THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Computational discovery of antibiotic resistance genes and their horizontal transfer

David Lund

Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden 2022

Computational discovery of antibiotic resistance genes and their horizontal transfer

David Lund

© David Lund, 2022

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

Typeset with LaTeX
Printed by Chalmers digitaltryck
Göteborg, Sweden 2022

Computational discovery of antibiotic resistance genes and their horizontal transfer

David Lund

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

Abstract

Antibiotic resistance is increasing among clinical infections and represents one of the most serious threats to public health. Pathogens often become resistant by acquiring mobile antibiotic resistance genes (ARGs) via horizontal gene transfer (HGT). To limit the spread of new ARGs, it is important that we identify emerging threats early, and that we improve our understanding of what drives the HGT of ARGs. The three papers encompassing this thesis aim to increase our knowledge about ARGs and their mobility. In paper I, computational screening of large genomic datasets was used to identify new resistance genes for macrolide antibiotics, and to clarify their evolution. A large diversity of new *erm* and *mph* genes was identified, including six new families of mobile ARGs carried by pathogens, that showed varied phylogenetic origins. Of the tested genes, 70% induced resistance in *Escherichia coli*. In paper II, we identified previously undiscovered mobile genes giving resistance to aminoglycoside antibiotics in pathogens, further demonstrating how computational methods can discover potential emerging ARGs. Close to one million bacterial genomes were screened for *aac* and *aph* genes, and the mobility of each predicted gene was evaluated. A total of 50 families of new mobile ARGs were identified in pathogens. When new ARGs were tested in *E. coli*, 86% were functional, with 39% giving clinical resistance. In paper III, the factors influencing the HGT of ARGs were investigated. Phylogenetic analysis was used to identify HGT events from a large set of ARGs. For each event, the genetic compatibility of the involved gene(s) and genomes, as well as the co-occurrence of donor and recipient in different environments, were computed and used as input to train random forest classifiers. The resulting models suggested that the most important factor for determining if a mobile ARG successfully undergoes horizontal transfer is the genetic compatibility between the gene and the recipient genome. The findings presented in this thesis increase our knowledge about new genes giving resistance to two important classes of antibiotics. Furthermore, the results provide new insights into the horizontal transfer of resistance genes.

Keywords: Antibiotic resistance, horizontal gene transfer, computational screening, phylogenetic analysis

List of publications

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

I. **Lund D**, Kieffer N, Parras-Moltó M, Ebmeyer S, Berglund F, Johnning A, Larsson D.G.J, Kristiansson E (2022). Large-scale characterization of the macrolide resistome reveals high diversity and several new pathogen-associated genes. *Microbial genomics*. 8(1):000770. doi: 10.1016/j.ejps.2021.105937

II. **Lund D**, Coertze R.D, Parras-Moltó M, Berglund F, Flach C-F, Johnning A, Larsson D.G.J, Kristiansson E (2022). Computational screening reveals previously undiscovered aminoglycoside resistance genes in human pathogens. *Manuscript*.

III. **Lund D**, Parras-Moltó M, Boström M, Inda-Díaz J.S, Benson L, Ebmeyer S, Larsson D.G.J, Johnning A, Kristiansson E (2022). Factors influencing the horizontal gene transfer potential of antibiotic resistance genes. *Manuscript*.

Additional papers not included in this thesis:

IV. Inda-Díaz J.S, **Lund D**, Parras-Moltó M, Johnning A, Bengtsson-Palme J, Kristiansson E. Latent antibiotic resistance genes are abundant, diverse, and mobile in human, animal, and environmental microbiomes (2022). *Submitted*.

V. Parras-Moltó M, **Lund D**, Johnning A, Kristiansson E (2022). Interphyla transfer of antibiotic resistance genes. *Manuscript*.

Author contributions

I. Participated in study design, collected data, created and optimized the models, performed the data analysis, created the phylogenetic trees, performed phylum enrichment analysis, performed genetic context analysis, analyzed the results, drafted and edited the manuscript.

II. Participated in study design, created and optimized the models, created the phylogenetic trees, performed phylum enrichment analysis, performed genetic context analysis, analyzed the results, drafted and edited the manuscript.

