D. R., Arumugam, M., Typas, A., and Bork, P. (2013). Country-specific antibiotic use practices impact the human gut resistome. *Genome research*, 23(7):1163–1169.

Gibson, M. K., Forsberg, K. J., and Dantas, G. (2015). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME journal*, 9(1):207.

Gillings, M. R. and Stokes, H. (2012). Are humans increasing bacterial evolvability? *Trends in ecology & evolution*, 27(6):346–352.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333.

Grüning, B. (2016). Amrplusplus. http://megares.meglab.org/amrplusplus.

Gullberg, E., Albrecht, L. M., Karlsson, C., Sandegren, L., and Andersson, D. I. (2014). Selection of a multidrug resistance plasmid by sublethal levels of antibiotics and heavy metals. *MBio*, 5(5):e01918–14.

Gupta, S. K., Padmanabhan, B. R., Diene, S. M., Lopez-Rojas, R., Kempf, M., Landraud, L., and Rolain, J.-M. (2014). Arg-annot, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrobial agents and chemotherapy*, 58(1):212–220.

Hall, B. G. and Barlow, M. (2004). Evolution of the serine β-lactamases: past, present and future. *Drug Resistance Updates*, 7(2):111–123.

Hall, T., Biosciences, I., and Carlsbad, C. (2011). Bioedit: an important software for molecular biology. *GERF Bull Biosci*, 2(1):60–61.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, 5(10):R245–R249.

Hao, H., Cheng, G., Iqbal, Z., Ai, X., Hussain, H. I., Huang, L., Dai, M., Wang, Y., Liu, Z., and Yuan, Z. (2014). Benefits and risks of antimicrobial use in food-producing animals. *Frontiers in microbiology*, 5:288.

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

Hendriksen, R. S., Munk, P., Njage, P., Van Bunnik, B., McNally, L., Lukjancenko, O., Röder, T., Nieuwenhuijse, D., Pedersen, S. K., Kjeldgaard, J., et al. (2019). Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nature communications*, 10(1):1124.

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., and Zamir, A. (1965). Structure of a ribonucleic acid. *Science*, 147(3664):1462–1465.

Hu, S.-h., Yuan, S.-x., Qu, H., Jiang, T., Zhou, Y.-j., Wang, M.-x., and Ming, D.-s. (2016). Antibiotic resistance mechanisms of myroides sp. *Journal of Zhejiang University-Science B*, 17(3):188–199.

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.

Huijbers, P. M., Flach, C.-F., and Larsson, D. J. (2019). A conceptual framework for the environmental surveillance of antibiotics and antibiotic resistance. *Environment International*, 130:104880.

Hunt, M., Mather, A. E., Sánchez-Busó, L., Page, A. J., Parkhill, J., Keane, J. A., and Harris, S. R. (2017). Ariba: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microbial genomics*, 3(10).

Hunter, S., Corbett, M., Denise, H., Fraser, M., Gonzalez-Beltran, A., Hunter, C., Jones, P., Leinonen, R., McAnulla, C., Maguire, E., et al. (2013). Ebi metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucleic acids research*, 42(D1):D600–D606.

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., and Tappu, R. (2016). Megan community edition-interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS computational biology*, 12(6):e1004957.

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., Lago, B. A., Dave, B. M., Pereira, S., Sharma, A. N., et al. (2016). Card 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic acids research*, page gkw1004.

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

Lakin, S. M., Dean, C., Noyes, N. R., Dettenwanger, A., Ross, A. S., Doster, E., Rovira, P., Abdo, Z., Jones, K. L., Ruiz, J., et al. (2016). Megares: an antimicrobial resistance database for high throughput sequencing. *Nucleic acids research*, 45(D1):D574–D580.

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

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357.

Larget, B. and Simon, D. L. (1999). Markov chain monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Molecular biology and evolution*, 16(6):750–759.

Larsson, D. J., Andremont, A., Bengtsson-Palme, J., Brandt, K. K., de Roda Husman, A. M., Fagerstedt, P., Fick, J., Flach, C.-F., Gaze, W. H., Kuroda, M., et al. (2018). Critical knowledge gaps and research needs related to the environmental dimensions of antibiotic resistance. *Environment international*, 117:132–138.

Li, D., Huang, Y., Leung, C.-M., Luo, R., Ting, H.-F., and Lam, T.-W. (2017). Megagta: a sensitive and accurate metagenomic gene-targeted assembler using iterative de bruijn graphs. *BMC bioinformatics*, 18(12):408.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676.

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32(14):2103–2110.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760.

Liu, B. and Pop, M. (2008). Ardb—antibiotic resistance genes database. *Nucleic acids research*, 37(suppl_1):D443–D447.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. *BioMed Research International*, 2012.

Marathe, N. P., Janzon, A., Kotsakis, S. D., Flach, C.-F., Razavi, M., Berglund, F., Kristiansson, E., and Larsson, D. J. (2018). Functional metagenomics reveals a novel carbapenem-hydrolyzing mobile beta-lactamase from indian river sediments contaminated with antibiotic production waste. *Environment international*, 112:279–286.

