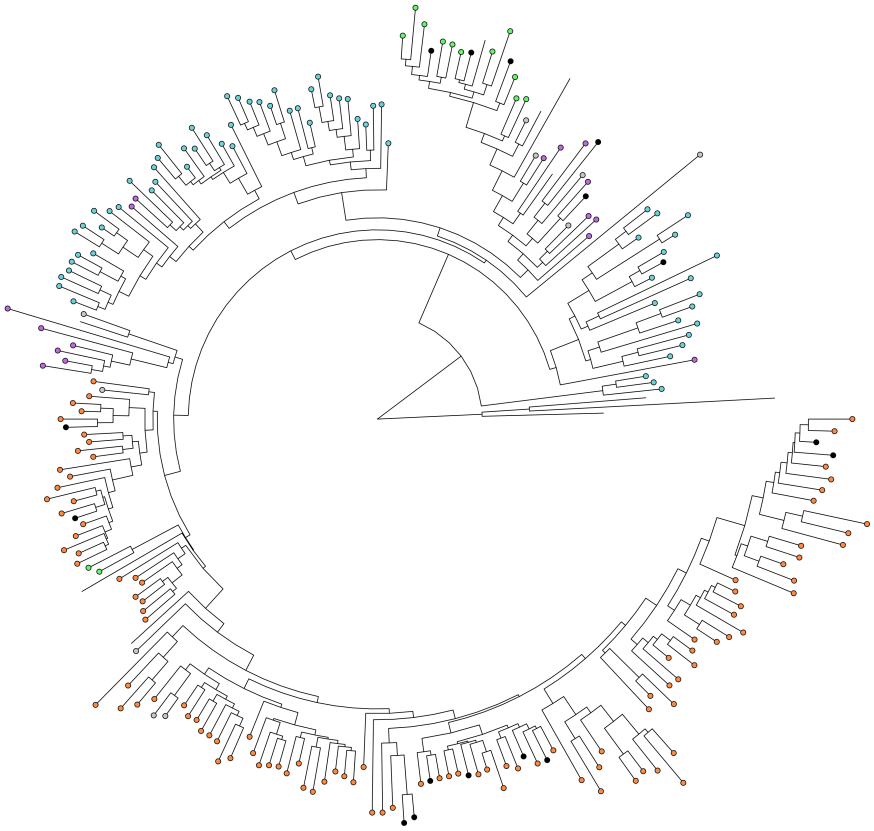
CHALMERS
UNIVERSITY OF TECHNOLOGY



Extensive Screening of Genomic and Metagenomic Data Identifies Novel Components of the Macrolide Resistome

Master's thesis in Biotechnology

David Lund

Department of Mathematical Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2020

Master's thesis 2020

Extensive Screening of Genomic and Metagenomic Data Identifies Novel Components of the Macrolide Resistome

David Lund

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2020

Extensive Screening of Genomic and Metagenomic Data Identifies Novel Components of the Macrolide Resistome
David Lund

Supervisor: Erik Kristiansson, Department of Mathematical Sciences
Co-supervisors: Anna Johnning, Department of Mathematical Sciences
Marcos Parras Moltó, Department of Mathematical Sciences
Examiner: Erik Kristiansson, Department of Mathematical Sciences

Master's Thesis 2020
Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Phylogenetic tree displaying Mph macrolide phosphotransferases identified in this study.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2020

Extensive Screening of Genomic and Metagenomic Data Identifies Novel Components of the Macrolide Resistome

David Lund
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Antibiotic resistance is growing among pathogenic bacteria all across the world, and has been called one of the most serious threats that humanity is facing. Typically, bacteria are able to develop resistance as a result of acquiring specific antibiotic resistance genes from other bacteria though so called horizontal gene transfer. One commonly used class of antibiotics for which resistance is spreading rapidly is macrolides. While a lot of research has been devoted to studying the genes that confer resistance to these antibiotics, the evolution of these macrolide resistance genes has not been determined. It has been suggested that resistance determinants that eventually find their way into the clinical environment originate from external environments, however the mechanisms behind this flow of resistance is not known. To prevent resistance to macrolide antibiotics from spreading further, it is therefore important to characterize how the resistance genes have evolved. Furthermore, knowledge about which genes are present in what environments will help with anticipating which genes might mobilize into the clinical environment in the future, and facilitate preemptive measures being taken.

This project aims to use a bioinformatic approach to characterize novel macrolide resistance genes, applying a computational method called fARGene. To achieve this, profile hidden Markov models have been developed that are able to identify two types of genes that confer resistance to macrolides, mediated by enzymes called Erm 23S rRNA methyltransferases and Mph macrolide phosphotransferases respectively, from biological sequencing data. The models have been used to analyze data representing over 400,000 bacterial genomes, and over 14 terabases of metagenomic data. Hundreds of gene families have been identified from the bacterial genomes, most of which are previously uncharacterized, and these have been analyzed based on their phylogenetic relationships. The results revealed a large variety of uncharacterized macrolide resistance genes that seem to have evolved primarily in bacteria from the phyla Firmicutes and Actinobacteria. In addition, several uncharacterized resistance genes that have potentially been mobilized have been identified from the results. No singular origin was determined for either of the analyzed gene classes, however the previously hypothesized evolutionary relationship between Erm methyltransferases and the housekeeping methyltransferase KsgA is supported by the results. In addition, the results from the analysis of metagenomic data indicate that the studied macrolide resistance genes are likely to mobilize from the human gut, naturally presenting a way through which the genes may enter the clinical environment.

Keywords: Antibiotic resistance, Macrolides, Bioinformatics, fARGene, Metagenomics, Evolution

Acknowledgements

I would like to thank Erik Kristiansson, for supervising this project and inviting me into his research group, Anna Johnning and Marcos Parras Mólto, for co-supervising the project and providing valuable insights, opinions and discussions, and Fanny Berglund, for creating the software around which the project is based, and providing feedback along the course of the project. None of this would have been possible without your help, and I look forward to keep working with all of you in the future.

A big thank you also goes out to the entire Kristiansson research group and colleagues at the Department of Mathematical Sciences, for welcoming me into the group and providing a great social environment during lunches and fikas.

Finally, I would like to thank my family and friends. The importance of your constant support can not be overstated.

David Lund, Gothenburg, June 2020

Contents

List of Figures

List of Tables

1

Introduction

1.1 Background

Antibiotic resistance is a phenomenon that continues to increase all over the world. Since the discovery of antibiotics, and their subsequent introduction into pharmaceutical treatments, pathogenic bacteria have emerged which are resistant to these antibiotics [1]. Diseases that can no longer be treated, due to the bacteria that cause them being resistant to multiple types of antibiotics, has been listed among the most serious threats to human health today [2, 3], and it is vital that strategies are developed to counteract this issue. To do this, more knowledge and understanding about the phenomenon must be obtained, especially as it relates to the antibiotics that are most frequently used for human treatment.

Macrolides make up a well-established class of antibiotics that has a long history of use in clinical settings [4]. Although there are many different types of macrolides, they are all composed of a large lactone ring of varying size, to which one or more sugars are attached via glycoside bonds. The lactone ring can also be substituted in various ways depending on the specific macrolide [5, 6]. To illustrate this, Figure 1.1 shows the chemical structure of erythromycin A, the first clinically relevant macrolide which was isolated from the bacterium *Saccharopolyspora erythraea*(formerly *Streptomyces erythreus*) in 1952 [7, 8]. The chemical structure that is shared among macrolides give them a hydrophobic nature that prevents them from efficiently permeating the outer membrane of Gram-negative organisms, and they are thereby primarily used as a treatment for infections caused by Gram-positive bacteria [9].

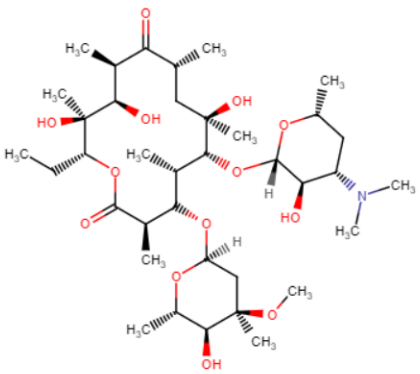


Figure 1.1: *Structure of erythromycin A.*

Since the discovery of erythromycin, more macrolides have been discovered and developed. Of note are the two semi-synthetic macrolides clarithromycin, which has increased activity against certain Gram-positive cocci, and azithromycin, which shows increased activity against Gram-negative bacteria while maintaining activity

against Gram-positive organisms and thereby bypasses one of the largest disadvantages of macrolides [10].

After entering the bacterial cell, macrolides act by binding to the 50S ribosomal subunit in the nascent peptide exit tunnel. There they act as a protein synthesis inhibitor [6], and while it was previously thought that the inhibition was a result of sterical hindrance interfering with the formation of the nascent peptides, it has since been shown that the mode of action is more complex. It instead seems macrolides can interact differently with different peptides by interfering with peptide bond formation rather than blocking the ribosomal exit tunnel. This is dependent on both the structure of the peptide, where if it lacks certain motifs its exit from the ribosome will be unobstructed by the presence of the drug, and the structure of the specific macrolide [11, 12].

Today, macrolides are among the most prescribed antibiotics in the world, with both human and animal applications [13, 14]. In addition to being the first choice antibiotic to combat certain infections, they also play a critical role as a preferred alternative treatment for patients where penicillin is not applicable due to allergies [13, 6]. Macrolides were present on the list that the World Health Organization published in 2017 describing the highest priority antibiotics for human medicine [15]. For these reasons, it is very problematic when pathogenic bacteria appear that show resistance to macrolide antibiotics.

How bacteria have evolved to become resistant to macrolides is something that remains largely unclear. While it is known that antibiotic resistance is often a result of acquiring specific antibiotic resistance genes (ARGs), and that bacteria can share these genes through so-called horizontal gene transfer [1], it is uncertain where the ARGs originated before making their way into pathogens in the clinical environment. While humanities overuse of antibiotics has certainly provided a selection pressure that promotes the acquisition of resistance determinants, the issue has been shown to be more complicated [16]. Understanding the evolution and mobilization of ARGs is of great interest to prevent these from spreading further. It has been suggested that there is a flow of ARGs from commensal and other harmless bacteria in external environments, including remote environments not polluted by antibiotics, into pathogens in the clinical environment, however further research is required to understand the process behind this transfer [17, 18]. This also highlights the fact that the macrolide resistome, the resistome meaning the complete collection of ARGs in bacteria [19], is likely much vaster than what is currently known, and through increasing the knowledge about the contents of the macrolide resistome, it can become easier to anticipate which ARGs might become problematic in the future and develop strategies to handle it [20].

1.2 Aim

The aim of this thesis is to characterize part of the macrolide resistome. As the resistome contains a massive amount of genes even when only considering ARGs that confer resistance to macrolide antibiotics, the project will be limited to the in-

vestigation of two types of macrolide ARGs. These genes encode so called Erm 23S rRNA methyltransferases and Mph macrolide phosphotransferases respectively, and the project aims to get a comprehensive overview of these resistance determinants by identifying novel gene families encoding these, as well as investigating the phylogenetic relationships of the identified genes, both new and known, in an attempt to elucidate how these genes have evolved in bacteria. This will be achieved through analysis of large amounts of sequencing data representing bacterial genomes, using profile hidden Markov Models that will be built for this purpose and a software called fARGene [21].

In addition, the project aims to identify environments from where these macrolide ARGs might mobilize. This aim will be achieved through the acquisition and analysis of several terabytes of shotgun metagenomic data from a variety of environments, using a similar approach to what is described above. Information about such environments can empower us to take measures to keep the flow of resistance determinants to a minimum.

2

Theory

In this chapter, the mechanisms that give rise to macrolide resistance are described. The two types of genes that the project is based around are given extra focus, with a summary of the current knowledge of their origin, spread, and evolution. In addition, the chapter also includes a section about the theory behind bioinformatic methods, specifically hidden Markov models, and how they can be applied to identify genes in DNA sequencing data.

2.1 Acquired resistance to macrolide antibiotics

As previously mentioned, resistance to macrolides has developed in pathogenic bacteria [1]. When compared to other types of antibiotic resistance, macrolide resistance is particularly problematic, since today macrolide resistant bacteria have become widespread throughout the world rather than being concentrated in specific areas [22]. The majority of macrolide resistance mechanisms fall into one of three broad categories; target modification, efflux, or enzymatic inactivation of the drug [23]. In addition, genetic mutations involving base substitutions in the macrolide binding site have also been shown to lead to a macrolide-resistant phenotype [24], however these will not be further discussed as they are not able to transfer horizontally. To limit the scope of this study, it will focus on two specific resistance mechanisms; target alteration of the 23S ribosomal subunit, mediated through *erm* family methyltransferases, and inactivation of macrolides through phosphorylation, mediated through *mph* family phosphotransferases [25, 26].

2.1.1 Target alteration by 23S rRNA methyltransferases

Bacteria are able to develop resistance to macrolide antibiotics by acquiring genes from the *erm* methyltransferase family. These genes encode enzymes which are able to either mono- or dimethylate the N6 position of nucleotide A2058 (*Escherichia coli* annotation) of the 23S rRNA subunit [27]. Methylation at this position prevents the drug from binding to the ribosomal RNA, and thereby renders the host bacteria resistant to it's inhibitory effects [25] (Figure 2.1). This mechanism has been stated to be the most common way that bacteria acquire resistance to macrolides [25], with the gene *erm(B)* being the most widespread variant [22]. Genes associated with this mechanism generally also confer resistance to lincosamide and streptogramin b-type

antibiotics, and the resulting phenotype is therefore called the MLSb phenotype [28].

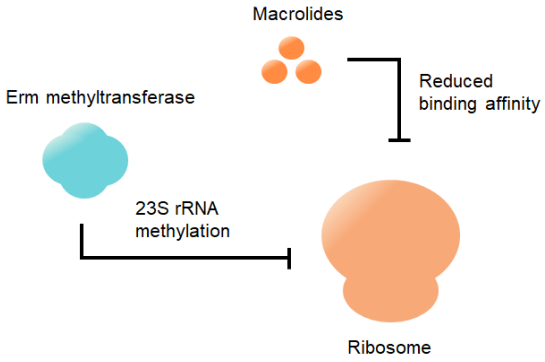


Figure 2.1: *Target methylation by Erm 23S rRNA methyltransferase enzymes.*

The origins of *erm* genes have not been completely characterized, however it has been reported that *erm* methyltransferases, together with efflux-pumps, make up the main mechanisms behind self-resistance in macrolide producers [29]. These producers are bacteria of the Actinobacteria phylum, and as a measure of protection against the antibiotics they produce they use *erm* mediated methylation of their own ribosomes. The most obvious example of this is the erythromycin producing bacteria *S. erythraea*, which harbors a resistance gene of this type called *erm(E)* [30]. While it could reasonably be speculated that all *erm* genes originate from macrolide producing bacteria and have since mobilized, it has also been suggested that there is an evolutionary relationship between *erm* methyltransferases and the housekeeping methyltransferase *ksgA*/*rsmA* [31].

KsgA is a highly conserved enzyme that methylates two adjacent adenosines in the 16S ribosomal subunit. In bacteria this enzyme is not essential for growth, however it does provide growth advantage and competitiveness [32]. Due to the high sequence homology between *ksgA* and *erm* genes, in addition to the high similarity in structural architecture and function between the two, it has been proposed that the two gene types either evolved separately from a common ancestor or that *erm* genes might have descended from one or several preexisting *ksgA* genes [31]. It has also been reported that some *erm* variants have arisen through recombination of two previously existing genes, one example being the gene *erm(33)* which is a result of recombination between *erm(A)* and *erm(C)* [33].

In total, 44 classes of *erm* genes have been characterized, some of which have been found on mobile genetic elements [34, 35]. Out of these, several classes have been found in pathogenic bacteria, including *erm(A)*, *erm(B)*, *erm(C)*, *erm(F)*, and *erm(39)* [36, 37]. These genes have predominantly been encountered in bacteria from the Firmicutes phylum, with *erm(A)* and *erm(C)* typically being associated with staphylococci and *erm(B)* being associated with streptococci and enterococci [38]. In addition, *erm(B)* has also been encountered in *Campylobacter* species isolated from fecal samples and has shown to always be associated with genes that confer resistance to multiple antibiotics [39]. By contrast, *erm(F)* is found in anaerobic

species such as *Bacteroides* but has also been encountered in *Haemophilus influenzae* [38, 36], while *erm(39)* has been identified in *Mycobacterium* species, which belong to the Actinobacteria phylum.

It comes as no surprise that these resistance genes are found in Actinobaceria however, for reasons described above. Other examples of such bacteria that have been shown to harbor *erm* genes are clinical *Corynebacteria* isolates, opportunistic pathogens which were found to harbor a combination of *erm(A)*, *erm(C)*, as well as *erm(X)* genes [40]. Select strains of *Corynebacteria* have been shown to carry *erm(B)* as well [41]. The fact that these genes appear across multiple phyla speaks to their mobility, and thereby the importance of characterizing from where they were originally mobilized.