III. Participated in study design, collected data, trained the machine learning models, analyzed the results, drafted and edited the manuscript.

Acknowledgements

First, and most importantly, I want to thank my supervisor Erik Kristiansson. Your constant positivity and enthusiasm about any progress, however minor it may seem to me, has been a much-needed source of motivation for me during my first years as a Ph.D. student. Without your guidance and support, I would not be where I am today, and for that, I am truly grateful. Next, I want to thank my co-supervisors Anna Johnning and Joakim Larsson. Thank you, Anna, for always being available to provide support, feedback, and discussion, research related or not. Thank you, Joakim, for providing expertise and encouragement. I also want to thank my examiner Fredrik Westerlund.

I want to thank my collaborators at Sahlgrenska: Fanny Berglund, Roelof Coertze, Stefan Ebmeyer, and Nicolas Kieffer, for your valuable input and contributions to this thesis. You have all been fantastic to work with.

I would like to thank the past and current members of Erik Kristiansson's research group: Anna Johnning, Mikael Gustavsson, Juan Inda, Marcos Parras Moltó, Patrik Svedberg, Astrid von Mentzer, Martin Boström, Styrbjörn Käll and, of course, Erik Kristiansson, for creating a fun and inspiring work environment. Additionally, I want to thank all my other colleagues at the Department of Mathematical Sciences.

Lastly, I would like to thank all of my friends and family for always supporting me. Special thanks go to my bandmates Joar, David, and Samuel, for helping me keep my sanity during trying times, and to my partner Julia, for always showing me the way when I am lost.

David Lund,
Gothenburg, December 2022

Contents

Abstract **iii**

List of publications **v**

Acknowledgements **vii**

Contents **ix**

1 Background **1**

1.1 Antibiotic resistance 1

1.2 Horizontal gene transfer 2

2 Aims **5**

3 Methods **7**

3.1 Computational prediction of antibiotic resistance genes from sequence data 7

3.2 Phylogenetic analysis 8

3.3 Detection of horizontal gene transfer events from sequence data 11

4 Summary of results **15**

4.1 Papers I and II 15
4.2 Paper III 16

5 Future work **19**

Bibliography **21**

Papers I-III

1 Background

The discovery of antibiotics, and their subsequent introduction into clinical use during the early 20th century, represents one of the most important historical advances in human health. The importance of these compounds can not be overstated, since in addition to their use for the treatment and prevention of infectious diseases they have also enabled the development of many modern medical and surgical procedures [1]. During the mid-20th century, sometimes referred to as the Golden Age of Antibiotics, a large variety of antibiotics were discovered. These mostly encompassed antimicrobial compounds naturally produced by bacteria or fungi, though some compounds were produced either synthetically or semi-synthetically [2]. However, the development of new antibiotics drastically slowed down after the 1960s, and since the end of the Golden Age, only two new classes of antibiotics have been introduced [3].

1.1 Antibiotic resistance

While human society continues to rely heavily on antibiotics for healthcare, human pathogens are gradually becoming immune to their antimicrobial effects, a phenomenon known as antibiotic resistance. This poses an obvious threat to human health, as it has the potential of making some infectious diseases harder, or even impossible to treat [4]. In general, bacteria develop antibiotic resistance either from mutations in their preexisting DNA as a result of adaptive evolution or through the acquisition of specific genes, so-called antibiotic resistance genes (ARGs). While mutations are an important source of resistance in some pathogens, they can only be transferred vertically between parent and offspring. By contrast, ARGs can be acquired from distantly related cells through horizontal gene transfer (HGT). Leveraging this, some bacteria have accumulated ARGs from different sources over time into large genetic constructs that, when acquired, confer resistance to many classes of antibiotics

[5].

Today, hundreds of ARGs have been identified, each specific to one or a few classes of antibiotics, and most are associated with specific mechanisms. Broadly, these include: reduced access of the drug to the cell, either via reduced membrane permeability or through efflux of the drug from the cell; modification of the drug target, where a functional group or protein binds to and occupies the site where the antibiotic would attach itself; or direct modification of the antibiotic, where an enzyme reacts with the drug molecule and inactivates it via some chemical reaction [6]. Since the introduction of antibiotics, ARGs conferring resistance to almost all classes of antibiotics used for treating infection have emerged in clinical pathogens. While some of these exist naturally in the chromosome of certain pathogens, the majority have been acquired through HGT. Indeed, new ARGs are regularly discovered in clinical infections, showing that the mobilization of resistance determinants is still ongoing. The evolutionary origins of these genes, however, are still not well understood, which hampers our ability to prevent the further spread of ARGs [7].