Marathe, N. P., Pal, C., Gaikwad, S. S., Jonsson, V., Kristiansson, E., and Larsson, D. J. (2017). Untreated urban waste contaminates indian river sediments with resistance genes to last resort antibiotics. *Water research*, 124:388–397.

Martínez, J. L. (2008). Antibiotics and antibiotic resistance genes in natural environments. *Science*, 321(5887):365–367.

Martínez, J. L. (2012). Bottlenecks in the transferability of antibiotic resistance from natural ecosystems to human bacterial pathogens. *Frontiers in microbiology*, 2:265.

Martínez, J. L., Coque, T. M., and Baquero, F. (2015). What is a resistance gene? ranking risk in resistomes. *Nature Reviews Microbiology*, 13(2):116.

Mathews, C. K., van Holde, K. E., and Ahern, K. G. (2000). *Biochemistry*. Benjamin/Cummings, San Francisco, California, 3 edition.

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., et al. (2014). The interpro protein families database: the classification resource after 15 years. *Nucleic acids research*, 43(D1):D213–D221.

Mullany, P. (2014). Functional metagenomics for the investigation of antibiotic resistance. *Virulence*, 5(3):443–447.

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research*, 40(20):e155–e155.

NCBI (2019). National Center for Biotechnology Information Genome Browser. `https://www.ncbi.nlm.nih.gov/genome/browse/`. [Online; accessed 2019-July-16].

Nei, M. and Kumar, S. (2000). *Molecular evolution and phylogenetics*. Oxford university press.

Nordmann, P., Poirel, L., and Dortet, L. (2012). Rapid detection of carbapenemase-producing enterobacteriaceae. *Emerging Infectious Disease*, 18(9).

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaspades: a new versatile metagenomic assembler. *Genome research*, 27(5):824–834.

O'Neill, J., Davies, S., Rex, J., White, L., Murray, R., et al. (2016). Review on antimicrobial resistance, tackling drug-resistant infections globally: final report and recommendations. *London: Wellcome Trust and UK Government*.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., et al. (2013). The seed and the rapid annotation of microbial genomes using subsystems technology (rast). *Nucleic acids research*, 42(D1):D206–D214.

Papp-Wallace, K. M., Endimiani, A., Taracila, M. A., and Bonomo, R. A. (2011). Carbapenems: past, present, and future. *Antimicrobial agents and chemotherapy*, 55(11):4943–4960.

Peng, Y., Leung, H. C., Yiu, S.-M., and Chin, F. Y. (2011). Meta-idba: a de novo assembler for metagenomic data. *Bioinformatics*, 27(13):i94–i101.

Peng, Y., Leung, H. C., Yiu, S.-M., and Chin, F. Y. (2012). Idba-ud: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428.

Powell, S., Szklarczyk, D., Trachana, K., Roth, A., Kuhn, M., Muller, J., Arnold, R., Rattei, T., Letunic, I., Doerks, T., et al. (2011). eggnog v3. 0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic acids research*, 40(D1):D284–D289.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature biotechnology*, 35(9):833.

Rhoads, A. and Au, K. F. (2015). Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5):278–289.

Riesenfeld, C. S., Goodman, R. M., and Handelsman, J. (2004a). Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environmental microbiology*, 6(9):981–989.

Riesenfeld, C. S., Schloss, P. D., and Handelsman, J. (2004b). Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.*, 38:525–552.

Rowe, W. P. and Winn, M. D. (2018). Indexed variation graphs for efficient and accurate resistome profiling. *Bioinformatics*, 34(21):3601–3608.

Rzhetsky, A. and Nei, M. (1992). A simple method for estimating and testing minimum-evolution trees.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425.

Sánchez, M. B., Hernández, A., Rodríguez-Martínez, J. M., Martínez-Martínez, L., and Martínez, J. L. (2008). Predictive analysis of transmissible quinolone resistance indicates stenotrophomonas maltophilia as a potential source of a novel family of qnr determinants. *BMC microbiology*, 8(1):148.

Sanger, F. (1960). Chemistry of insulin. *British medical bulletin*, 16(3):183–188.

Segawa, T., Takeuchi, N., Rivera, A., Yamada, A., Yoshimura, Y., Barcaza, G., Shinbori, K., Motoyama, H., Kohshima, S., and Ushida, K. (2013). Distribution of antibiotic resistance genes in glacier environments. *Environmental microbiology reports*, 5(1):127–134.

Somboro, A. M., Sekyere, J. O., Amoako, D. G., Essack, S. Y., and Bester, L. A. (2018). Diversity and proliferation of metallo-β-lactamases: a clarion call for clinically effective metallo-β-lactamase inhibitors. *Appl. Environ. Microbiol.*, 84(18):e00698–18.

Sommer, M. O., Dantas, G., and Church, G. M. (2009). Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science*, 325(5944):1128–1131.

Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Stokes, H. W. and Gillings, M. R. (2011). Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into gram-negative pathogens. *FEMS microbiology reviews*, 35(5):790–819.

Tacconelli, E., Carrara, E., Savoldi, A., Harbarth, S., Mendelson, M., Monnet, D. L., Pulcini, C., Kahlmeter, G., Kluytmans, J., Carmeli, Y., et al. (2018). Discovery, research, and development of new antibiotics: the who priority list of antibiotic-resistant bacteria and tuberculosis. *The Lancet Infectious Diseases*, 18(3):318–327.

Thaker, M., Spanogiannopoulos, P., and Wright, G. D. (2010). The tetracycline resistome. *Cellular and Molecular Life Sciences*, 67(3):419–431.

Thomas, C. M. and Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews microbiology*, 3(9):711.

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

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

Ventola, C. L. (2015). The antibiotic resistance crisis: part 1: causes and threats. *Pharmacy and Therapeutics*, 40(4):277.

Wachino, J.-i., Yoshida, H., Yamane, K., Suzuki, S., Matsui, M., Yamagishi, T., Tsutsui, A., Konda, T., Shibayama, K., and Arakawa, Y. (2011). Smb-1, a novel subclass b3 metallo-β-lactamase, associated with iscr1 and a class 1 integron, from a carbapenem-resistant serratia marcescens clinical isolate. *Antimicrobial agents and chemotherapy*, 55(11):5143–5149.

Walsh, C. (2003). *Antibiotics*. American Society of Microbiology.

Walsh, F. (2013). Investigating antibiotic resistance in non-clinical environments. *Frontiers in Microbiology*, 4:19.

Walsh, T. R., Weeks, J., Livermore, D. M., and Toleman, M. A. (2011). Dissemination of ndm-1 positive bacteria in the new delhi environment and its implications for human health: an environmental point prevalence study. *The Lancet infectious diseases*, 11(5):355–362.

Wang, Q., Fish, J. A., Gilman, M., Sun, Y., Brown, C. T., Tiedje, J. M., and Cole, J. R. (2015). Xander: employing a novel method for efficient gene-targeted metagenomic assembly. *Microbiome*, 3(1):32.

Wang, Z., Wang, Y., Fuhrman, J. A., Sun, F., and Zhu, S. (2019). Assessment of metagenomic assemblers based on hybrid reads of real and simulated metagenomic sequences. *Briefings in bioinformatics*.

Watson, J. D., Crick, F. H., et al. (1953). Molecular structure of nucleic acids. *Nature*, 171(4356):737–738.

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

Whittle, G., Hund, B. D., Shoemaker, N. B., and Salyers, A. A. (2001). Characterization of the 13-kilobaseermf region of the bacteroidesconjugative transposon ctndot. *Appl. Environ. Microbiol.*, 67(8):3488–3495.

Wichmann, F., Udikovic-Kolic, N., Andrew, S., and Handelsman, J. (2014). Diverse antibiotic resistance genes in dairy cow manure. *MBio*, 5(2):e01017–13.

Widmann, M., Pleiss, J., and Oelschlaeger, P. (2012). Systematic analysis of metallo-β-lactamases using an automated database. *Antimicrobial agents and chemotherapy*, 56(7):3481–3491.

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

Wu, R. and Taylor, E. (1971). Nucleotide sequence analysis of dna: Ii. complete nucleotide sequence of the cohesive ends of bacteriophage λ dna. *Journal of molecular biology*, 57(3):491–511.

Wu, Y.-W., Rho, M., Doak, T. G., and Ye, Y. (2012). Oral spirochetes implicated in dental diseases are widespread in normal human subjects and carry extremely diverse integron gene cassettes. *Appl. Environ. Microbiol.*, 78(15):5288–5296.

Yang, Y., Li, B., Ju, F., and Zhang, T. (2013). Exploring variation of antibiotic resistance genes in activated sludge over a four-year period through a metagenomic approach. *Environmental science & technology*, 47(18):10197–10205.

Yang, Y., Li, B., Zou, S., Fang, H. H., and Zhang, T. (2014). Fate of antibiotic resistance genes in sewage treatment plant revealed by metagenomic approach. *Water research*, 62:97–106.

Yang, Z. and Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature reviews genetics*, 13(5):303.

Yin, X., Jiang, X.-T., Chai, B., Li, L., Yang, Y., Cole, J. R., Tiedje, J. M., and Zhang, T. (2018). Args-oap v2. 0 with an expanded sarg database and hidden markov models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. *Bioinformatics*, 34(13):2263–2270.

Yong, D., Toleman, M. A., Giske, C. G., Cho, H. S., Sundman, K., Lee, K., and Walsh, T. R. (2009). Characterization of a new metallo-β-lactamase gene, blandm-1, and a novel erythromycin esterase gene carried on a unique genetic structure in klebsiella pneumoniae sequence type 14 from india. *Antimicrobial agents and chemotherapy*, 53(12):5046–5054.

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.