Other *erm* genes have been identified on mobile elements, though they are not yet known to have transferred into pathogens. One such example is *erm(48)*, that was identified on a plasmid in *Staphylococcus xylosus* isolated from bovine mastis milk [42]. While *erm* genes being found in bacteria isolated from milk and fecal samples suggests that the genes might originate from the human and animal microbiota, genes of this type have been also been encountered in the environment. For example, the gene *ermG* was discovered in *Bacillus sphaericus* isolated from soil samples [43], however the origin of these soil samples were unclear from the original study and thereby there is a possibility that the gene was transferred to the soil via humans or animals. Furthermore, *erm(B)* and *erm(F)* have also been encountered in bacteria isolated from water samples taken from surface drainage connected to swine farms [44], strengthening the hypothesis that these genes might originate from the animal microbiome. These genes have also been found in isolates from a wastewater treatment plant in Zagreb, along with several other variants [45].

2.1.2 Drug inactivation by macrolide 2'-phosphotransferases

As previously mentioned, another mechanism by which resistance can occur is through inactivation of the drug. One way that macrolides can be inactivated is through phosphorylation by Mph 2'-macrolide phosphotransferase enzymes. These enzymes interact with 14- 15- and 16-membered macrolides by attaching a phosphate group to the 2'-OH group on the macrolide (Figure 2.2). This makes the drug unable to interact with A2058 on the 23S rRNA subunit, thereby making bacteria that harbor genes that encode these enzymes resistant [27].

mph genes are present in bacteria from a range of different origins, and it is known that at least four variants (*mph(A)*, *mph(B)*, *mph(C)* and *mph(E)*) have mobilized into pathogens [26]. Furthermore, it is known that these enzymes do not all have the same substrate specificity, notably MphI and MphK show a more narrow substrate range than other Mph enzymes, lacking the capability to phosphorylate macrolides containing a C3 cladinose [26]. Additionally, it has been shown that MphB has a broader substrate range than MphA, being able to act upon 16-membered macrolides in addition to 14- and 15-membered macrolides [46], while simultaneously being unable to act upon azithromycin unlike MphA [47]. There are conflicting reports

about the latter statement however, with some sources claiming that MphB does have an affinity towards azithromycin, although this may also be a result of amino acid substitutions in the enzyme [48, 49].

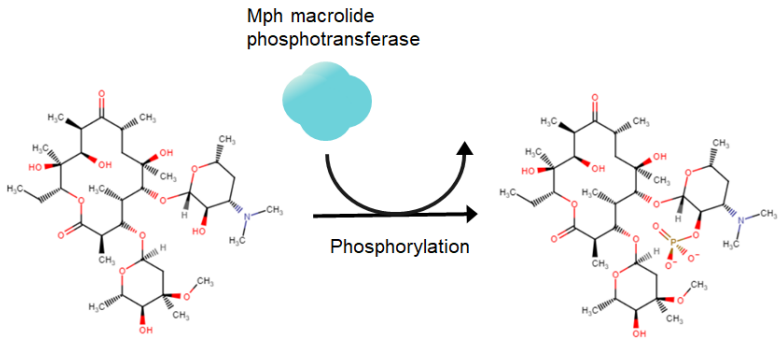


Figure 2.2: *Phosphorylation of macrolides catalyzed by Mph macrolide phosphotransferase enzymes.*

It has been shown that the functionality of Mph enzymes is dependent on the organism that harbors them. As an example of this, the gene *mph(J)*, a close homolog of the previously mentioned *mph(I)*, exists in bacteria from the *Brevibacillus* genus that are still susceptible to erythromycin. This would suggest that the acquisition of MphJ does not result in a macrolide resistant phenotype, however when heterologously expressing *mph(J)* in *E. coli* it showed the capacity to phosphorylate a wide range of macrolides [50]. In addition to showing the functionality of the MphJ enzyme, this also suggests that sequence homology is not necessarily correlated with substrate specificity.

Not much is known about how these genes have evolved or where they originate. As previously mentioned, four Mph homologs are known to have mobilized into human pathogens, including *E. coli*, *Acinetobacter baumanii* and *Klebsiella pnemoniae*, however when studying ancient strains of these bacteria it was shown that these did not carry the *mph* genes in question [26]. This suggests that a gene transfer event of these occurred from some other bacteria long ago.

On this point it is also of note that the gene *mph(C)* often appears on mobile genetic elements together with other macrolide resistance genes, notably the efflux pump *msr(A)* very often appears in its genetic vicinity. However, a cluster of genes related to antibiotic and metal resistance was identified on the chromosome of a clinical isolate of the Gram-negative bacterium *Stentrophomonas maltophilia*, which points towards the exchange of *mph(C)* between Gram-negative and Gram-positive bacteria, since the gene has also been known to appear in bacteria from the *Staphylococcus* genus [35].

As is the case for *erm* genes, *mph* genes have also been found in the environment. *mph(G)* has been reportedly found in marine bacteria in Japan [51], while *mph(A)* and *mph(B)* have been found in wastewater treatment plant isolates from Zagreb [45]. In samples from the Haihe river in China, *mph(A)* was identified as the most frequently occurring macrolide resistance gene on plasmids, and this gene was also found to often appear on plasmids together with the methyltransferase gene *erm(B)*.

This might suggest that *mph* genes have marine origins, however it is likely that these genes were promoted in these environments as a result of antibiotic selection pressure or other human intervention. Further research is needed to definitively characterize the origins from where these genes have mobilized.

2.2 Application of bioinformatics tools for antibiotic resistance research

Scientists have been aware of the problem with emerging antibiotic resistance for a long time [52]. Much research has therefore been devoted to the problem, traditionally using experimental methods to discover resistance determinants, characterize their properties [53, 54], and study their evolution [55]. While such experimental studies are still essential, and can not be replaced, today they can be supplemented by computational methods such as bioinformatics. Bioinformatics is an interdisciplinary field combining biology, mathematics and computer science to create tools that can be used to analyze biological data [56], and the insights obtained from such analysis can then be used to help design better experiments, and devise new ways of combating antibiotic resistance [57].

As a result of the decreasing costs of DNA sequencing associated with the introduction of next-generation sequencing technologies, the amount of publicly available sequencing data has seen a considerable increase over the last decade [58]. Many different software tools and approaches have been developed to process this data, however as it relates to identifying novel genes of specific functions from bacterial DNA, perhaps the most popular approach makes use of hidden Markov models [59].

2.2.1 Hidden Markov models for ARG finding

A hidden Markov model (HMM) can be explained as a model describing a probability distribution over a potentially infinite number of sequences [60]. The model consists of two stochastic processes, where one is the so called **observed process**, meaning that it produces an observable output, and the other is the underlying **hidden process** from which the model derives its name [61]. The hidden process X can not be observed, and takes the form of a finite-stage homogeneous Markov chain [62], and while it may be of higher orders this section will only concern first-order HMMs. A first-order Markov chain is defined as a random process of jumping between states in a sample space $S = \{s_1, ..., s_N\}$. The definition of Markov means that the probability of what the next state j in the chain will be, the so-called transition probability a_{ij}, is only dependent on state i that the process is currently in [63]. Transition probabilities can be denoted as

$$a_{ij} = P(X_t = s_j | X_{t-1} = s_i),\ s_i,\ s_j\ \in\ S. \tag{2.1}$$

The starting state i of the chain is determined through so-called initial probabilities

π_i [64], here denoted as

$$\pi_i = P(X_1 = s_i),\ s_i\ \in\ S. \tag{2.2}$$

The observed process Y generates output as a function of the hidden states, and in general lacks the Markov property. This means that, in any given state, the hidden process is independent of the observed process while the opposite is not true since the observed process is dependent on the current state of the hidden process and deterministically on the sequence of the observed process [62], i.e.,

$$Y_1^T : Y_t = f(X_t). \tag{2.3}$$

The probabilities that the observed process will generate an output Y_t given a state of the hidden process X_t are called emission probabilities. The collection of these probabilities is called the emission distribution [65], and can be denoted as

$$b_j(Y_t|Y_1^{t-1}) = P(Y_t|Y_1^{t-1}, X_t = j). \tag{2.4}$$

The above described probabilities comprise the parameters that need to be estimated in any given HMM before it can be used as a predictive tool. Parameters are estimated by training the model on a set of training data, which is typically the most difficult step when creating HMMs, and then evaluating the model to see whether it is a decent enough representation of reality through calculating the probability that the observed sequence was generated by the model given the estimated parameters [65]. The basics of HMM theory has been adapted for a variety of applications, and one variation of HMMs that has proved useful for gene finding is so-called profile HMMs.

2.2.1.1 Profile HMMs

This section is a summary of what was written by SR Eddy in 1998 [60], as I consider it to be a well-written, comprehensive overview of the theory behind profile HMMs, and want to adhere to established notations. Profile HMMs are a type of HMM architecture introduced by Krogh et. al in 1994, adapted for representing profiles of multiple sequence alignments [66]. Given such an alignment, each consensus column corresponds to a 'match' state wherein the distribution of residues allowed in the column is modeled. For a protein sequence alignment these residues would represent amino acids in the sequences. In order to allow for insertions between residues in the alignment, as well as deletions of consensus residues at a given column, corresponding 'insert' and 'delete' states also exist within the given state space (Figure 2.3)

Profile HMMs can be used to score a sequence within an alignment by converting the probability parameters of HMMs to additive log-odds scores. This is different from how alignments are typically scored. In traditional gapped alignments an insert of x residues is usually scored using an affine gap penalty $a+b(x-1)$, where a is the score for the first residue and b is the score of each subsequent residue in the insertion.

In a profile HMM, for an insertion of length x there is a state transition into an insert state which costs $\log(t_{MI})$ (where t_{MI} is the state transition probability for moving from the match state to the insert state), $(x-1)$ state transitions for each subsequent insert state that cost $\log(t_{II})$ and a state transition for leaving the insert state that costs $\log(t_{IM})$. This is analogous to the traditional affine gap penalties, with the open gap cost expressed as $a = \log(t_M I) + \log(t_{IM})$ and the gap extend cost expressed as $b = \log(t_{II})$.

The difference, however, is that while the affine gap penalty in traditional gapped alignment can be considered arbitrary, this is not the case with profile HMMs. As an example of why this matters, one could theoretically optimize the model parameters such that the score of a sequence in the alignment would be maximized by assigning the gap costs a value of zero. In this scenario, while the score would be high, the alignment itself would be atrocious. Contrasting this to profile HMMs, here the probability associated with transitioning from a match state to an insert state is connected to the probability to transition from one match state to another, meaning that there is actually a cost associated with a match-match transition that does not exist in traditional alignment. The gap cost can be lowered by increasing the value of t_{MI} towards 1.0, however this in turn leads to t_{MM} decreasing towards 0, and the cost for sequences with no insertion approaching negative infinity. For this reason, there is a trade-off when assigning state transition probabilities as sequences without insertion have to be balanced against sequences with insertion.

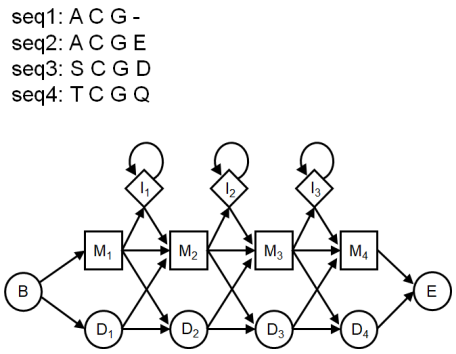


Figure 2.3: *Schematic describing the hidden process of an arbitrary multiple sequence alignment (top left). Arrows indicate transition probabilities between states in the state space. The state space comprises three types of states; match states (squares) which have a number of emission probabilities corresponding to the possible residues at that position, insertion states (diamonds) with a similar number of corresponding emission probabilities and delete states (circles) without any associated emission probabilities.*

Another difference when comparing profile HMMs to traditional alignments, is how insert residues are handled. While in traditional alignment there is no cost associated with inserts aside from the gap penalty, profile HMMs include emission probabilities in the insert states as well. Given that these insert state emission probabilities are equal to the background amino acid frequency, the score of inserted residues become $\log(f_x|f_x) = 0$, effectively similar to what is seen for traditional alignment. However, for insertions that do not share the same amino acid distribution as proteins in general, this information can be incorporated into the model through the insert state emission probabilities. As this is often the case, with insertions being more common in surface loops of protein structures, this results in profile HMMs being able to score sequences in a way that more accurately describes biological reality.

3

Methods

This chapter contains a detailed description of the methodology and workflow applied during the project. This includes the creation of profile HMMs, their subsequent application on various datasets as well as the post-processing of the output sequences yielded by the analysis.

3.1 Model creation and optimization

Amino acid sequences representing known macrolide resistance genes were downloaded from NCBI GenBank, based on Genbank IDs provided on the official Tetracycline and MLS nomenclature website [67, 34]. In total, sequences representing 38 of the 41 listed *erm* genes were downloaded, as the representative protein sequences for two of the Erm enzymes, namely ErmI and Erm37, could not be located. Furthermore, research revealed that one of the listed resistance determinants, Erm32, deviated from other Erm variants in both function and protein structure [23], and this gene was therefore excluded from all further analysis (NOTE: In February 2020 the official list of *erm* macrolide resistance determinants was updated, adding two new genes, denoted *erm(50)* and *erm(51)* [34], however for obvious reasons these were also not included in the creation of the models). Similarly, representative protein sequences for 13 of the 15 listed Mph variants were downloaded, since the sequences corresponding to MphD and MphH could not be identified from the GenBank IDs provided by the official list, and thereby these were also excluded during the creation of the models. To avoid bias when creating models, such that only the regions responsible for interaction with the macrolide

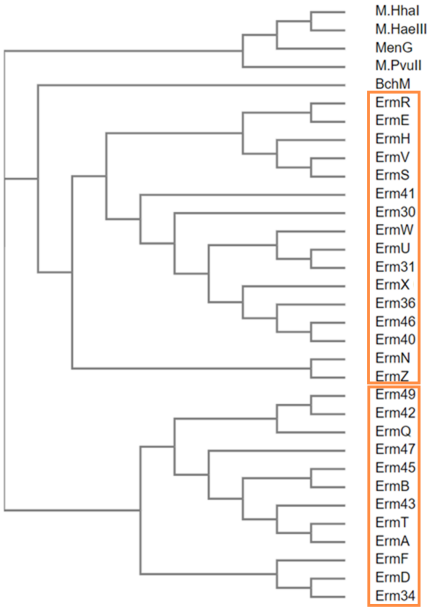
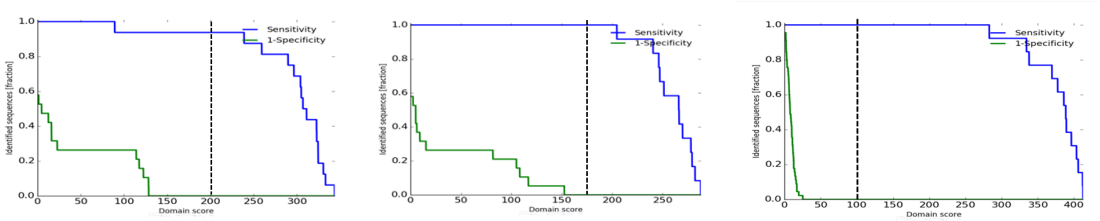


Figure 3.1: *Phylogenetic tree displaying the Erm sequences that showed an amino acid similarity < 70%. The two groups that were used to create separate models are highlighted in the tree.*

would be considered, the sequences were clustered at 70% amino acid sequence similarity using usearch with parameters '-cluster_fast', '-id 0.7' [68]. This resulted in 28 representative Erm sequences being left after clustering, while all 13 Mph enzymes showed an amino acid similarity $< 70\%$ and were therefore kept.

To investigate the similarity and evolutionary relationship of the remaining genes, multiple sequence alignment was performed using the online client of clustal omega [69]. Similar sequences without the resistance functionality were included in the alignment to act as outgroups in the resulting phylogenetic trees. From these trees a few observations were made, most importantly in the *erm*-tree where the genes appeared to cluster together in two distinct clades, with genes in one of these showing more similarity to one of the negative genes than they did to the other clade containing known *erm* genes (Fig 3.1). Taking this into consideration, it was decided to generate two separate HMMs for identifying *erm* genes. One of these would be built from the 16 protein sequences in the topmost group in Figure 3.1, this model will henceforth be referred to as the Erm group 1 model, while the other, built from the 12 protein sequences in the bottom-most group in 3.1, will be referred to as the Erm group 2 model. A similar observation was made about the *mph* genes, where three genes clustered outside of the intended outgroup in the tree, however since the number of available sequences were lower for this type of genes it was decided to only make one model representing the entire class of genes.