While the evolutionary details of most ARGs remain unclear, it is known that these genes existed well before humans started to treat infections with antibiotics. Indeed, it has been suggested that some ARGs first evolved billions of years ago [8]. Since they first arose, ARGs have evolved and diversified over long periods, which has resulted in a vast resistome, i.e. the complete collection of ARGs, that can be observed today. In addition to the resistance determinants that are commonly encountered in pathogens, the current resistome encompasses a genetic diversity that far exceeds what has hitherto been recorded in the clinical environment. These diverse ARGs exist in many different types of environments, both external environments like soil and water and host-associated environments like human microbiomes and sewage [9]. Though the evolution of the resistome happened independently of human interference, the problem has been exacerbated by humanity's excessive use of antibiotics over the last century. The increased concentrations of antibiotics in different environments have provided enough selection pressure for resistance genes to develop, mobilize and be shared within and between bacterial communities at rates that were likely not attained in the pre-antibiotic era [10].

1.2 Horizontal gene transfer

A large proportion of ARGs that are carried by clinical pathogens today is mobile, meaning that the genes can be shared between bacteria through HGT. This allows for ARGs to be shared between unrelated species, for example

between harmless commensal species and pathogenic species. This makes HGT one of the main causes of the spread of resistance, and today the HGT phenomenon is actively studied in relation to antibiotic resistance [11].

Typically, mobile genes are associated with and/or carried by mobile genetic elements (MGEs). These MGEs are genes encoding proteins that enable the genetic material to move within and between cells. There are many different MGEs associated with ARGs, including e.g. conjugative elements, integrons, and insertion sequences/transposons. It is not uncommon for several of these to exist together on larger mobile genetic constructs like plasmids. The movement of DNA between cells can happen through one of three mechanisms: natural transformation, or uptake of free (non-cell bound) DNA into the cell, transduction, where the transfer is mediated by bacteriophages, and, perhaps most importantly, conjugation, whereby a sex pilus is formed between adjacent bacteria through which the DNA moves from the donor cell to the recipient. Conjugation requires a specific set of genes to initiate the transfer which are often located together on conjugative plasmids together with other genetic material. Where transformation and transduction can occur as side-effects of other biological processes, plasmid conjugation, by comparison, is a more efficient and reliable way for the recipient to acquire foreign DNA directly from the donor [12, 13]. In some instances, conjugation also enables the recipient to develop resistance towards multiple antibiotics through a single HGT event, by acquiring a large multidrug-resistance plasmid [14].

Though the mechanisms by which HGT happens are relatively well-studied, much remains unclear about why HGT happens and between which species. It is clear that we need to increase our knowledge about the dissemination of ARGs via HGT to combat the spread of new forms of multidrug-resistant pathogens.

2 Aims

The overall aim of this thesis is to increase our knowledge about mobile antibiotic resistance genes, and how these genes moved from external environments into clinical pathogens. Each of the three papers encompassing the thesis contributes to this via the following aims:

- Identify new genes conferring resistance to macrolide and aminoglycoside antibiotics, elucidate their evolutionary history, and experimentally validate their functionality (paper I and II).
- Detect new mobile resistance genes carried by pathogens (paper II).
- Identify the main factors influencing horizontal gene transfer of antibiotic resistance genes (paper III).

3 Methods

This chapter provides a brief description of some of the main methods that were used in the papers presented in this thesis.

3.1 Computational prediction of antibiotic resistance genes from sequence data

Currently, a large number of ARGs, conferring resistance to all classes of antibiotics used for clinical infection treatment, is known to circulate among human pathogens [15]. When new resistance genes emerge, we are often unable to discover them before they are widely disseminated among bacterial communities using traditional methods. For example, this was the case for the beta-lactamase NDM-1 and the colistin resistance determinant MCR-1 [16, 17]. To overcome the drawbacks of traditional surveillance, several computational methods have recently been developed that can identify ARGs, including new variants, from whole genome sequencing (WGS) and metagenomic sequencing data [18].