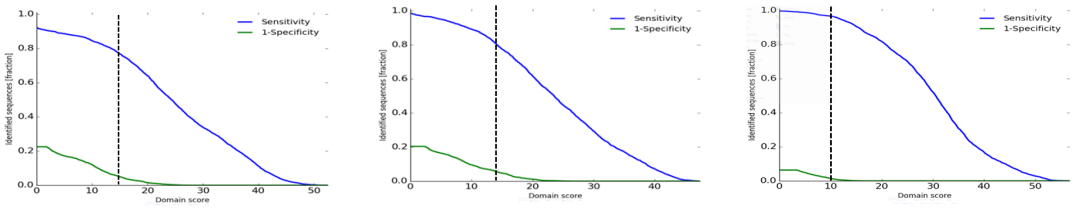
(a) Erm group 1 model **(b)** Erm group 2 model **(c)** Mph model

Figure 3.2: *The sensitivity and specificity of the three obtained HMMs as a function of threshold score, for full-length genes. The blue lines represent the sensitivity, the green lines represent 1 - specificity and the dashed black lines represent the chosen threshold scores used for analysis for each model.*

To estimate the specificity when creating the aforementioned models, sequences from the same protein superfamilies as the macrolide resistance genes were identified using NCBI's Conserved Domain Database [70], and then obtained from GenBank. In total, 19 various sequences from the AdoMet MTase superfamily to which the *erm* genes belong were obtained. This also included 5 different sequences representing the *ksgA* gene, which is suspected to have an evolutionary relationship to *erm* as described in the background section. From the superfamily to which the *mph* genes belong, 45 sequences representing homoserine kinase II were obtained. These negative sequences were used as input, together with the positive sequences for each specific model, to create profile HMMs for the three genetypes using the command 'fargene_model_creation' from the fARGene package [21]. The created models were first evaluated based on their sensitivity and specificity for full-length genes, which

were determined through leave one out cross-validation and misclassification of the confirmed negative sequences respectively.

It was observed that the Mph model displayed an overall better performance in terms of both sensitivity and specificity than either of the Erm models (Figure 3.2). This allowed for a lower threshold score to be set when using this model for analysis where these parameters would both still retain as high a value as possible, and it was decided to set the threshold at a score of 100 for the Mph model. As can be seen from Figure 3.2a, there was no score that allowed for both a sensitivity and a specificity of 1 for the Erm group 1 model. It was decided that a threshold score of 200 would be used, which would sacrifice some sensitivity but retain a high specificity. Finally, for the Erm group 2 model it was decided to use a threshold score of 175, which was within the narrow range where both parameters were evaluated to be optimal as is displayed in Figure 3.2b. The models were also evaluated in the same way based on their ability to correctly classify 33 AA long fragments from resistance genes, while not misclassifying fragments of equal length from confirmed negative genes. These fragments are supposed to represent metagenomic data, which is fragmented in its nature.



(a) Erm group 1 model **(b)** Erm group 2 model **(c)** Mph model

Figure 3.3: *The sensitivity and specificity of the three obtained HMMs as a function of threshold score, for 33 AA long gene fragments. The blue lines represent the sensitivity, the green lines represent 1 - specificity and the dashed black lines represent the chosen threshold scores used for analysis for each model.*

The Mph model once again proved to have superior performance to the other two models (Figure 3.3). After considering the results of the evaluation of all models, the threshold scores chosen to use for classification of metagenomic fragments were 10 for the Mph model, 15 for the Erm group 1 model and 14 for the Erm group 2 model. Once the evaluation of the models was complete, they were used to search both genomic and metagenomic data for the presence of macrolide resistance genes.

3.2 fARGene analysis and post-processing

All bacterial genomes present in NCBI GenBank in October of 2019 were downloaded [67]. This included 15,438 complete genomes and 412,095 draft genomes, which were analyzed separately. The analysis of this data was performed using fARGene, applying the newly constructed HMMs with threshold scores for full-length

as specified above (Figure 3.2), but otherwise default parameters, which resulted in a set of predicted sequences corresponding to each analyzed genetype.

To extract meaningful information from the predicted ARG sequences, a number of post-processing steps were performed (Figure 3.4). For the first analysis, the predicted protein sequences of the two genetypes were aligned separately using mafft v7.23 [71] with default parameters, and a phylogenetic tree representing each macrolide ARG was generated from the alignments using FastTree v2.1.10 [72], again using default parameters. At this point the sequences found by the two Erm models were analyzed together, and for the tree representing these genes one sequence representing the methyltransferase *ksgA* was used as outgroup due to the suspected evolutionary relationship between these genes. Correspondingly, three sequences representing aminoglycoside phosphotransferases were used as outgroup in the *mph*-tree based on a publication by Pawlowski et. al, where these had been used for the same purpose [26].

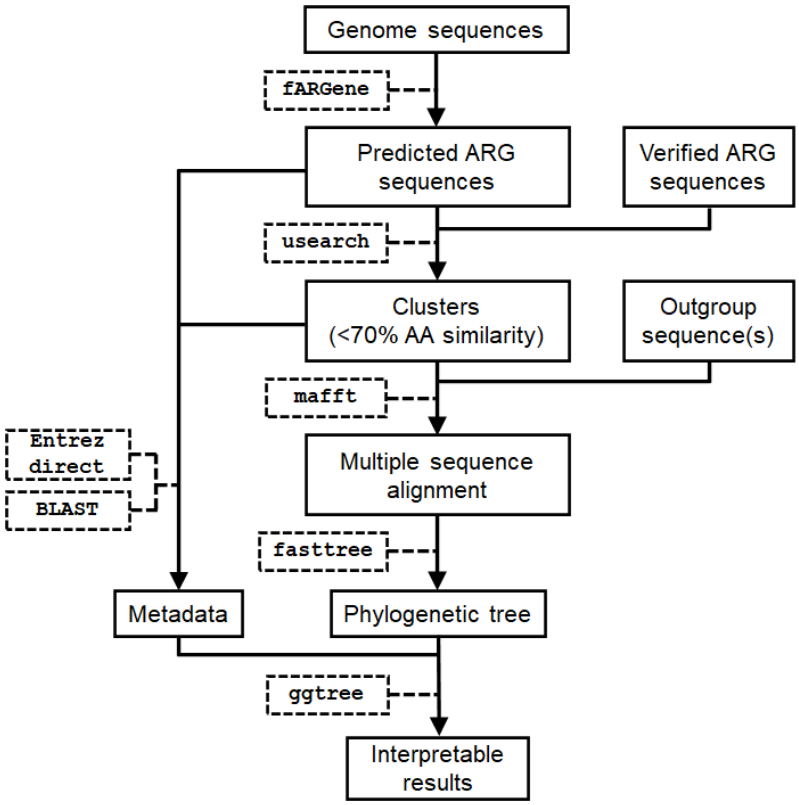


Figure 3.4: *Graphic representation of the workflow used when analyzing and post-processing genomic data. Boxes with solid borders represent various files, boxes with dashed borders represent external software used during each part of the analysis.*

The trees were visualized using ggtree v2.0.1 [73, 74], with additional information about the leaves being included to improve the interpretability of the trees. To visualize the distribution of previously known ARGs in the trees, a BLAST search was performed for each protein sequence against the reference sequences using blast+ v2.6.0 [75]. The best hits from the BLAST searches were extracted, and if they showed $\leq$ 79% amino acid similarity to any known gene they were considered as

the same gene based on the principles of MLS gene nomenclature [76]. Additionally, metadata about the source from which the bacteria that correspond to the genomes in GenBank had been isolated was retrieved using Entrez direct v13.3 [77]. This information, for the genomes where it was available, was included in the final tree alongside information about the phylum to which the bacteria that harbored each ARG belonged. To investigate whether any specific type of bacteria were enriched for the presence of the analyzed macrolide ARGs, phylum enrichment analysis was performed by counting the number of unique species harboring a macrolide ARG and comparing that number to the total number of species from that phylum in the database using Fisher's exact test.

Next, the novel gene families that were identified by the fARGene analysis were studied. To this end, the predicted protein sequences were clustered together with the verified ARG sequences of the corresponding type at 70% amino acid similarity using usearch v8.0.1445 with parameters '-cluster_fast', '-id 0.7' [68]. Information about the contents of each cluster was gathered, including which organisms the sequences corresponded to, and if one or more of the previously known sequences clustered together with sequences identified in the data, the family was considered as known, otherwise the gene family was considered novel. After clustering, phylogenetic trees were created from the representative centroid sequences of each cluster using a similar methodology as described above, with the exception that information about where the bacteria were isolated from was excluded, and the addition that the trees were rerooted at the outgroup using the Interactive Tree of Life web client [78].

Based on the results from the genome analysis, metagenomic datasets that were suspected to contain macrolide ARGs were selected. This included metagenomes from the Human Microbiome Project (HMP) [79], gut microbiomes of patients with Parkinson's disease [80] and type 2 diabetes [81], pig gut microbiomes [82], wastewater treatment plants (WWTP) in Sweden [83], the antibiotic-polluted Pune river in India [84], marine environments from the Tara oceans expedition [85], oil contaminated bacterial communities in marine sediments [86], antarctic soil [87], soil from forests in China [88], feces from wild baboons[89] (as it has been noted that baboons have a similar gut microbiome composition to humans [90]), gut microbiomes of rhinos [91], lake Hazen in northern Canada [92], and a river in the Amazon rainforest [93]. These were all analyzed using fARGene, with the threshold scores of the profile HMMs set as previously specified (Figure 3.3, Figure 3.2), and the sequences of the identified ARGs were post-processed in the same way as the sequences found in the genomic data.

4

Results

In this chapter, all of the results from the study are presented. Section 4.1 contains the results relating to the analysis of genome sequences, while section 4.2 contains the results obtained from analyzing the metagenomic datasets. Each section is divided into subsections where the results relating to the two genetypes are presented separately, and mainly contains figures and tables, as well as the information required to interpret them. The interpretations and discussions of the presented information will be addressed in the following chapter.

4.1 Analysis of bacterial genomes

After analyzing the genomes in NCBI GenBank, the results revealed that both *erm* and *mph* type macrolide ARGs are widely present in many different types of bacteria. Genes of type *erm* were found across more species (875) than genes of type *mph* (568), and in addition more different *erm* gene families (< 70% AA similarity) were identified compared to *mph*. However, more individual instances of *mph* genes were found, with 3.19% of all genomes in GenBank harboring an *mph* gene as compared to 2.77% of all genomes in GenBank.

Table 4.1: *Phylum distributions of the predicted genes from the two gene types. Genes refer to the individual occurrences of ARGs found within the data, while classes refer to clusters of < 70% AA identity.*

	Erm		**Mph**	
Dataset	**Genes**	**Families[a,b]**	**Genes**	**Families[a,b]**
NCBI RefSeq	330	10/21	1107	13/59
NCBI Assembly	12423	27/316	14033	14/210
Total:	**12753**	**28/320[c]**	**15140**	**14/221[c]**

a AA similarity < 70%
b Known/new
c Non-redundant

Interestingly, the number of different new gene families discovered greatly outnumbered the number of families that have been characterized for both *erm* and *mph* type genes to date (Table 4.1). While this indicates the presence of a vast and

uncharacterized macrolide resistome, the majority of the predicted unknown families contained only a few predicted sequences originating from very similar, non-pathogenic, organisms. This can be compared to some of the gene families that are known to be widespread, which in some instances were identified in thousands of different, often pathogenic, genomes. Most likely this means that these newly identified genes have not mobilized, and thereby that they currently are not of clinical interest.

When comparing the species of bacteria that harbored the two different genetypes, distinct differences were found. Looking towards *erm* methyltransferase-genes, the number of genes found in bacteria from the Firmicutes phylum vastly outnumbered the number of genes found bacteria from the Proteobacteria, Actinobacteria or Bacteroidetes phyla, which all harbored a number of genes comparable to each other. By contrast, *mph* phosphotransferase-genes were found to be most widespread in bacteria from the Proteobacteria phylum, with Firmicutes also containing a large number of these genes while the number of *mph* genes found in both Actinobacteria and Bacteroidetes proved to be comparatively very small (Table 4.2). It is also of note that while a very small number of *mph*-genes were identified in bacteria assigned to any phylum other than the four previously mentioned, this number was substantially larger in the case of *erm*-genes. In addition, a larger number of genomes not assigned to any phylum in the database was shown to harbor *mph* genes compared to the corresponding number for *erm* genes.

Table 4.2: *Phylum distributions of the predicted genes from the two gene types.*

	Erm		**Mph**	
Phylum	**Genes**	**Families[a]**	**Genes**	**Families[a]**
Firmicutes	9535	102	4377	111
Proteobacteria	711	5	8211	16
Actinobacteria	877	135	280	67
Bacteroidetes	855	8	36	15
Miscellaneous	243	50	34	11
NA	532	50	2202	15
Total:	**12753**	**348[b]**	**15140**	**235[b]**

a AA similarity $< 70\%$
b Non-redundant

Regarding the diversity of sequences, Table 4.2 shows a few interesting observations. First, with respect to the amount of ARGs found within Proteobacteria, the diversity among these is surprisingly small for both gene types. Aside from this, the diversity of identified sequences within a given phylum with respect to the number of predicted sequences is in general larger for *mph* type genes.

4.1.1 Erm 23S rRNA methyltransferases

To account for the number and diversity of the analyzed genomes, phylum enrichment analysis of known and new Erm sequences was performed for species be-

longing the four most abundant phyla using Fisher's exact test (Figure 4.1). The over- or underrepresentation of a specific phylum can be determined by whether the calculated odds ratio is above or below one, given that the test was significant. Interestingly, when comparing the odds ratios of known and new genes, there was a distinct difference between these. For known Erm variants, Firmicutes and Bacteroidetes displayed a large (ratio=4.56, p-value$<10^{-15}$) and moderate (ratio=1.98, p-value=6.26×10^{-8}) over-representation respectively, while Proteobacteria were significantly underrepresented (ratio=0.109, p-value$<10^{-15}$). This was contrasted against the previously uncharacterized Erm variants, where Actinobacteria were the most significantly overrepresented phylum (ratio=4.74, p-value$<10^{-15}$), with Firmicutes being moderately overrepresented (ratio=2.08, p-value=2.74×10^{-11}) and Proteobacteria being heavily underrepresented (ratio=0.024, p-value$<10^{-15}$). Two tests proved to not be significant, these corresponded to Actinobacteria for known genes and Bacteroidetes for new genes.

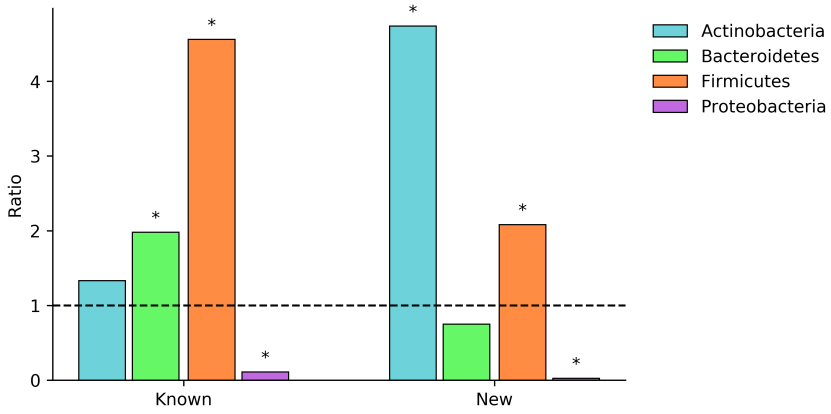


Figure 4.1: *Phylum analysis of species containing known and new Erm sequences identified from NCBI GenBank. The odds ratios where calculated using Fisher's exact test, and a star above the bar denotes whether the corresponding test was significant using a p-value cut-off of 0.001.*

To get a measure of the distribution of known Erm classes, and ratio of known to new variants identified in the data, all predicted sequences were compared against the reference sequences of all known variants using BLAST. From the results, it was shown that a majority of the identified sequences corresponded to three known genes (Figure 4.2). Out of these, ErmB was shown to be the most widely present one, followed by ErmA and ErmC. Afterwards, the sequences were clustered into gene families sharing 70% sequence identity, the similarity being intentionally set much stricter than what is necessary for a gene to be classified as new to avoid problems with accidentally misclassifying known genes as novel, and to avoid having several families representing variants of the same gene. This did, however, also mean that in some cases several known genes clustered into the same family as their similarity was higher than 70%. In total, 28 known families were formed representing 34 out of the 42 known *erm* genes, while 341 gene families representing previously uncharacterized Erm variants were formed.

It is important to recognize that the results yielded from the Erm models are not

without a certain degree of false positives. This is indicated in the phylogenetic tree formed from the aforementioned gene families (Figure 4.3), as a number of unlabeled sequences cluster together with the outgroup (KsgA). Upon investigation, it was revealed that these gene families indeed represented KsgA variants rather than novel Erm classes, and the close homology between these genetypes resulted in some misclassification of these sequences. Thankfully, these false positives were easily identified from their location in the tree, and no other sequences showed any indication of being incorrectly classified. Accordingly they have not been considered in any of the results presented in this chapter, with the exception of the phylogenetic Erm trees, and will not be further mentioned.