These methods use different computational frameworks for the prediction of resistance genes. In general, they create models based on resistance gene databases like ResFinder [19] and CARD [20], that can identify homologs to the reference genes present in these databases based on similarities in gene sequence or protein structure [21]. One of the most well-established computational frameworks for gene prediction is based on profile hidden Markov models (HMMs). These models are built from multiple sequence alignments and can identify homologous genes based on conserved genetic regions rather than overall sequence similarity [22]. This allows for the discovery of previously unknown gene variants but is restricted to the identification of genes with similar functions as the reference genes used to build the models. One

method that uses profile HMMs is fARGene, which in addition to the prediction of ARGs in WGS data also enables gene prediction in metagenomic data without the need for prior assembly [23]. This method has repeatedly shown a high performance for predicting functional new ARGs, in addition to their well-characterized counterparts, from a variety of datasets [24, 25, 26], proving the reliability of HMM-based ARG predictions.

More recently, methods have been developed that apply machine learning algorithms for predicting ARGs. A prominent example is deepARG, which takes a deep learning approach, and uses algorithms and models that can discriminate between true ARGs and genes that contain some ARG-like regions without conferring resistance [27]. Another example is PCM, a method that uses machine learning to make predictions based on protein structure [28]. Similar to the HMM-based methods, the machine learning models are also able to identify previously uncharacterized homologs but are unable to predict ARGs associated with novel resistance mechanisms.

3.2 Phylogenetic analysis

When new genes are identified, it is often of interest to deduce their evolutionary history and relationships with other similar genes to better understand their function and origin. Typically, this analysis is based on an inferred phylogeny, calculated through computational phylogenetics. Phylogenetic analysis aims to divulge the evolutionary relationships of genes or taxa by reconstructing phylogenetic trees, simulations of the evolutionary tree computed from molecular sequences. As the amount of available DNA sequencing data has increased, so has the demand for phylogenetic analysis, leading to the development of more efficient and sophisticated methods [29].

A phylogenetic tree is in essence a branching diagram used to visualize the evolutionary relationships between different taxa or genes. The observed sequences that are used to construct the tree are referred to as leaves. These are placed at the tips of branches which are connected by internal nodes in the tree. Together, all leaves that extend from the same node are called a clade. The node that is determined to be the oldest from an evolutionary standpoint is referred to as the root of the tree and can either be inferred from the tree-building algorithm or deliberately selected based on assumptions [30] (Fig. 3.1). Placing a root is not required when building a tree, however, the root provides an evolutionary direction. Consequently, an un-rooted tree can only be used to infer the relatedness of the leaves, and not the evolutionary history [31].

The methods used for phylogenetic tree reconstruction can broadly be divided into two categories; distance-based methods, such as neighbor-joining and least-squares, and character-based methods, including maximum parsimony, maximum likelihood, and Bayesian algorithms. Distance-based methods use measures of genetic distance computed from a multiple sequence alignment to derive the phylogenetic tree. Assuming that all genetic divergence events throughout history are accurately recorded in the sequence that we can observe today, the distance (or the amount of dissimilarity between two aligned sequences) could be used to reconstruct the true evolutionary tree. Among the distance-based methods, the neighbor-joining algorithm is the most widely used [32].

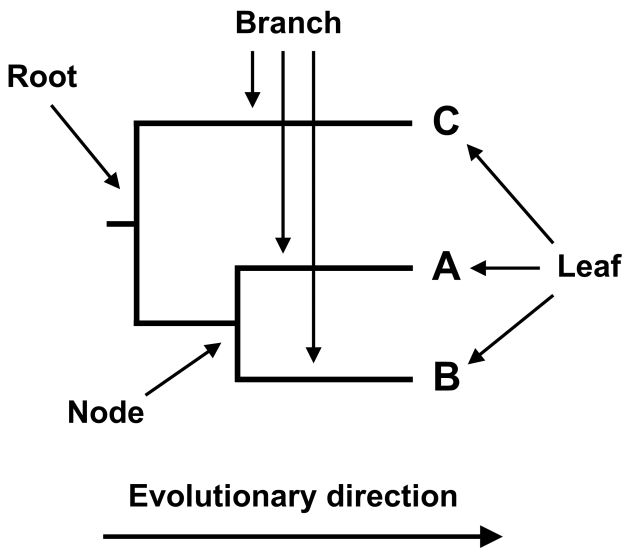

Figure 3.1: Basic illustration of a phylogenetic tree, including the nomenclature commonly used to denote the different components of the tree.