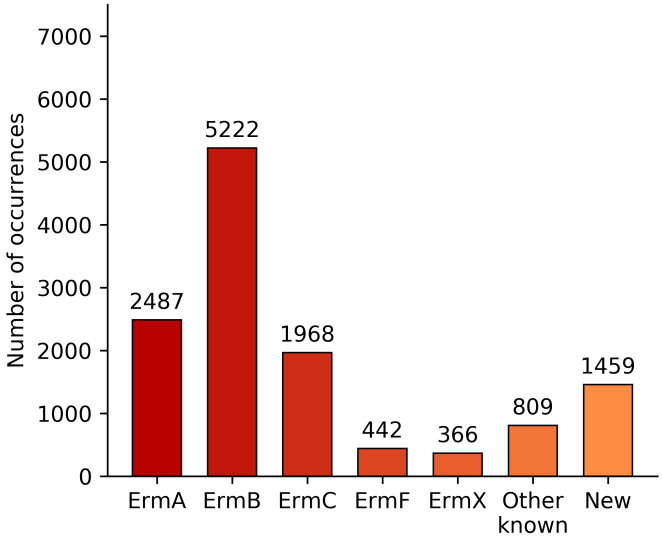


Figure 4.2: *Distribution of known Erm sequences identified in the genomic data. The five most frequently identified variants are displayed as individual bars, all sequences that showed >79% AA similarity to any known Erm variant other than these five are compiled into the bar titled 'Other known', while all sequences that showed <79% AA similarity are compiled into the bar titled 'New'.*

When studying the evolutionary relationships of the predicted gene families from the phylogenetic tree (Figure 4.3), a number of interesting observations can be made. The tree is largely structured based on what organisms the sequences originate from. Largely, the tree can be seen as consisting of eight groups, where four groups, separated from each other in pairs, correspond to genes mostly originating from bacteria of the Firmicutes phylum, two adjacent groups where the majority of the sequences were identified in bacteria from the Actinobacteria phylum, one small group containing gene families from the Bacteroidetes phylum and one lager group that is mostly represented by sequences from other, more exotic phyla. It is of note that the previously known Erm classes are well spread out throughout the tree, speaking to the overall diversity of this type of genes. It should also be noted that there at several locations in the tree are indications of horizontal gene transfer events having occurred, as indicated both by the black tips, which represent either gene families that were shown to contain sequences identified in bacteria from more

than one phylum, gene families containing sequences identified on plasmids, as well as previously known variants that are known to be mobile. In addition, there are instances where sequences originating from bacteria belonging to a certain phylum appear in parts of the tree where the majority of sequences come from another phylum, also indicating mobility.

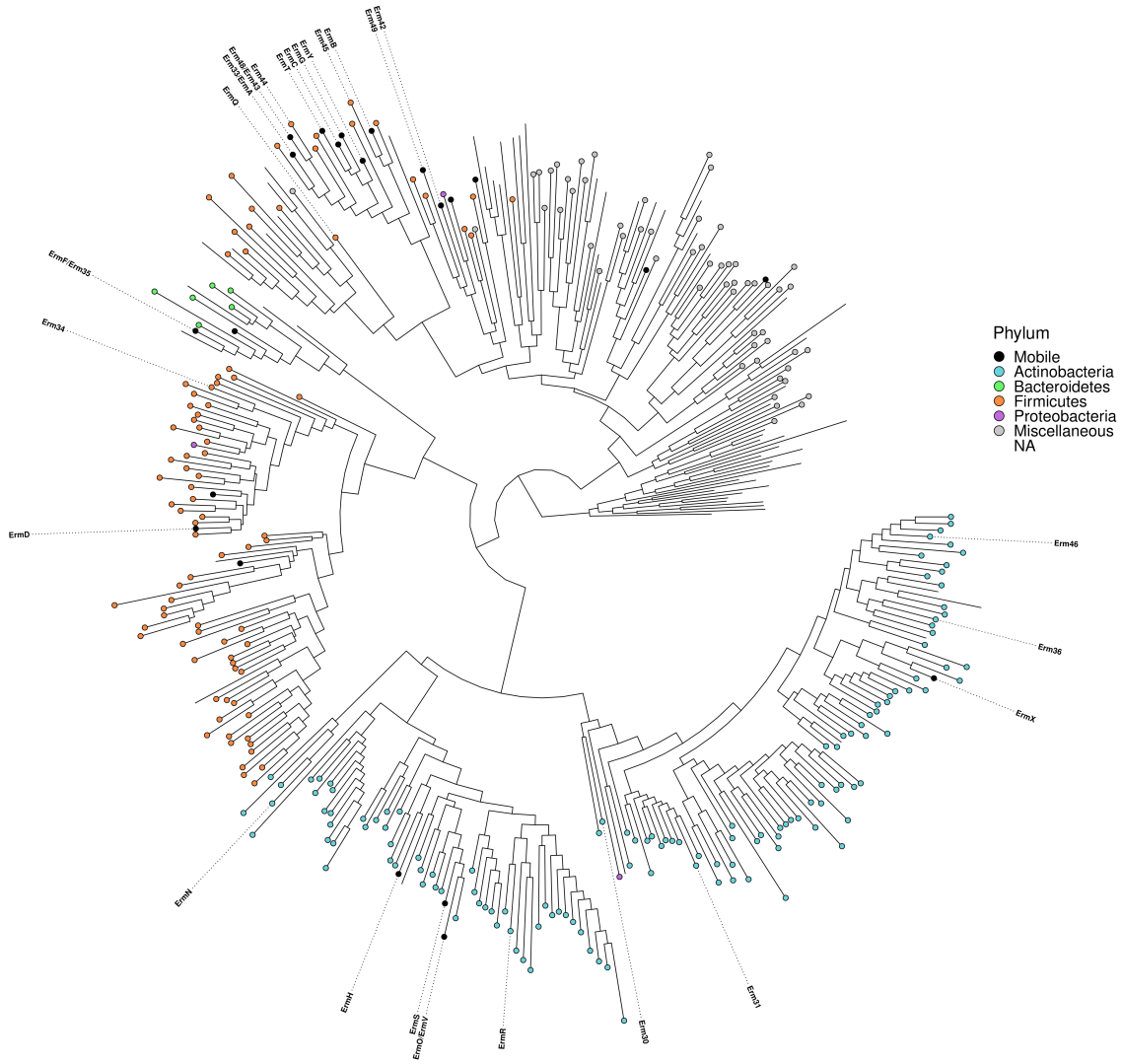


Figure 4.3: *Phylogenetic tree displaying representative centroid sequences obtained by clustering the predicted Erm sequences at 70% similarity. The color of the tips represent the phylum of the organism from which the representative sequence originates. All known Erm enzymes that were found in the data have been placed out in the tree.*

When studying the clustered ARG families, some information gets lost as only the representative centroid sequence for each cluster gets considered. As interesting metadata regarding where the bacteria corresponding to many genomes in GenBank had been isolated from was available, it was decided to generate a cladogram from the complete set of predicted Erm sequences and include this information as a map around the tree (Figure A.1, Appendix A).

The majority of the known *erm* genes identified in the database originated from bacteria that seemingly had been isolated from humans or animals. In some cases it was explicitly stated that the bacteria were isolated from the clinical environment. The novel genes, by contrast, showed a greater tendency towards being harbored by bacteria that had been isolated from various environmental samples, though there exceptions to this, notably a Bacteroidetes-associated cluster of novel genes where all members of the cluster were hosted by bacteria of either human or animal origin. It should be noted that this information was not available for a substantial amount of the analyzed genomes, and furthermore sometimes the information was very vague even if it did exist.

4.1.2 Mph macrolide phosphotransferases

Looking towards Mph macrolide phosphotransferases, the same analysis pipeline as previously described was employed. Phylum enrichment analysis of known and new Mph sequences again revealed distinct differences between the types of bacteria harboring these. Firmicutes were significantly overrepresented among the hosts of known *mph* genes (ratio=4.35, p-value$<10^{-15}$), while Actinobacteria and Bacteroidetes were significantly underrepresented (ratio=0.14, p-value=2.55×10^{-12}, ratio=0.21, p-value=2.28×10^{-6} resp.) (Figure 4.1). By contrast, Actinobacteria were significantly overrepresented among the hosts of previously uncharacterized *mph* genes (ratio=2.35, p-value=1.79×10^{-12}), along with Firmicutes (ratio=3.68, p-value$<10^{-15}$), while Proteobacteria were significantly underrepresented (ratio=0.14, p-value$<10^{-15}$). The tests that did not prove significant related to Bacteroidetes for new Mph sequences and Proteobacteria for known Mph variants.

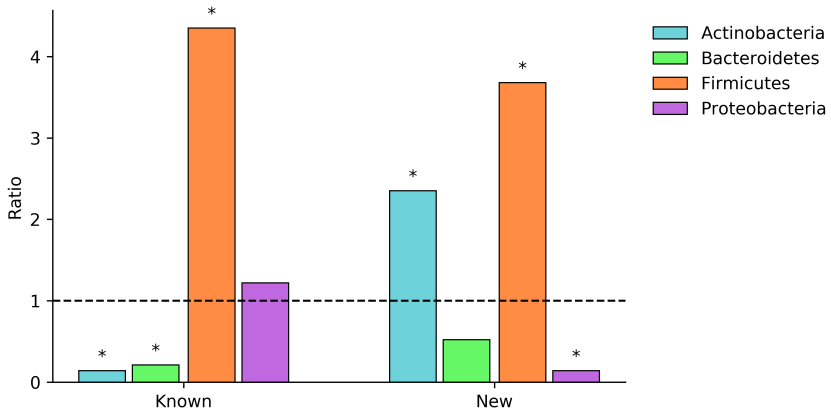


Figure 4.4: *Phylum analysis of species containing known and new identified Mph sequences from NCBI GenBank. The odds ratios where calculated using Fisher's exact test, and a star above the bar denotes whether the corresponding test was significant using a p-value cut-off of 0.001.*

Similarly to what was observed for the Erm type genes, a substantial majority of the identified Mph sequences corresponded to previously known genes (Figure 4.5). In particular, close to half of the identified genes corresponded to a single

gene, MphA, with two other known variants, MphE and MphC, also being widely occurring. In contrast to the Erm genes, where about a fourth of the known genes were not identified in the analyzed genomes, all 14 of the known Mph genes (with the exception of MphD, for which no reference sequence was ever located) were identified. After clustering the sequences into gene families with <70% sequence identity, 14 families representing the known Mph variants were obtained (in contrast to the Erm sequences no family contained multiple known sequences) along with 221 families representing previously uncharacterized Mph variants.

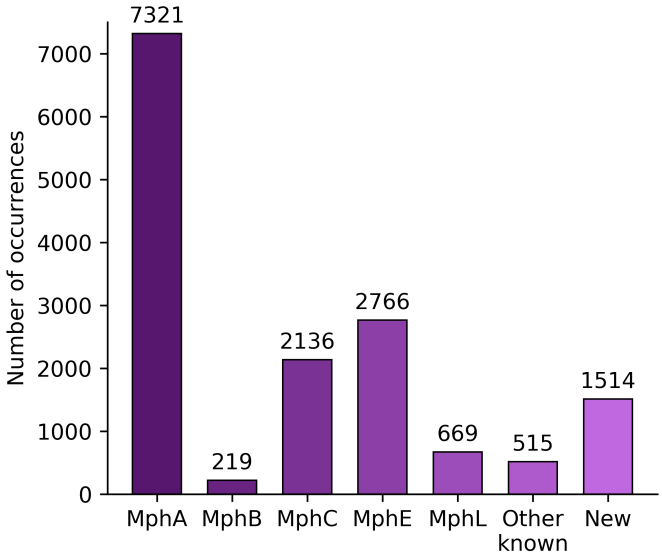


Figure 4.5: *Distribution of known Mph sequences identified in the genomic data. The five most frequently identified variants are displayed as individual bars, all sequences that showed >79% AA similarity to any known Mph variant other than these five are compiled into the bar titled 'Other known', while all sequences that showed <79% AA similarity are compiled into the bar titled 'New'.*

The phylogenetic tree describing the evolutionary relationships of these gene families is displayed below (Figure 4.6). The tree is overall similarly structured to the tree displayed in Figure 4.3, with genes identified in similar organisms clustering together. All parts of the tree contain known Mph classes, with the only notable exception being the clade at the top right of the tree, containing sequences identified in Actinobacteria. As this group of sequences clusters outside of the main tree in addition to containing no known Mph variants this could be an indication of these sequences being false positives, however unlike the previous case a further investigation of these sequences did not give any indication about this being the case. There is again multiple indications of horizontal gene transfer having occurred some time in the past, though compared to Figure 4.3 this tree contains fewer such indications.

As was the case with the previous genetype, it was also of interest to study the complete set of predicted Mph sequences. Specifically it was still of interest to analyze any trends in the isolation sources of the bacteria harboring these genes, as that could provide a basis for metagenomic analysis. To achieve this, a cladogram

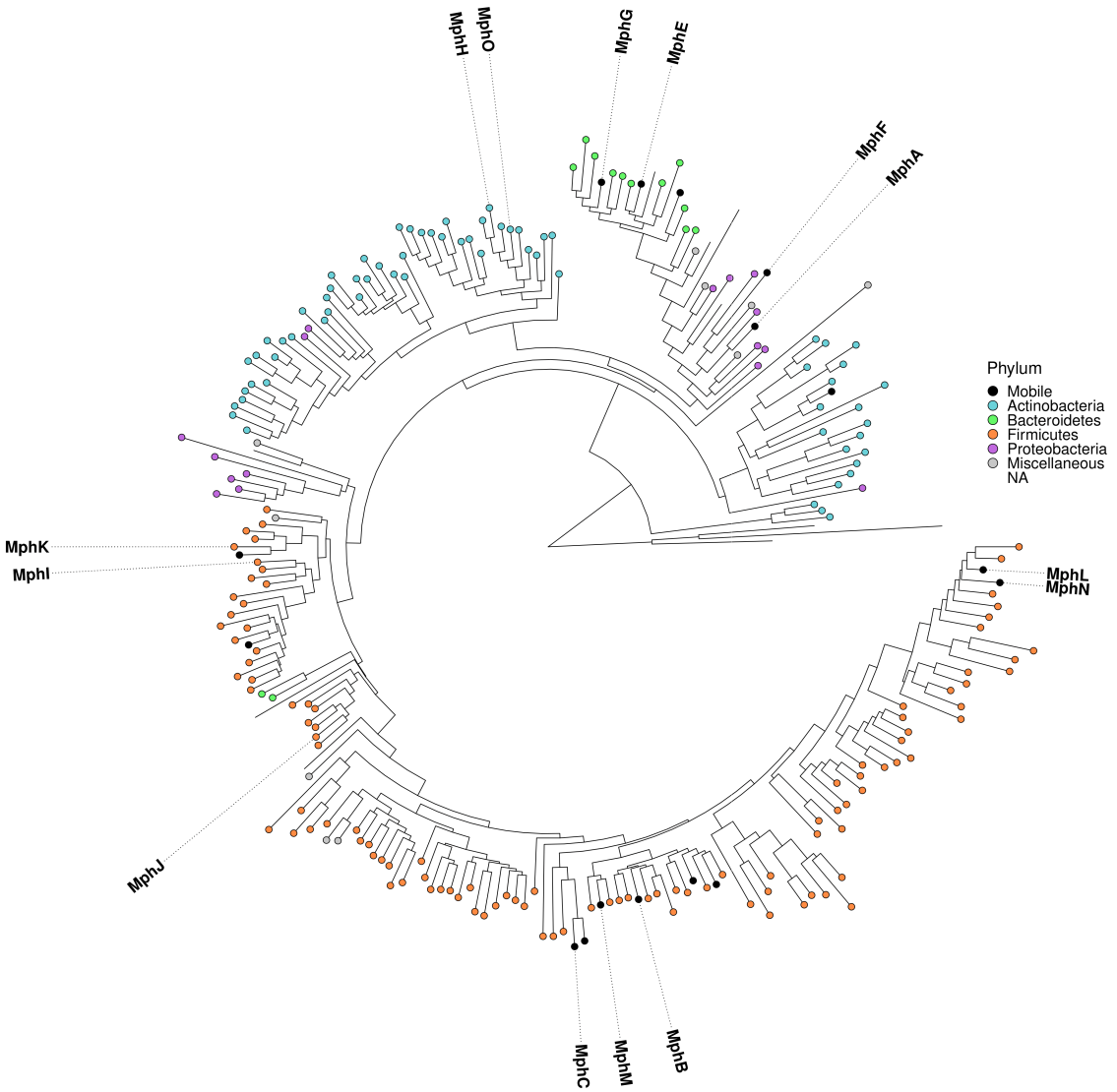


Figure 4.6: *Phylogenetic tree displaying representative centroid sequences obtained by clustering the predicted Mph sequences at 70% similarity. The color of the tips represent the phylum of the organism from which the representative sequence originates. All known Mph enzymes that were found in the data have been placed out in the tree.*

constructed from all of the predicted sequences, where the isolation source was added as a map around the tree (Figure A.2).

Similar to the bacteria harboring *erm* genes, the majority of the bacteria harboring known *mph* genes were isolated from human or animal microbiomes. Here, there seemed to be a variety among the bacteria containing novel genes however, with some being isolated from environmental microbiomes, some being isolated from humans or animals and some being isolated from food. There was no clear pattern among these, though the majority seemed to be environmentally associated. Taking these results into consideration together with what was previously observed for the predicted *erm* genes, it was clear that both genetypes were present in various types of samples,

including human, animal and environmental, and therefore a wide range of different types of metagenomes were chosen for the next step in the analysis.

4.2 Analysis of metagenomic data

To identify environments that serve as a reservoir from which macrolide ARGs can mobilize, metagenomic datasets were searched for the presence of these genes. After analyzing more than 14 terabases of metagenomic data, it was found that *erm* genes were much more prevalent than *mph* genes (Table 4.3). Notably, the human-associated (HMP, Human gut 1, Human gut 2) and human-adjacent (Pig gut, WWTP, Pune river) metagenomes contained the majority of the identified genes of both type *erm* and *mph*, with both genetypes being most prevalent in the Pig gut metagenome. Comparatively, the environmental metagenomes contained almost no macrolide ARGs, and this was also true for wild animal microbiomes. Two notable exceptions to this were the forest soil metagenome, which contained a number of *mph* genes that was comparable to the number found in the human-associated microbiomes, and the samples from lake Hazen, which curiously contained a relatively large number of uncharacterized *erm* genes.

Table 4.3: *Number of macrolide ARGs identified in each analyzed metagenome.*

		Erm		**Mph**		
Metagenome	**Size (nt)**	**Genes**	**Families[a,b]**	**Genes**	**Families[a,b]**	**Ref**
HMP	4.69×10^{12}	82	6/8	8	1/1	[79]
Human gut 1	1.93×10^{11}	15	4/7	2	1/1	[80]
Human gut 2	1.32×10^{12}	14	6/4	2	1/1	[81]
Pig gut	1.74×10^{12}	145	8/11	17	1/0	[82]
Baboon gut	1.37×10^{11}	0	0/0	0	0/0	[89]
Rhino gut	6.21×10^{10}	0	0/0	0	0/0	[91]
WWTP	4.82×10^{11}	49	6/35	8	4/4	[83]
Pune river	3.91×10^{11}	45	6/33	13	4/7	[84]
Tara oceans	4.89×10^{12}	2	0/2	1	0/1	[85]
Antarctic soil	6.25×10^{9}	0	0/0	0	0/0	[87]
Forest soil	1.99×10^{11}	6	1/5	6	3/2	[88]
Oilspill	2.75×10^{11}	0	0/0	0	0/0	[86]
Lake Hazen	2.75×10^{11}	32	0/21	0	0/0	[92]
Amazon river	2.88×10^{11}	0	0/0	0	0/0	[93]
Total	$\mathbf{14.38 \times 10^{12}}$	**389**	**8/100[c]**	**57**	**9/13[c]**	

a AA similarity < 70%
b Known/new
c Non-redundant

4.2.1 Erm 23S-rRNA methyltransferases

When analyzing the number of Erm sequences that were assembled from the various metagenomic datasets (Figure 4.7), a clear pattern could be seen. In general, the number of predicted genes per gigabase was higher in environments where antibiotics may be present, with the highest density of known genes being observed in the pig gut microbiota and the highest density of new *erm* genes being found in samples from the Pune river in India. As previously noted, the lake Hazen metagenome proved to be an interesting deviation from the trend, as it harbored the third highest density of new *erm* genes, while simultaneously containing no known variants. This is interesting as the only other metagenome that harbored new *erm* genes without any known genes being found within the same dataset was the Tara oceans metagenomic dataset, however since only two separate occurrences being found within this data, which also was the single largest dataset analyzed, containing samples from a multitude of locations [85], this does not seem comparable.

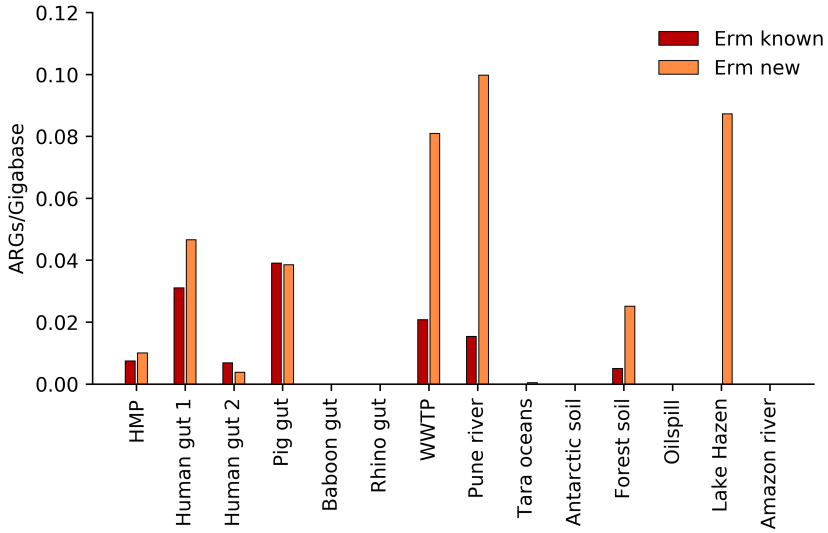


Figure 4.7: *Number of predicted erm genes per gigabase in all of the analyzed metagenomic datasets*

Far from all known Erm variants were identified in the analyzed metagenomic data, however it is noteworthy that all but one of the known variants that were found were present in the pig gut environment. A total number of 11 known Erm variants were identified; ErmA, ErmB, ErmC, ErmF, ErmG, ErmQ, ErmT, ErmX, Erm47, Erm49 and Erm50. Some of the most prevalent classes included ones that have known clinical significance, such as ErmA and ErmB, and variants that do not have such significance such as ErmQ. To study the relationships between the sequences identified from different environments, a phylogenetic tree was generated from all of the reconstructed Erm sequences together with reference sequences for known variants (Figure 4.8). Mostly, sequences identified from the same environment naturally tended to cluster together, however it was notable that in some cases very similar, or even identical, uncharacterized Erm sequences were reconstructed from different datasets.

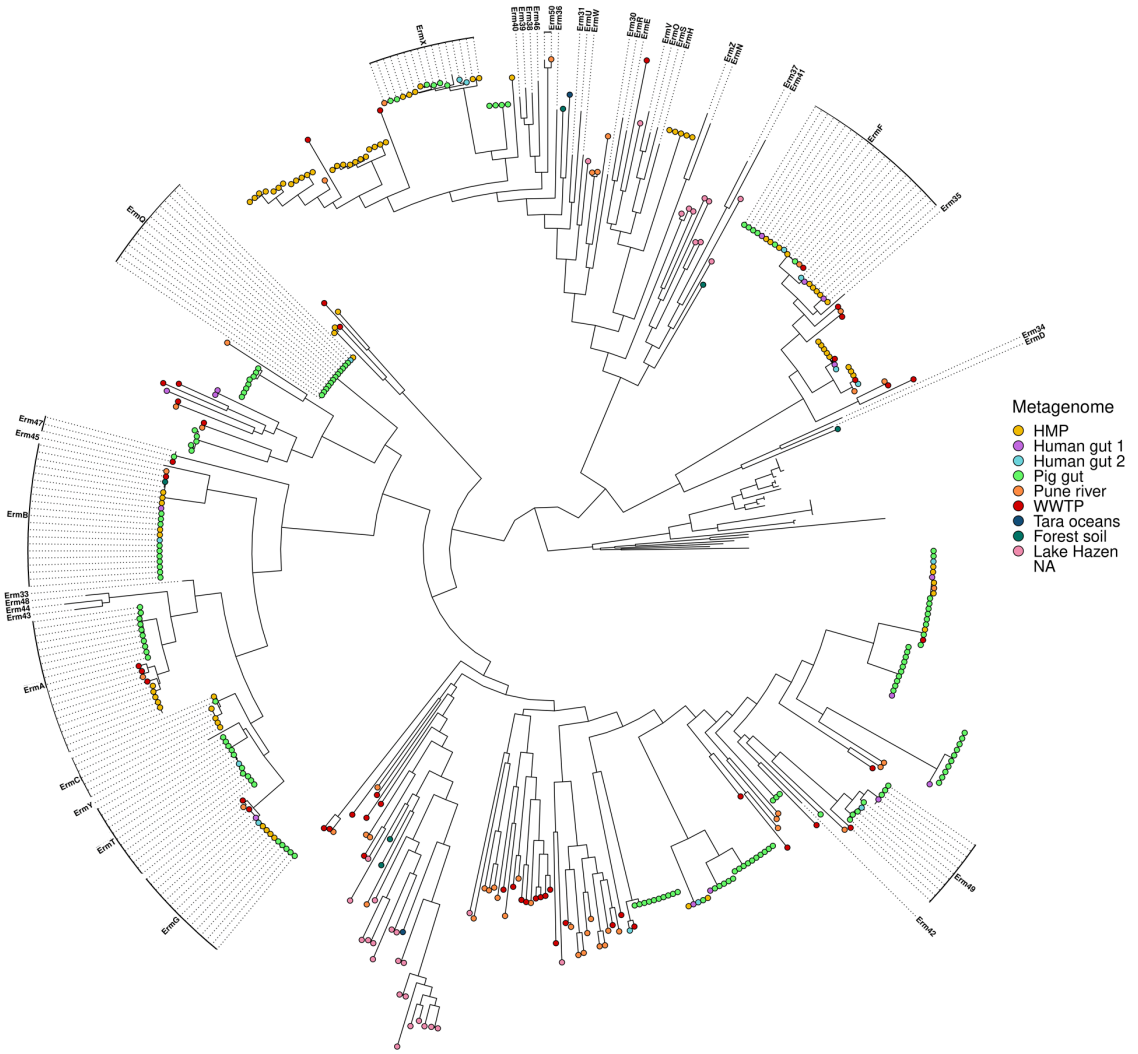


Figure 4.8: *Phylogenetic tree displaying the Mph sequences found in the metagenomic data. To improve the interpretability of the tree, it has been spiked with reference sequences of known Mph classes.*

4.2.2 Mph macrolide phosphotransferases

When observing the number of assembled Mph sequences per gigabase of metagenomic DNA (Figure 4.9), the results proved to be both similar to, and different from the corresponding results from the Erm sequences. Similarly, the environments from wastewater treatment plants (WWTP) and the Pune river were among the densest environmnets with respect to Mph sequences, with the Pune river metagenome being the densest when considering previously uncharacterized Mph sequences. Where the *mph* genes differed was considering the previously known genes, as the highest number of assembled known *mph* genes per gigabase was found in the forest soil metagenome, sampled from pristine forests in China.

As noted from Table 4.3, the identified Mph sequences could be divided into a total of 22 gene families sharing <70% AA similarity. 9 of these represented previously

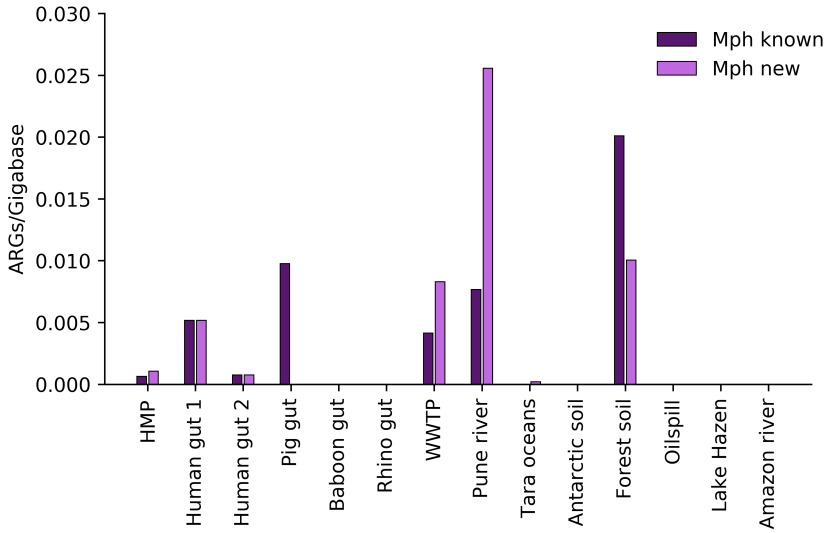


Figure 4.9: *Number of predicted mph genes per gigabase in all of the analyzed metagenomic datasets*

known Mph variants, namely MphA, MphB, MphC, MphE, MphF, MphG, MphI, MphL and MphM. To study the evolutionary relationships between the assembled genes, a phylogenetic tree was created from all of the reconstructed sequences together with all of the reference Mph sequences (Figure 4.10). The tree presents a number of interesting observations, not the least of which being that MphB, which was the known gene with the most occurrences across the data, was almost exclusively identified in the pig gut environment.

As the tree is not composed of clusters of identified genes, but rather the raw output sequences from fARGene, the similarity among the sequences can be studied. The more level the leaves that cluster together are, the more similar the assembled sequences are, and if the points are located next to each other on a line that means that they are identical. While if this occurs within the same datasets it can be attributed to a number of factors, it can be observed that here many identical sequences have been predicted across different datasets as well. Most notably five identical sequences have been identified in both the WWTP and the Pune river samples, all but one representing new gene families. For the identical pair representing MphE, these were also identical to the reference sequence for this gene familiy. Another point of note in the tree is the set of 7 identical sequences at the top left, representing a new gene family, as these were identified among all three human gut microbiomes.

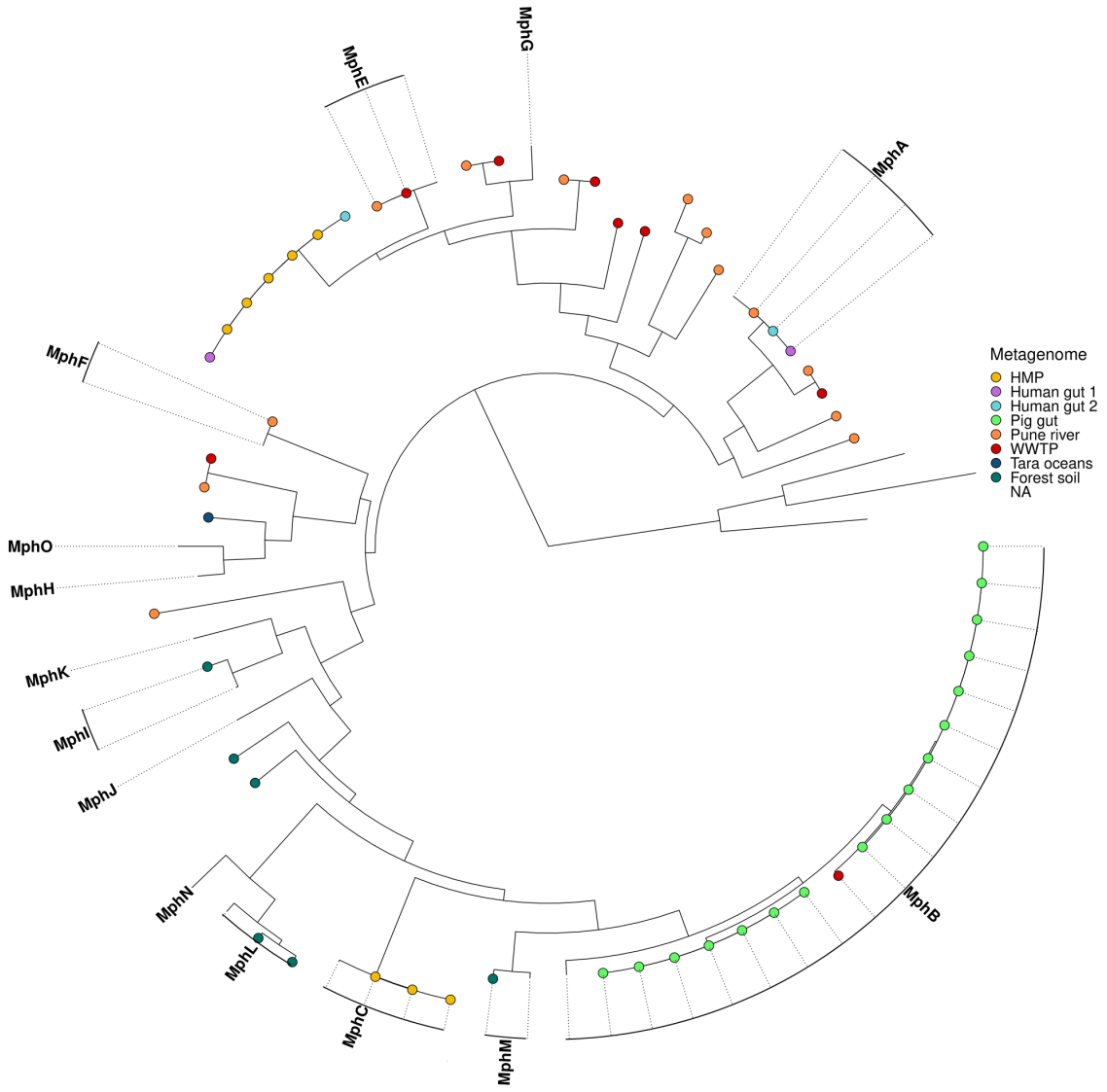


Figure 4.10: *Phylogenetic tree displaying the Mph sequences found in the metagenomic data. To improve the interpretability of the tree, it has been spiked with reference sequences of known Mph classes.*

5

Discussion

In this chapter the results described in the previous chapter will be further interpreted and discussed with respect to reliability and implications. To make the chapter less confusing, the two genetypes will be discussed separately, including the findings from both the genomic and metagenomic data.