Briefly, neighbor-joining starts by assuming an un-rooted, bifurcating tree with N leaves. A pair of neighbors is defined as two leaves connected through one interior node in the tree. The topology of the tree can then be defined by successively merging pairs of neighbors, producing new pairs of neighbors that are again merged until a consensus topology is achieved.

During each iteration, a distance matrix D is calculated from all pairwise distances of the leaves in the tree. For each pair of leaves a, b, the matrix is then used to compute the sum of branch lengths in the tree after merging the leaves as the sum of the least square estimates of branch lengths. The pair of leaves producing the smallest sum are then joined to make a combined leaf $(a - b)$. The distance between $(a - b)$ and another leaf c is given by

$$D_{(a-b)c} = \frac{1}{2}(D_{ac} + D_{bc}) \tag{3.1}$$

and from this, a new distance matrix is generated. After each iteration, the number of leaves in the tree is reduced by 1, and this is repeated until the remaining number of leaves in the trees becomes 3, at which point there is only one un-rooted tree topology remaining [33].

Distance-based methods have the advantage of being simple and computationally efficient, however, they are based on questionable assumptions since genetic mutations can be reversed over time, after which the genetic divergence is erased from the observable sequences. This is reflected in their performance, which is generally worse when compared to more complex character-based methods like maximum likelihood (ML) and Bayesian methods, which rely on probability-based algorithms based on a predefined model of sequence evolution [34].

The ML-based methods aim to maximize the probability of the observed data under a given substitution model. This is achieved in two steps, first the likelihood $L(\theta)$, where θ is an unknown parameter based on the parameters of the substitution model and the branch lengths in the tree, is maximized for each tree topology. The tree space is then searched for the topology that produced the highest likelihood, i.e. the tree that makes the observed data most likely to occur, which is then selected as the "correct" topology, though there is no guarantee that this accurately reflects the actual evolution [32, 35]. Mainly, if a substitution model that poorly reflects the evolution of the sequences in question is used, this can severely impact the performance of ML algorithms and lead to wrongful interpretations [36]. Furthermore, ML-based methods are generally based on the assumption that mutations at each site and lineage occur independently, meaning that the likelihood is the product of the probabilities of observing the data at each site, which is not necessarily true to the biological reality [37, 38].

Closely related to ML-based methods are the phylogenetic analysis methods based on Bayesian statistics. The Bayesian approach to phylogeny is based around the posterior probability P(T|D), that is, the probability that the tree topology T is correct given the observed data D, the prior probability P(T), and a likelihood function P(D|T), where

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

While the posterior probability is easy to formulate, it is almost impossible to compute analytically due to the high dimensionality of the trees and model parameters. For this reason, algorithms like Markov chain Monte Carlo are usually applied to approximate the posterior instead [39, 40].

3.3 Detection of horizontal gene transfer events from sequence data

As noted in chapter 1, horizontal gene transfer (HGT) has a very central role in the spread of antibiotic resistance. Therefore, when studying new types of ARGs it is important to identify genes that have undergone HGT, particularly those that have already moved into pathogens since these ARGs present a more immediate threat to human health [41]. The most common approach to identifying mobile genes is to search the genetic regions up and downstream of the gene in question for MGEs [42]. However, several computational methodologies have also been developed for the detection of gene transfer events directly from the sequence data. Broadly, these fall into two categories: parametric and phylogenetic methods [43].