5.1 Erm 23S-rRNA methyltransferases

The distribution of Erm sequences identified from the genomes in NCBI GenBank seems to be largely consistent with what has been reported previously. Indeed, *erm(B)* has been described as the most widespread gene of this type, and *erm(A)*, *erm(C)* and *erm(F)* are all known to have mobilized into pathogens and would therefore be expected to be more widespread than many other variants. Looking towards the types of bacteria that these were identified in, ErmA, ErmB and ErmC were all predominantly identified in bacteria from the Firmicutes phylum, with ErmA and ErmC primarily being found among the *Staphylococcus* genus and ErmB being found among the *Streptococcus*/*Enterococcus* genera. These results are exactly in line with what has been stated about these genes in the literature, and accordingly *erm(F)* was also primarily identified where it would be expected, namely in bacteria from the Bacteroidetes phylum, specifically among the *Bacteroides* genus. As these findings seemingly correspond so well to reality, it speaks to the reliability of the remaining results, with the exception of obvious false positives.

The fact that the models missclassified some *ksgA*-genes can likely be explained by the high similarity and close evolutionary relationship between these two types of genes. When examining the scores assigned to the false positives it could be seen that they were of a comparable magnitude to those assigned to some of the verified known Erm variants, and thereby the score threshold could not be raised to completely remove the false positives without simultaneously misclassifying true positive Erm variants as negatives. By contrast, *mph* genes are more evolutionary close to eukaryotic genes than they are to most bacterial genes, explaining why the Mph model allowed for a much lower threshold score to be set while not gaining any obviously missclassified sequences.

The overall distribution of bacteria harboring the predicted *erm* genes, as displayed in Table 4.2 and Figure 4.1, is both reasonable and rather interesting. As bacteria

of the Firmicutes phylum are Gram-positive and thereby very susceptible to the effects of most macrolides, in addition to the phylum containing many pathogens that are prime targets for the clinical use of macrolides, it is clear that having these genes can provide an evolutionary advantage. Therefore it makes sense that Firmicutes are heavily overrepresented when it comes to harboring both known and new macrolide ARGs. A similar rationale can be applied regarding the overrepresentation of Actinobacteria among hosts of uncharacterized Erm variants, as these are also Gram-positive bacteria with some prominent pathogens, notably *Mycobacterium tuberculosis*, present within the phylum. The phylum contains a large diversity of different bacteria with different functions however, many of which exist as environmental bacteria [94]. It is also of note that all macrolide producing bacteria belong to the Actinobacteria phylum, and as mentioned in the Theory chapter it has been described that one of the main mechanisms used by such producers to achieve self-resistance is through Erm type methyltransferases. This information can be used to explain the overrepresentation of Actinobacteria, as they also obtain an evolutionary advantage from harboring Erm methyltransferases.

Proteobacteria and Bacteroidetes on the other hand are Gram-negative bacteria, and therefore not nearly as susceptible to macrolide antibiotics in general. This can help explain the massive underrepresentation of Proteobacteria harboring both known and new *erm* genes, as they are already intrinsically resistant to all naturally occurring macrolide antibiotics and would therefore only benefit from having these genes in environments were macrolides made or modified by humans, notably azithromycin, are being used. This is reflected in the results as most genes that were identified in Proteobacteria where known mobile variants such as *erm(B)* that exist in the clinical environment. There were a few exceptions to this however, which serves as an indication of some genes being mobile and/or gene transfer events having occurred some time in the past, as the sequences identified in Proteobacteria appeared in parts of the tree that clearly had an evolutionary connection to another phylum. This implies that while a relatively small number of *erm* genes have mobilized into Proteobacteria, the diversity of the gene families that have transferred is quite large as they originate from all across the phylogenetic tree.

The subject of evolution of Erm 23S-methyltransferases, as indicated by the phylogenetic tree presented in Figure 4.3 presents a number of insights, but also a number of questions. The two top left clusters in the tree, containing 12 known Erm sequences, are in part dominated by genes found in Firmicutes bacteria that have an association with the human and animal gut microbiota, such as *Clostridiales* and *Lachnospiraceae*, and in part dominated by sequences from other types of Firmicutes bacteria, notably *Staphylococcus*. Of the genes found in this part of the tree all but five variants (*erm(Q)*, *erm(43)*, *erm(44)*, *erm(45)*, and *erm(47)*) are known to be mobile, but their locations in the tree indicate that they all mobilized from very similar or the same ancestors in the past regardless of the hosts they are currently found in. This becomes clear when observing *erm(G)*, a gene that was most predominantly found in bacteria from the *Bacteroides* genus, its location in the tree clearly indicates that it has mobilized from bacteria from the Firmicutes phylum, which are genetically very different from Bacteroidetes.

The large group of sequences at the top right of Figure 4.3 is for this reason very curious. Here, new *erm* genes originating from most of the more exotic bacteria, not belonging to any of the four main phyla, cluster together. These bacteria are phylogenetically very diverse, hailing from a range of different phyla including more than 20 different '*Candidatus*' phyla, a notation that describes a type of bacteria that is not very well studied and also unculturable, but do not necessarily share a close taxonomic relationship [95]. What is perhaps more interesting than the fact that genes from these exotic bacteria cluster together however, is the fact that the previously described Firmicutes-based clusters are connected to this part of the tree. In between these parts are a group of gene families containing sequences from Firmicutes, Proteobacteria and Actinobacteria, which also includes the known genes *erm(42)* and *erm(49)*. This might indicate that these genes originate from these more exotic bacteria and have since transferred to other, more common bacteria.

The remaining parts of the tree are easier to summarize. The pair of Firmicutes-derived clusters to the left are heavily dominated by bacteria that would typically be found in the environment, mostly various types of *Bacilli*. It is interesting to observe this group in relation to the previously mentioned Firmicutes clusters, since those genes seemingly share a closer relationship with the genes identified in the miscellaneous exotic bacteria than with the genes in this cluster. Instead, the genes in the left cluster seemingly share a closer relationship with the genes identified in Bacteroidetes, which locate to a small cluster above. Aside from those harboring the known ErmF/Erm35 sequences, the remaining hosts represented in this cluster are dominated by environmentally associated species from Bacteroidetes, and this might provide an indication that these bacteria have shared genes with the aforementioned *Bacilli* within these environments long ago. The two Actinobacteria clusters that locate to the bottom of the tree, containing several known Erm variants, are seemingly closest to the environmentally associated Firmicutes and Bacteroidetes in terms of evolutionary distance. This is reasonable since Actinobacteria are typically associated with environmental samples as well, and points to a large diversity of *erm* genes hailing from the environment when considering the number of gene families that are associated with it. As a final note, there is seemingly no discernible difference between the organisms whose genes make up the two Actinobacteria clusters, meaning that there must be another reason for their division.

The metagenomic analysis further elucidated which environments *erm* methyltransferase genes currently exist in, and where they might mobilize from. By observing the results presented in Table 4.3, it becomes clear that the gut microbiota serves as a reservoir for both new and previously known Erm variants. Notably, most of the clinically relevant Erm variants were identified in both human and pig gut microbiome, as well as the samples from WWTP and the Pune river. Likely, the high levels of antibiotics present within these areas have resulted in a selection pressure leading to the need for many bacteria to acquire ARGs to gain a competitive advantage.

There are several implications that come with macrolide ARGs mobilizing from the human gut. The most major implication is that by having humans as a reservoir, these genes will inevitably end up in the clinical environment if they become mobile.

This is supported by the fact that all known mobile genes that were identified in human samples have been encountered in the clinical environment. Further, the fact that the effluent from WWTPs and polluted rivers such as the Pune contains macrolide ARGs is likely a result of the presence of human fecal matter in these environments, showcasing how ARGs can spread from the gut environment. The implication here is that these effluents flow out into the more remote environments, promoting the spread of macrolide ARGs at various locations across the world. For example, it is not unlikely that many of the environmentally associated species that were shown to harbor Erm variants from the genomic data originated from environments exposed to antibiotic selection pressure. However, the genes reconstructed from the forest soil and lake Hazen metagenomes further support that antibiotic resistance can develop even in pristine environments.

Considering all of the obtained results, this study was able to substantially contribute to the knowledge about *erm* type macrolide ARGs. Discounting the confirmed false positives, a total of 320 novel gene families of this type were identified, across a large number of diverse bacterial genomes and over a dozen terabases of metagenomic data. While it seems likely that many of these genes have not yet mobilized, and hopefully many of them never will and thereby will not cause any problem in the future, it is important to recognize that these genes exist in a variety far greater than previously known, and have the possibility of mobilizing from many different environments. Since a main reservoir is indicated to unfortunately be the gut microbiome, it is of utmost importance that we in the future limit our use of antibiotics to not further promote the spread of *erm* genes.

5.2 Mph macrolide phosphotransferases

The Mph model, similarly to the other two models, produced results that correspond well to what is stated in literature about these genes. In the genomes in NCBI GenBank, the most frequently found gene by far was *mph(A)*, followed by *mph(E)* and *mph(C)*. As these are all known to be mobile genes of clinical significance, these results are not surprising. Primarily MphA and MphE sequences were identified in Proteobacteria, with *Escherichia* and *Acinetobacter* being the most prevalent host genus respectively, though both variants were present in *Klebsiella* and *Salmonella* as well. Contrarily, MphC was most frequently occurring in bacteria from the Firmicutes phylum, specifically of the *Staphylococcus* genus. All of these three variants also had a small number of occurrences in bacteria from phyla other than the one primarily associated with them, confirming their mobility across phyla and contributing to the reliability of the results as this has been reported before (see the Theory chapter). In addition, as previously mentioned, the dissimilarity of *mph* phosphotransferase-genes to most other bacterial genes is likely what contributed to the lack of obvious false positives for this type of gene, despite a much lower threshold score being used during the analysis.

Comparing the distribution of species carrying *mph* genes presented in Table 4.1 and Figure 4.4, to the corresponding information for *erm* genes, the results are

both similar and different. The first thing to note is that Proteobacteria carrying *mph* genes are seemingly much more prevalent, however it should be noted that close to all occurrences of *mph* genes in Proteobacteria are represented by the two mobile genes *mph(A)* and *mph(E)*. This explains the small diversity of the genes found in Proteobacteria with respect to the total number of identified sequences, as well as the significant underrepresentation of Proteobacteria among the hosts of uncharacterzed *erm* genes visible in Figure 4.4. It is very reasonable that there is a large number of different Mph variants harbored by bacteria from the Firmicutes phylum, for the same reasons that they collectively host a large number of Erm variants. It is interesting however, that the number of occurrences in Actinobacteria is not very high, but highly diverse with respect to the overall count. This is reflected by their significant overrepresentiation among carriers of new Mph variants. Unlike Erm-mediated resistance, Mph-mediated resistance has not been reported as responsible for self resistance in macrolide producing Actinobacteria, however as they are Gram-positive bacteria that are susceptible to macrolides this means that they are indeed likely to gain a competitive advantage from harboring these genes. Finally, while the number of genes identified in bacteria from the Bacteroidetes phylum is comparatively very small, the diversity of the predicted ARG sequences is surprisingly high, suggesting that some of these bacteria obtained this type of genes long ago, and that they may have evolved along with the organisms.

The phylogenetic tree presented in Figure 4.6 tells an interesting and curious tale about the evolution of this type of macrolide resistance. It becomes clear from the tree that a quite large variety of *mph*-genes has evolved within Gram-negative bacteria, as two clusters mostly comprised of sequences identified in Proteobacteria and one cluster comprised mainly of sequences identified in Bacteroidetes can be clearly seen to the top right, left and top of the tree respectively. This is highly interesting since it has been shown that some Gram-negative bacteria, e.g *E. coli*, are intrinsically resistant to macrolides on account of the structure of their outer cellular membrane [96]. However, it has also been shown that this does not apply to all Gram-negative organisms, as *Bacteroides fragilis*, an anaerobic bacteria from the Bacteroidetes phylum that has a complete outer membrane, has proven to be susceptible to macrolide antibiotics [97].

This suggests that there are more complex factors that determine weather a bacterium is intrinsically resistant to macrolides than it just being Gram-negative. Looking towards the types of bacteria present within the aforementioned Gram-negative clusters, the majority of the gene families, besides the previously known ones that were present in pathogens, originated from the genomes of bacteria about which not a lot of information was available. Examples of genera represented are *Corallococcus* (Proteobacteria), *Methylophaga* (Proteobacteria), and *Sphingobacteria* (Bacteroidetes), and it is not unreasonable to assume that since macrolide resistance genes have evolved in these genera, they are likely among the Gram-negative bacteria that are not intrinsically resistant to macrolides. Regardless, it sees apparent from the phylogenetic tree that MphA and MphF have evolved in Proteobacteria, and that MphE and MphG have likely evolved in Bacteroidetes and then mobilized into Proteobacteria.

Despite a relatively high diversity of uncharacterized Mph variants being identified in bacteria of the Actinobacteria phylum, it is interesting to note that only two known variants, MphH and MphO, were identified in bacteria of this phylum. The sequences from Actinobacteria also notably clustered into two separate groups, with one being placed higher up in the tree, closer to the outgroup. These groups are seemingly defined by the species of Actinobacteria that the sequences were found in, with the larger group containing the two known variants being dominated by species like *Brevibacterium*, *Brachybacterium* and *Arthrobacter*, and some sequences identified in Deltaproteobacteria locating to this part of the tree as well. These misplaced Proteobacteria are particularly noteworthy, as they mainly belong to the the genus *Myxococcus*. These bacteria are predators that tend to feed on soil-dwelling bacteria [98], such as the Actinobacteria represented in these clades, and the fact that these predators have obtained genes that have evolved in their prey might suggest that genes are able to transfer to new hosts through consumption of the original host. This might be a way for genes to cross the phylum barrier, however it is of course also possible that these genes have simply mobilized into *Myxococcus* using other means as they exist in the same environments as the original hosts, unrelated to their predatory nature.

The smaller group corresponding to Actinobacteria that is located higher up in the tree is primarily comprised of sequences identified in bacteria from the genera *Streptomyces* and *Saccharopolyspora*. As mentioned in the introduction, these genera include bacteria that produce macrolides. Even though this type of resistance has not been reported in macrolide producers, the type of bacteria represented in this cluster, as well as its position in the tree, indicate that the gene families present within this group may be derivative of ancestral Mph sequences. These may have developed in producers or similar bacteria, and have later spread among other phyla and evolved into the variations that we know today.

By observing the remaining parts of the phylogenetic tree, it seems apparent that the majority of known *mph* genes have evolved in bacteria from the Firmicutes phylum. Similarly to the groups representing Actinobacteria, the Firmicutes-derived sequences also are divided into two groups, however when analyzing the species represented within these groups it was revealed that both groups were comprised of sequences that were primarily identified in different types of *Bacilli*. Instead it is likely that the groups here are defined by the substrate specificity of the encoded enzymes, as it MphK and MphI, which exist in the same group, are both known to have a narrow substrate range compared to most Mph variants. It is therefore likely that the uncharacterized Mph sequences within this group represent enzymes with similarly narrow substrate ranges, that likely evolved from broad-range enzymes that over time lost the affinity for some substrates. It is also of note that MphJ, the closest known variant in the tree, is not considered to have a narrow substrate range, and may thereby be representative of the Mph enzymes that evolved into the narrow-range enzymes over time given their close evolutionary distance [26].