The parametric approach to detecting HGT events aims to identify genetic regions that clearly deviate from the host genome average based on one or more features. Such a region is implied to be of foreign origin, i.e. it has been acquired through HGT. Commonly studied features include nucleotide composition, codon bias, and structural features [44]. These approaches are based on the notion that the features of the genome have been determined by evolutionary pressures specific to each species. This in turn has resulted in each species developing a recognizable genomic signature that newly acquired genes will not conform to [45]. Nucleotide composition is commonly expressed as GC-content (the proportion of the genomic DNA that corresponds to either guanine or cytosine), which in bacteria can vary between as low as $< 20\%$ to as high as $> 70\%$ [46]. However, this method has a rather low resolution, since while the difference in GC-content can be pronounced even between closely related species, it can also be very small between distantly related species. This leads to some HGT events being undetectable by focusing on GC-content alone. A higher resolution can be obtained by instead analyzing codon-frequencies,

since species can share a very similar nucleotide composition, but show clear differences in their preferred codon usage [47]. Codon bias is not as easily computed as the genomic GC-content but can be modeled by, for example, applying Markov chains [48].

The other approach for inferring HGT events is the phylogenetic approach. As the name suggests, these methods make use of phylogenetic trees and aim to identify anomalies in the tree that cannot be explained by vertical evolution. More specifically, for genes that have undergone HGT, the gene tree (describing the evolution of the genes) will not agree with the species tree (describing the evolution of the corresponding host species), but rather the transferred gene will appear as closely related to a gene from the donor species (Fig. 3.2). Since it is not feasible to build larger species trees from complete genomes, the species trees are typically reconstructed from well-conserved housekeeping or informational genes [49].

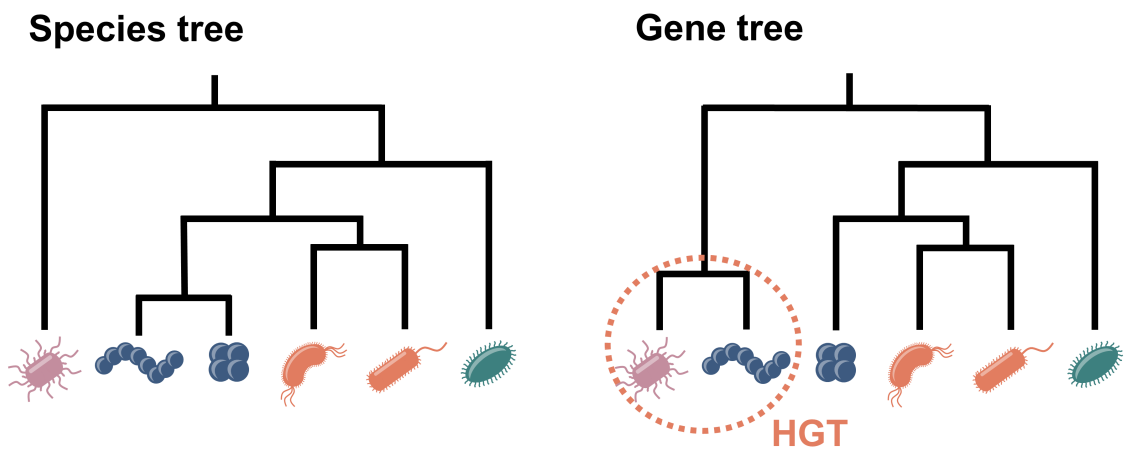

Figure 3.2: Illustration of a basic gene tree and corresponding species tree. The inconsistency between the gene tree and the species tree, which is encircled in the figure, can be inferred as a HGT event.

The phylogenetic methods can be further divided into two subcategories: explicit and implicit methods. Explicit phylogenetic methods make direct use of both gene and species trees [43]. One such method is to apply statistical tests to all of the sites in the two trees to identify significant disagreements between them, which are then interpreted as HGT events [50]. Explicit methods can quickly become very computationally expensive as the trees become larger, and to combat this one can opt to instead use an implicit phylogenetic method for inferring HGT events. While both approaches are based on the same basic principles, the implicit methods do not make direct use of a species tree. Instead, the species tree is implied from sequence similarity or measures of the evolutionary distance of the host species [51]. Transfer events between evolutionary

distant species can be inferred from the host taxonomy represented in the gene tree, but HGT between more similar species might not be immediately obvious [49]. A more sensitive measure of evolutionary distance between two taxa can be estimated from a pairwise sequence alignment using Maximum-Likelihood, and subjected to a statistical test (likelihood-ratio test) to see if the difference is significant enough for HGT to be invoked as the explanation for the observed gene tree [52].