The most noteworthy sequence that locates to the second Firmicutes-derived group is MphB, which corresponds to the sixth most prevalent of the genes in the data. This gene is very interesting, as it clearly has evolved in Firmicutes based on its

placement in the tree, yet more than 90% of the occurrences came from Proteobacteria, mostly *E. coli*. The remaining occurrences were identified in bacteria from the *Clostridium* genus, and it is likely that the gene mobilized from these bacteria into pathogens from Proteobacteria as a result of selection pressure or other factors some time in the past. Regardless, it is interesting to compare this gene, which is present in these Proteobacteria and known to be mobile, to another gene with the same properties. The perfect candidate for this comparison is MphA, a mobile gene which is also widely present in *E. coli*, but also seemingly has evolved in Proteobacteria rather than spreading there as a result of horizontal gene transfer. As was noted from Figure 4.5, MphA is so widespread that it represented close to half of all identified genes in the data, and *E. coli* genomes harboring MphA were more than 20 times as prevalent as *E. coli* genomes harboring MphB. There are likely several reasons for MphA being more prominent, one being that since MphB originates from a different phylum the codon structure of the gene *mph(B)* is not as optimal for *E. coli* as the one of *mph(A)*, since evolution within the same phylum has likely promoted the most efficient codons over time. This was confirmed during an analysis of these two genes, and another interesting fact that was revealed was that the two genes tended to be located on plasmids of the same incompatibility group, that being IncFIB. Since bacteria can not simultaneously promote two plasmids of the same incompatibility group, that means that in this case MphA, being able to be more efficiently transcribed and as previously noted in the Theory chapter as having a better affinity towards phosphorylating azithromycin, which is both the most prescribed macrolide as well as the one that is engineered to be used against Gram-negative bacteria, will provide the highest advantage and will therefore be promoted under selection pressure.

Observing the results of the metagenomic analysis, the environments where *mph* genes were present were mostly the same as the environments where *erm* genes were present, however the quantity of *mph* genes was much lower. This is in line with what has been stated in the literature, as it is known that Erm-mediated resistance is the most common type of macrolide resistance, and therefore not surprising. This means that *mph* genes may also mobilize from the gut microbiota and that antibiotic pollution has promoted the spread of *mph* genes. Aside from this however, there are a few interesting observations that can be made from the phylogenetic tree presented in Figure 4.10.

It is highly interesting that so many instances of previously uncharacterized sequences were assembled from different metagenomes that were identical to each other. As most of these instances were pairs of sequences were one sequence was assembled from the WWTP metagenome and the other from the Pune river metagenome. How these genes have transferred across the world between Sweden and India is possibly a result of humans traveling from one location to the other, bringing with them bacteria that have these genes that can then transfer them to the bacteria that exist in the other location. This likely also has the implication of these genes existing in other parts of the world, considering the distance between the locations corresponding to the samples in which they were identified during the metagenomic analysis here, and might be worth investigating further in case they have a possibility of be-

coming problematic. One of these pairs in particular clusters together with MphA in the tree, and given that that implies that they have similar properties this gene would be a prime candidate for further investigation. Another similar group of identical, unknown sequences clusters together with MphE, and was identified in all human metegenomes that were analyzed. Since the human samples came from both America and China, this might mean that this gene might exist in humans all over the world, and might mobilize from the gut microbiota in the future.

The final interesting point relates to the gene *mph(B)*. This gene accounted for almost a third of all identified Mph variants, while simultaneously being almost exclusively found in the pig gut microbiota. As mentioned above, it seems likely that the original hosts of *mph(B)* come from the genus *Clostridium* before the gene was mobilized into Proteobacteria. This agrees with the metagenomic results, as it has been shown that *Clostridium sp.* is very prevalent within the pig gut microbiota [99], and might imply that the original mobilization of this gene happened within the pig gut environment. Considering the amount of antibiotics that are used when breeding such animals today, this would provide a suitable selection pressure for this gene to become advantageous and widespread across the bacteria in the gut, and if this is in fact what happened this is a fine example of why antibiotic use should be limited not just for human applications, but for animal applications as well.

6

Conclusion

Through analysis of both genomic and metagenomic data, we were able to predict 320 novel *erm* gene families, not including false positives, as well as 221 novel *mph* gene families. This potentially represents a more than seven-fold increase in the number of known *erm* genes, as well as a more than fourteen-fold increase in the number of known *mph* genes. It is important to recognize however, that none of the predicted gene families can be considered for inclusion in the list of known macrolide ARGs before their functionality has been experimentally validated. This would therefore be a future prospect for this study.

Furthermore, through phylogenetic analysis we were able to elucidate the evolutionary relationships between the predicted ARG families, and suggest how they might have evolved. Finally, through metagenomic analysis we were able to conclude that the gut microbiome of humans and animals acts as a reservoir for both of the studied genetypes, where pathogens are likely to acquire mobile macrolide ARGs. This highlights the importance of taking proper measures in the future, as humans are likely to have spread these genes to environments across the world, as highlighted by two ARG hotspots identified in the metagenomic data, where one was located in India and the other in Sweden. More metagenomes would need to be analyzed to identify environments where these genes originally evolved and originally mobilized from, which would be useful information to have in order to, if possible, take precautions that would prevent new macrolide ARGs from mobilizing into the clinical environment. In total, this study highlights that the macrolide resistome is vaster and far more diverse than we are currently aware of and that it would be interesting to study it further, including resistance determinants that were not a part of this study.

Bibliography

[1] Julian Davies and Dorothy Davies. Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.*, 74(3):417–433, 2010.

[2] Jessica MA Blair, Mark A Webber, Alison J Baylay, David O Ogbolu, and Laura JV Piddock. Molecular mechanisms of antibiotic resistance. *Nature reviews microbiology*, 13(1):42, 2015.

[3] World Health Organization et al. Antibiotic resistance. *Fact sheet*, 2016.

[4] Jerome O Klein. History of macrolide use in pediatrics. *The Pediatric infectious disease journal*, 16(4):427–431, 1997.

[5] T Mazzei, E Mini, A Novelli, and P Periti. Chemistry and mode of action of macrolides. *Journal of Antimicrobial Chemotherapy*, 31(suppl_C):1–9, 1993.

[6] W. Schönfeld, H.A. Kirst, Michael J. Parnhma, and Jaques Bruinvels. *Macrolide Antibiotics*. Springer Basel AG, 2002.

[7] Fordyce R Heilman, Wallace E Herrell, William E Wellman, Joseph E Geraci, et al. Some laboratory and clinical observations on a new antibiotic, erythromycin (ilotycin). In *Proceedings of Staff Meetings of the Mayo Clinic*, volume 27, pages 285–304, 1952.

[8] David P Labeda. Transfer of the type strain of streptomyces erythraeus (waksman 1923) waksman and henrici 1948 to the genus saccharopolyspora lacey and goodfellow 1975 as saccharopolyspora erythraea sp. nov., and designation of a neotype strain for streptomyces erythraeus. *International Journal of Systematic and Evolutionary Microbiology*, 37(1):19–22, 1987.

[9] J-C Pechère. Macrolide resistance mechanisms in gram-positive cocci. *International journal of antimicrobial agents*, 18:25–28, 2001.

[10] SC Piscitelli, LH Danziger, and KA Rodvold. Clarithromycin and azithromycin: new macrolide antibiotics. *Clinical pharmacy*, 11(2):137–152, 1992.

[11] Krishna Kannan, Pinal Kanabar, David Schryer, Tanja Florin, Eugene Oh, Neil Bahroos, Tanel Tenson, Jonathan S Weissman, and Alexander S Mankin. The general mode of translation inhibition by macrolide antibiotics. *Proceedings of the National Academy of Sciences*, 111(45):15958–15963, 2014.

[12] Nora Vázquez-Laslop and Alexander S Mankin. How macrolide antibiotics work. *Trends in biochemical sciences*, 43(9):668–684, 2018.

[13] Rustam Aminov. History of antimicrobial drug discovery: Major classes and health impact. *Biochemical pharmacology*, 133:4–19, 2017.

[14] S Schwarz, C Kehrenberg, and TR Walsh. Use of antimicrobial agents in veterinary medicine and food animal production. *International journal of antimicrobial agents*, 17(6):431–437, 2001.

[15] World Health Organization et al. Critically important antimicrobials for human medicine: ranking of antimicrobial agents for risk management of antimicrobial resistance due to non-human use. 2017.

[16] Anna Knöppel, Joakim Näsvall, and Dan I Andersson. Evolution of antibiotic resistance without antibiotic exposure. *Antimicrobial agents and chemotherapy*, 61(11):e01495–17, 2017.

[17] Heather K Allen, Justin Donato, Helena Huimi Wang, Karen A Cloud-Hansen, Julian Davies, and Jo Handelsman. Call of the wild: antibiotic resistance genes in natural environments. *Nature reviews microbiology*, 8(4):251, 2010.

[18] Chandan Pal, Johan Bengtsson-Palme, Erik Kristiansson, and DG Joakim Larsson. The structure and diversity of human, animal and environmental resistomes. *Microbiome*, 4(1):54, 2016.

[19] Gerard D Wright. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nature Reviews Microbiology*, 5(3):175–186, 2007.

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

[21] Fanny Berglund, Tobias Österlund, Fredrik Boulund, Nachiket P Marathe, DG Joakim Larsson, and Erik Kristiansson. Identification and reconstruction of novel antibiotic resistance genes from metagenomes. *Microbiome*, 7(1):52, 2019.

[22] Fiona M Walsh and Sebastian GB Amyes. Microbiology and drug resistance mechanisms of fully resistant pathogens. *Current opinion in microbiology*, 7(5):439–444, 2004.

[23] Cláudia Gomes, Sandra Martínez-Puchol, Noemí Palma, Gertrudis Horna, Lidia Ruiz-Roldán, Maria J Pons, and Joaquim Ruiz. Macrolide resistance mechanisms in enterobacteriaceae: focus on azithromycin. *Critical reviews in microbiology*, 43(1):1–30, 2017.

[24] Birte Vester and Stephen Douthwaite. Macrolide resistance conferred by base substitutions in 23s rrna. *Antimicrobial agents and chemotherapy*, 45(1):1–12, 2001.

[25] G Maravic. Macrolide resistance based on the erm-mediated rrna methylation. *Current Drug Targets-Infectious Disorders*, 4(3):193–202, 2004.

[26] Andrew C Pawlowski, Peter J Stogios, Kalinka Koteva, Tatiana Skarina, Elena Evdokimova, Alexei Savchenko, and Gerard D Wright. The evolution of substrate discrimination in macrolide antibiotic resistance enzymes. *Nature communications*, 9(1):112, 2018.

[27] George P Dinos. The macrolide antibiotic renaissance. *British Journal of Pharmacology*, 174(18):2967–2983, 2017.

[28] A Portillo, M Lantero, I Olarte, F Ruiz-Larrea, and C Torres. Mls resistance phenotypes and mechanisms in β-haemolytic group b, c and g streptococcus isolates in la rioja, spain. *Journal of antimicrobial chemotherapy*, 47(1):115–116, 2001.

[29] Hiroshi Ogawara. Comparison of antibiotic resistance mechanisms in antibiotic-producing and pathogenic bacteria. *Molecules*, 24(19):3430, 2019.

[30] Alena Stsiapanava and Maria Selmer. Crystal structure of erme-23s rrna methyltransferase in macrolide resistance. *Scientific reports*, 9(1):1–9, 2019.

[31] Ae Kyung Park, Ho Kim, and Hyung Jong Jin. Phylogenetic analysis of rrna methyltransferases, erm and ksga, as related to antibiotic resistance. *FEMS microbiology letters*, 309(2):151–162, 2010.

[32] Tatsuhiko Kyuma, Hayato Kizaki, Hiroki Ryuno, Kazuhisa Sekimizu, and Chikara Kaito. 16s rrna methyltransferase ksga contributes to oxidative stress resistance and virulence in staphylococcus aureus. *Biochimie*, 119:166–174, 2015.

[33] Stefan Schwarz, Corinna Kehrenberg, and Kayode K Ojo. Staphylococcus sciuri gene erm (33), encoding inducible resistance to macrolides, lincosamides, and streptogramin b antibiotics, is a product of recombination between erm (c) and erm (a). *Antimicrobial agents and chemotherapy*, 46(11):3621–3623, 2002.

[34] MC Roberts. Tetracycline and mls nomenclature, 2006.

[35] Andrea T Feßler, Yang Wang, Congming Wu, and Stefan Schwarz. Mobile macrolide resistance genes in staphylococci. *Plasmid*, 99:2–10, 2018.

[36] Corey Fyfe, Trudy H Grossman, Kathy Kerstein, and Joyce Sutcliffe. Resistance to macrolide antibiotics in public health pathogens. *Cold Spring Harbor perspectives in medicine*, 6(10):a025395, 2016.

[37] Su-Young Kim, Seong Mi Moon, Byung Woo Jhun, O Jung Kwon, Hee Jae Huh, Nam Yong Lee, Seung Heon Lee, Sung Jae Shin, Shannon H Kasperbauer, Gwen A Huitt, et al. Species distribution and macrolide susceptibility of mycobacterium fortuitum complex clinical isolates. *Antimicrobial agents and chemotherapy*, 63(6):e02331–18, 2019.

[38] Roland Leclercq. Mechanisms of resistance to macrolides and lincosamides: nature of the resistance elements and their clinical implications. *Clinical Infectious Diseases*, 34(4):482–492, 2002.

[39] Dejun Liu, Weiwen Liu, Ziquan Lv, Junjie Xia, Xing Li, Yuxin Hao, Ying Zhou, Hong Yao, Zhihai Liu, Yang Wang, et al. Emerging erm (b)-mediated macrolide resistance associated with novel multidrug resistance genomic islands in campylobacter. *Antimicrobial agents and chemotherapy*, 63(7):e00153–19, 2019.

[40] Magdalena Szemraj, Anna Kwaszewska, and Eligia M Szewczyk. New gene responsible for resistance of clinical corynebacteria to macrolide, lincosamide and streptogramin b. *Polish journal of microbiology*, 67(2):237–240, 2018.

[41] Alberto Ortiz-Pérez, Nieves Z Martin-de Hijas, Jaime Esteban, María Isabel Fernández-Natal, José Ignacio García-Cía, and Ricardo Fernández-Roblas. High frequency of macrolide resistance mechanisms in clinical isolates of corynebacterium species. *Microbial Drug Resistance*, 16(4):273–277, 2010.

[42] Juliette RK Wipf, Matthew C Riley, Stephen A Kania, David A Bemis, Sabrina Andreis, Sybille Schwendener, and Vincent Perreten. New macrolide-lincosamide-streptogramin b resistance gene erm (48) on the novel plasmid pjw2311 in staphylococcus xylosus. *Antimicrobial agents and chemotherapy*, 61(7):e00066–17, 2017.

[43] Ml Monod, S Mohan, and David Dubnau. Cloning and analysis of ermg, a new macrolide-lincosamide-streptogramin b resistance element from bacillus sphaericus. *Journal of bacteriology*, 169(1):340–350, 1987.

[44] Elizabeth Luby Rieke, Thomas B Moorman, Elizabeth L Douglass, and Michelle L Soupir. Seasonal variation of macrolide resistance gene abundances in the south fork iowa river watershed. *Science of The Total Environment*, 610:1173–1179, 2018.

[45] Johan Bengtsson-Palme, Milena Milakovic, Helena Švecová, Marin Ganjto, Viktor Jonsson, Roman Grabic, and Nikolina Udikovic-Kolic. Industrial wastewater treatment plant enriches antibiotic resistance genes and alters the structure of microbial communities. *Water research*, 162:437–445, 2019.

[46] Tolou Golkar, Michał Zieliński, and Albert M Berghuis. Look and outlook on enzyme-mediated macrolide resistance. *Frontiers in Microbiology*, 9:1942, 2018.

[47] Olivier Chesneau, Krassimira Tsvetkova, and Patrice Courvalin. Resistance phenotypes conferred by macrolide phosphotransferases. *FEMS microbiology letters*, 269(2):317–322, 2007.

[48] Rafael Szczepanowski, Burkhard Linke, Irene Krahn, Karl-Heinz Gartemann, Tim Guetzkow, Wolfgang Eichler, Alfred Pühler, and Andreas Schlueter. Detection of 140 clinically relevant antibiotic-resistance genes in the plasmid

metagenome of wastewater treatment plant bacteria showing reduced susceptibility to selected antibiotics. *Microbiology*, 155(7):2306–2319, 2009.

[49] Kazuo Taniguchi, Akio Nakamura, Kazue Tsurubuchi, Aki Ishii, Koji O'Hara, and Tetsuo Sawai. Identification of functional amino acids in the macrolide 2-phosphotransferase ii. *Antimicrobial agents and chemotherapy*, 43(8):2063–2065, 1999.

[50] Andrew C Pawlowski, Erin L Westman, Kalinka Koteva, Nicholas Waglechner, and Gerard D Wright. The complex resistomes of paenibacillaceae reflect diverse antibiotic chemical ecologies. *The ISME journal*, 12(3):885, 2018.

[51] Yuta Sugimoto, Satoru Suzuki, Lisa Nonaka, Chanchai Boonla, Nop Sukpanyatham, Hsin-Yiu Chou, and Jer-Horng Wu. The novel mef (c)–mph (g) macrolide resistance genes are conveyed in the environment on various vectors. *Journal of global antimicrobial resistance*, 10:47–53, 2017.

[52] Maxwell Finland. Emergence of antibiotic resistance in hospitals, 1935–1975. *Reviews of infectious diseases*, 1(1):4–21, 1979.