4 Summary of results

This chapter provides a summary of the aims and findings of the three papers included in this thesis.

4.1 Papers I and II

Today, antibiotic resistance genes (ARGs) conferring resistance to all of the major classes of antibiotics used to treat infections have been detected in clinical pathogens. This issue is made worse by the fact that new ARGs keep moving into the clinic from external sources. Often, these are acquired through horizontal gene transfer (HGT) from harmless commensal or environmental bacteria, which are known to maintain a large diversity of ARGs [53]. Currently, the lack of knowledge about the resistome, the complete collection of ARGs carried by bacteria, makes it difficult to anticipate and manage new clinical ARGs. Indeed, new ARGs are usually discovered only after they have become widely disseminated among pathogens [54], at which point further spread is difficult to prevent. Papers I and II, therefore, aimed to expand the the knowledge about the resistome, and to demonstrate how large-scale computational screening can be used for early detection of new ARGs that have been acquired by pathogens through HGT before they spread widely.

In paper I, *Large-scale characterization of the macrolide resistome reveals high diversity and several new pathogen-associated genes*, we performed a systematic investigation of the macrolide resistome to characterize its size and diversity. We created and optimized profile hidden Markov models (HMMs) for identification of *erm* and *mph* macrolide resistance genes in genomic and metagenomic sequencing data, and used them to screen over 16TB of data for macrolide resistance genes. This yielded a diversity of *erm* and *mph* genes several times larger than previously described. Next, we aimed to elucidate the evolutionary

history of *erm* and *mph* genes through phylogenetic analysis. Our results show evidence that the mobile *erm* genes that are commonly carried by pathogens today all originate in species from the Firmicutes phylum, while *mph* genes have successfully mobilized from several phyla. From their placement in the respective phylogenetic trees, we identified five new *erm* genes and one new *mph* gene that were suggested to have moved into pathogens through HGT. Analysis of the genetic contexts of these six genes revealed the presence of mobile genetic elements (MGEs), providing further evidence that these are mobile ARGs that are likely to transfer into the clinic in the future. A total of ten predicted new resistance genes were selected for experimental validation through growth assays in *Escherichia coli*, where 70% of the tested genes were shown to confer increased resistance to macrolides in this host. This paper provides new insights into the evolutionary history of the macrolide resistome and reveals a set of new mobile macrolide resistance genes carried by pathogenic hosts.

In paper II, *Computational screening reveals previously undiscovered aminoglycoside resistance genes in human pathogens*, we further developed the methodology used in paper I by also evaluating the mobility of predicted genes on a large scale. Here, the target was the identification of new ARGs that show evidence of being emerging in clinical pathogens. To demonstrate this, we created and optimized profile HMMs for the identification of *aac* and *aph* aminoglycoside resistance genes, and used them to predict a massive collection of ARGs from ~ 1 million bacterial genomes. This included a previously unreported diversity of *aac* and *aph* genes carried by pathogenic species. Next, we retrieved the genetic regions directly up and downstream of all predicted ARGs and screened them for MGEs. This analysis revealed a total of 50 previously unknown resistance genes carried by pathogenic host species that were also found to co-localize with MGEs. Moreover, 21 of these ARGs were found in clinical isolates, showing that they have been able to move into clinical pathogens undetected. When expressed in *E. coli* and assessed through disk diffusion tests, 86% of the tested genes produced a resistant phenotype. The results from this paper demonstrate the usefulness of computational screening as a tool for identifying new ARGs as they are potentially emerging in clinical pathogens.

4.2 Paper III

The acquisition of foreign genetic material through HGT is one of the main ways through which antibiotic resistance is spreading. In a single transfer event, a cell can develop resistance to multiple antibiotics, which enables rapid evolution of multidrug-resistant pathogens under appropriate selection

pressure [55]. However, while the negative impact the HGT of ARGs has on human health is undeniable, much is still unknown about to what extent different factors affect the process itself. To overcome the threat posed by increasing antibiotic resistance levels, it is vital that we learn more about what influences bacteria to engage in HGT.