[53] ML Diaz-Torres, R McNab, DA Spratt, A Villedieu, N Hunt, M Wilson, and P Mullany. Novel tetracycline resistance determinant from the oral metagenome. *Antimicrobial agents and chemotherapy*, 47(4):1430–1432, 2003.

[54] Peter J Petersen, NV Jacobus, WJ Weiss, PE Sum, and RT Testa. In vitro and in vivo antibacterial activities of a novel glycylcycline, the 9-t-butylglycylamido derivative of minocycline (gar-936). *Antimicrobial agents and chemotherapy*, 43(4):738–744, 1999.

[55] Erdal Toprak, Adrian Veres, Jean-Baptiste Michel, Remy Chait, Daniel L Hartl, and Roy Kishony. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nature genetics*, 44(1):101, 2012.

[56] Andreas D Baxevanis, Gary D Bader, and David S Wishart. *Bioinformatics*. John Wiley & Sons, 2020.

[57] Archana Tiwari et al. Applications of bioinformatics tools to combat the antibiotic resistance. In *2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI)*, pages 96–98. IEEE, 2015.

[58] Elaine R Mardis. Dna sequencing technologies: 2006–2016. *Nature protocols*, 12(2):213, 2017.

[59] Bagos' PG, Th D Liakopoulos, and SJ Hamodrakas. Efficient training of hidden markov models for protein sequence analysis. In *International Conference of Computational Methods in Sciences and Engineering (ICCMSE 2004)*, page 53. CRC Press, 2019.

[60] Sean R. Eddy. Profile hidden markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763, 1998.

[61] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[62] Yariv Ephraim and Neri Merhav. Hidden markov processes. *IEEE Transactions on information theory*, 48(6):1518–1569, 2002.

[63] Magdi A Mohamed and Paul Gader. Generalized hidden markov models. i. theoretical frameworks. *IEEE Transactions on fuzzy systems*, 8(1):67–81, 2000.

[64] Phil Blunsom. Hidden markov models. *Lecture notes, August*, 15(18-19):48, 2004.

[65] Sean R Eddy. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365, 1996.

[66] Anders Krogh, Michael Brown, I Saira Mian, Kimmen Sjolander, and David Haussler. Hidden markov models in computational biology. applications to protein modeling. *Journal of molecular biology*, 235(5):1501–1531, 1994.

[67] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic acids research*, 37(suppl_1):D26–D31, 2008.

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

[69] Fábio Madeira, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian Tivey, Simon C Potter, Robert D Finn, Rodrigo Lopez, et al. The embl-ebi search and sequence analysis tools apis in 2019. *Nucleic acids research*, 2019.

[70] Aron Marchler-Bauer, Myra K Derbyshire, Noreen R Gonzales, Shennan Lu, Farideh Chitsaz, Lewis Y Geer, Renata C Geer, Jane He, Marc Gwadz, David I Hurwitz, et al. Cdd: Ncbi's conserved domain database. *Nucleic acids research*, 43(D1):D222–D226, 2014.

[71] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, 30(14):3059–3066, 2002.

[72] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490, 2010.

[73] Guangchuang Yu, David Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8:28–36, 2017.

[74] Guangchuang Yu, Tommy Tsan-Yuk Lam, Huachen Zhu, and Yi Guan. Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Molecular Biology and Evolution*, 35:3041–3043, 2018.

[75] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[76] Marilyn C Roberts, Joyce Sutcliffe, Patrice Courvalin, Lars Bogo Jensen, Julian Rood, and Helena Seppala. Nomenclature for macrolide and macrolide-lincosamide-streptogramin b resistance determinants. *Antimicrobial agents and chemotherapy*, 43(12):2823–2830, 1999.

[77] Jonathan Kans. Entrez direct: E-utilities on the unix command line. In *Entrez Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US), 2020.

[78] Ivica Letunic and Peer Bork. Interactive tree of life (itol) v4: recent updates and new developments. *Nucleic acids research*, 47(W1):W256–W259, 2019.

[79] T Human. Project m. structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–14, 2012.

[80] Janis R Bedarf, Falk Hildebrand, Luis P Coelho, Shinichi Sunagawa, Mohammad Bahram, Felix Goeser, Peer Bork, and Ullrich Wüllner. Functional implications of microbial and viral gut metagenome changes in early stage l-dopa-naïve parkinson's disease patients. *Genome medicine*, 9(1):39, 2017.

[81] Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, 2012.

[82] Liang Xiao, Jordi Estellé, Pia Kiilerich, Yuliaxis Ramayo-Caldas, Zhongkui Xia, Qiang Feng, Suisha Liang, Anni Øyan Pedersen, Niels Jørgen Kjeldsen, Chuan Liu, et al. A reference gene catalogue of the pig gut microbiome. *Nature microbiology*, 1(12):1–6, 2016.

[83] Johan Bengtsson-Palme, Rickard Hammaren, Chandan Pal, Marcus Östman, Berndt Björlenius, Carl-Fredrik Flach, Jerker Fick, Erik Kristiansson, Mats Tysklind, and DG Joakim Larsson. Elucidating selection processes for antibiotic resistance in sewage treatment plants using metagenomics. *Science of the Total Environment*, 572:697–712, 2016.

[84] Nachiket P Marathe, Chandan Pal, Swapnil S Gaikwad, Viktor Jonsson, Erik Kristiansson, and DG Joakim Larsson. Untreated urban waste contaminates indian river sediments with resistance genes to last resort antibiotics. *Water research*, 124:388–397, 2017.

[85] Eric Karsenti, Silvia G Acinas, Peer Bork, Chris Bowler, Colomban De Vargas, Jeroen Raes, Matthew Sullivan, Detlev Arendt, Francesca Benzoni, Jean-Michel Claverie, et al. A holistic approach to marine eco-systems biology. *PLoS biology*, 9(10), 2011.

[86] Olivia U Mason, Terry C Hazen, Sharon Borglin, Patrick SG Chain, Eric A Dubinsky, Julian L Fortney, James Han, Hoi-Ying N Holman, Jenni Hultman, Regina Lamendella, et al. Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to deepwater horizon oil spill. *The ISME journal*, 6(9):1715–1727, 2012.

[87] Mukan Ji, Chris Greening, Inka Vanwonterghem, Carlo R Carere, Sean K Bay, Jason A Steen, Kate Montgomery, Thomas Lines, John Beardall, Josie Van Dorst, et al. Atmospheric trace gases support primary production in antarctic desert surface soil. *Nature*, 552(7685):400–403, 2017.

[88] Bin Ma, Kankan Zhao, Xiaofei Lv, Weiqin Su, Zhongmin Dai, Jack A Gilbert, Philip C Brookes, Karoline Faust, and Jianming Xu. Genetic correlation network prediction of forest soil microbial functional organization. *The ISME journal*, 12(10):2492–2505, 2018.

[89] Jenny Tung, Luis B Barreiro, Michael B Burns, Jean-Christophe Grenier, Josh Lynch, Laura E Grieneisen, Jeanne Altmann, Susan C Alberts, Ran Blekhman, and Elizabeth A Archie. Social networks predict gut microbiome composition in wild baboons. *elife*, 4:e05224, 2015.

[90] Katherine R Amato, Elizabeth K Mallott, Daniel McDonald, Nathaniel J Dominy, Tony Goldberg, Joanna E Lambert, Larissa Swedell, Jessica L Metcalf, Andres Gomez, Gillian AO Britton, et al. Convergence of human and old world monkey gut microbiomes demonstrates the importance of human ecology over phylogeny. *Genome biology*, 20(1):201, 2019.

[91] Keylie M Gibson, Bryan N Nguyen, Laura M Neumann, Michele Miller, Peter Buss, Savel Daniels, Michelle J Ahn, Keith A Crandall, and Budhan Pukazhenthi. Gut microbiome differences between wild and captive black rhinoceros–implications for rhino health. *Scientific reports*, 9(1):1–11, 2019.

[92] Graham A Colby, Matti O Ruuskanen, Kyra A St Pierre, Vincent L St Louis, Alexandre J Poulain, and Stéphane Aris-Brosou. Climate change lowers diversity and functional potential of microbes in canada's high arctic. *bioRxiv*, page 705178, 2019.

[93] Michael Tessler, Johannes S Neumann, Ebrahim Afshinnekoo, Michael Pineda, Rebecca Hersch, Luiz Felipe M Velho, Bianca T Segovia, Fabio A Lansac-Toha, Michael Lemke, Rob DeSalle, et al. Large-scale differences in microbial biodiversity discovery between 16s amplicon and shotgun sequencing. *Scientific reports*, 7(1):1–14, 2017.

[94] Vivian Miao and Julian Davies. Actinobacteria: the good, the bad, and the ugly. *Antonie Van Leeuwenhoek*, 98(2):143–150, 2010.

[95] RGE Murray and E Stackebrandt. Taxonomic note: implementation of the provisional status candidatus for incompletely described procaryotes. *International Journal of Systematic and Evolutionary Microbiology*, 45(1):186–187, 1995.

[96] Roland Leclercq and Patrice Courvalin. Intrinsic and unusual resistance to macrolide, lincosamide, and streptogramin antibiotics in bacteria. *Antimicrobial agents and chemotherapy*, 35(7):1273, 1991.

[97] YOSHINORI Muto, KAORI Bandoh, KUNITOMO Watanabe, NAOKI Katoh, and KAZUE Ueno. Macrolide accumulation by bacteroides fragilis atcc 25285. *Antimicrobial agents and chemotherapy*, 33(2):242–244, 1989.

[98] Andrew D Morgan, R Craig MacLean, Kristina L Hillesland, and Gregory J Velicer. Comparative analysis of myxococcus predation on soil bacteria. *Appl. Environ. Microbiol.*, 76(20):6920–6927, 2010.

[99] Scot E Dowd, Yan Sun, Randy D Wolcott, Alexander Domingo, and Jeffery A Carroll. Bacterial tag–encoded flx amplicon pyrosequencing (btefap) for microbiome studies: Bacterial diversity in the ileum of newly weaned salmonella-infected pigs. *Foodborne pathogens and disease*, 5(4):459–472, 2008.

A

Supplementary Figures

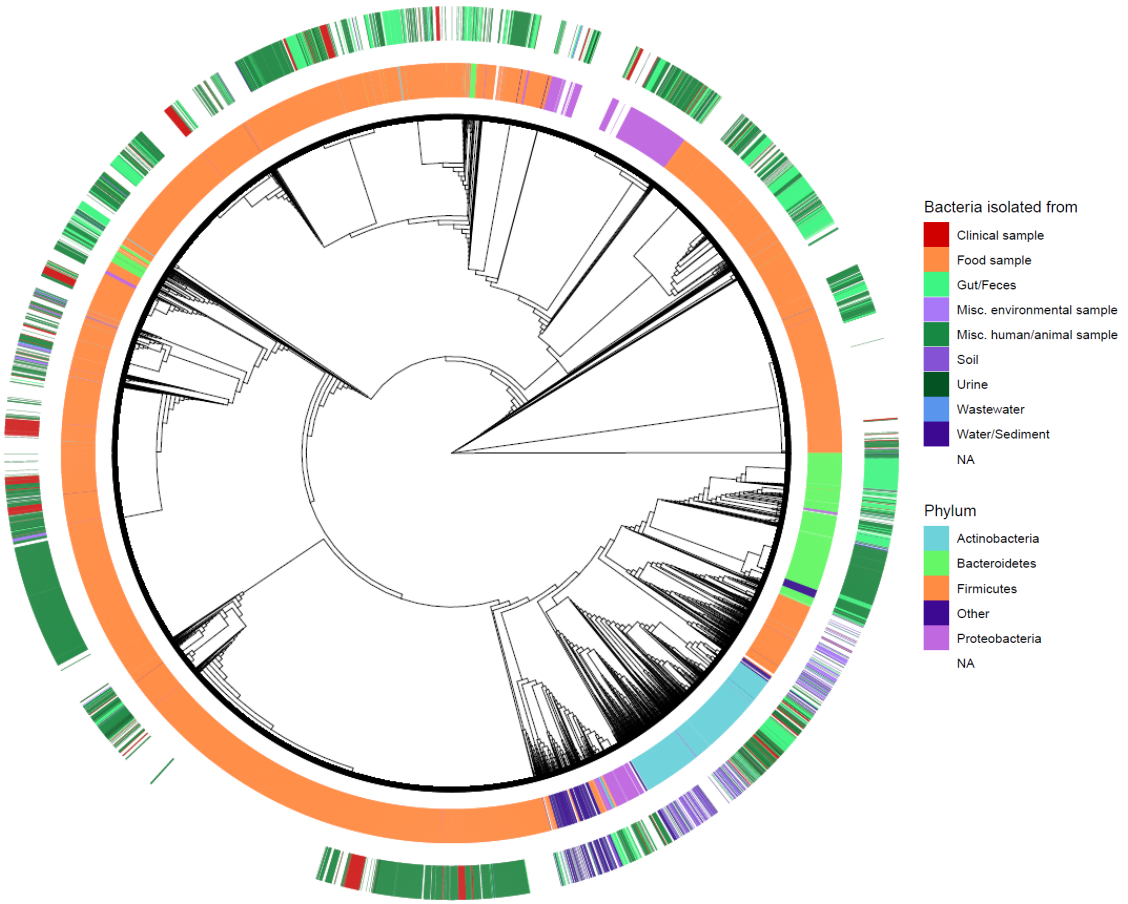


Figure A.1: *Cladogram displaying all of the genes predicted by the two erm models. The innermost of the map represents the phylum of the organism where the sequence was identified, and the outermost map represents the source from which the organism was isolated.*

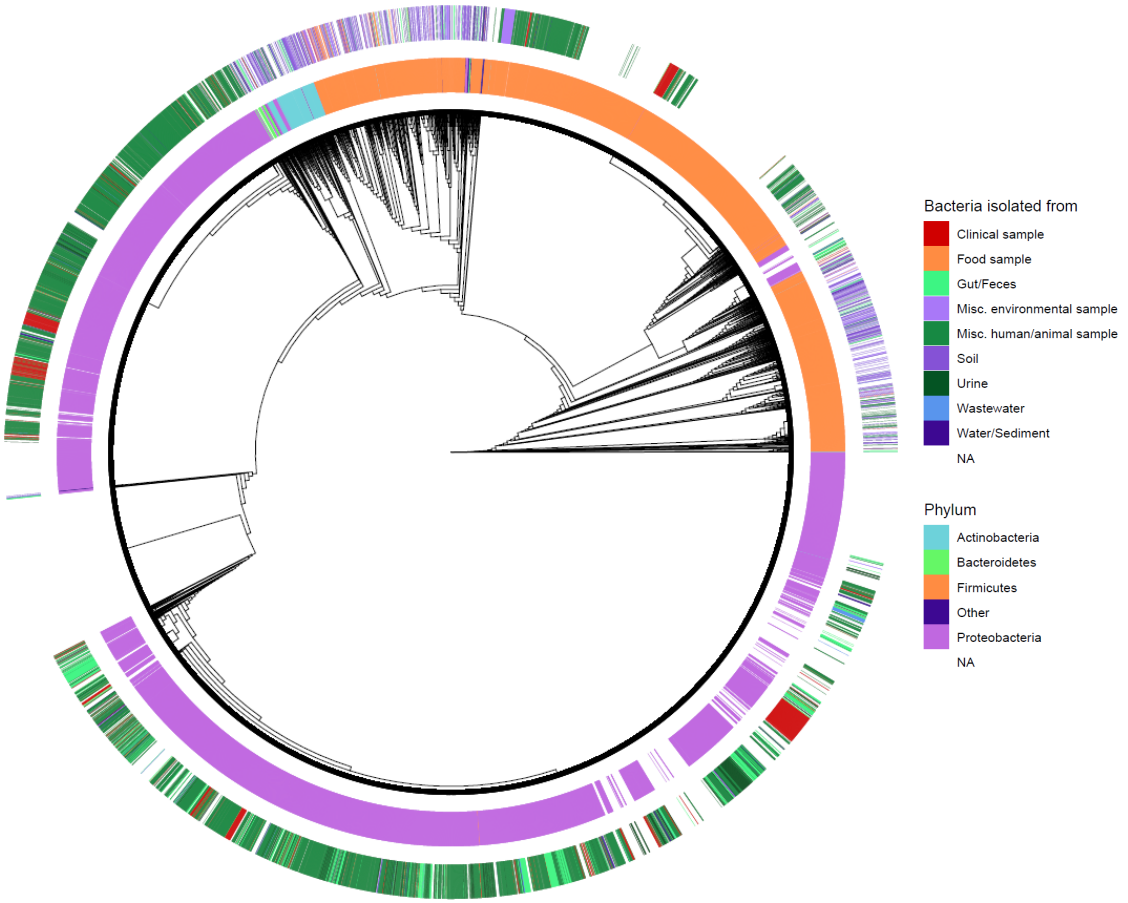


Figure A.2: *Cladogram displaying all of the genes predicted by the mph model. The innermost map represents the phylum of the organism where the sequence was identified, and the outermost map represents the source from which the organism was isolated.*