In paper III, *Factors influencing the horizontal gene transfer potential of antibiotic resistance genes*, we aimed to quantify how different factors influence the HGT of ARGs over large phylogenetic distances. We used fARGene to predict ARGs conferring resistance to aminoglycoside, beta-lactam, macrolide, quinolone, and tetracycline antibiotics, and from the predicted protein sequences we reconstructed phylogenetic trees representing each included gene class. Based on these trees, we detected a large set of events where ARGs had been transferred between species belonging to at least different taxonomic orders. For each of these events, features were collected representing the genetic compatibility of the ARG and genomes involved, as well as the estimated co-occurrence of donor and recipient genome(s) in different environments. These features were used as input to train random forest classifiers for the prediction of the HGT compatibility between bacterial genomes. Inference of these models revealed the genetic similarity between the acquired ARG and the recipient genome (based on comparison of *k*mers) to be the most influential feature. Here, ARGs that showed a similar nucleotide composition to the recipient genome showed a higher likelihood to be successfully transferred. By comparison, the co-occurrence between donor and recipient was suggested to play a less important role. These findings represent new insights into how different factors influence the HGT of ARGs.

5 Future work

The methodology applied in paper II was shown to be highly successful for the identification of new mobile ARGs in clinical pathogens. Therefore, this method can be expanded beyond aminoglycoside resistance genes, to identify new ARGs emerging in pathogens on a larger scale. Here, all gene models included in fARGene could be used, as well as additional models created to find genes not currently covered by fARGene. Some of this work is already planned, for example, an investigation of genes conferring resistance to colistin antibiotics will commence shortly.

Paper III represents a work in progress, and, consequently, there are improvements and additional implementations envisioned for the final paper. The random forest classifiers will be re-trained using an expanded null distribution to ensure the robustness of our findings. Moreover, a more thorough exploration of how the features impact the classifications will be undertaken to improve our understanding of this phenomenon. Ultimately, the aim is to then use the finalized classifiers for large-scale evaluation of the HGT compatibility of bacterial species carrying ARGs, to analyze the dissemination of ARGs.

Bibliography

[1] Matthew I Hutchings, Andrew W Truman, and Barrie Wilkinson. Antibiotics: past, present and future. *Current opinion in microbiology*, 51:72–80, 2019.

[2] Kathrin I Mohr. History of antibiotics research. *How to Overcome the Antibiotic Crisis*, pages 237–272, 2016.

[3] Anthony RM Coates, Gerry Halls, and Yanmin Hu. Novel classes of antibiotics or more of the same? *British journal of pharmacology*, 163(1): 184–194, 2011.

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

[5] Jose M Munita and Cesar A Arias. Mechanisms of antibiotic resistance. *Microbiology spectrum*, 4(2):4–2, 2016.

[6] Jessica Blair, Mark A Webber, Alison J Baylay, David O Ogbolu, and Laura JV Piddock. Molecular mechanisms of antibiotic resistance. *Nature reviews microbiology*, 13(1):42–51, 2015.

[7] Stefan Ebmeyer, Erik Kristiansson, and DG Larsson. A framework for identifying the recent origins of mobile antibiotic resistance genes. *Communications biology*, 4(1):1–10, 2021.

[8] Julie Perry, Nicholas Waglechner, and Gerard Wright. The prehistory of antibiotic resistance. *Cold Spring Harbor perspectives in medicine*, 6(6): a025197, 2016.

[9] Chandan Pal, Johan Bengtsson-Palme, Erik Kristiansson, and DG Larsson. The structure and diversity of human, animal and environmental resistomes. *Microbiome*, 4(1):1–15, 2016.

[10] Michael R Gillings. Evolutionary consequences of antibiotic use for the resistome, mobilome and microbial pangenome. *Frontiers in microbiology*, 4:4, 2013.

[11] Dongchang Sun, Katy Jeannot, Yonghong Xiao, and Charles W Knapp. Horizontal gene transfer mediated bacterial antibiotic resistance. *Frontiers in microbiology*, 10:1933, 2019.

[12] Laura S Frost, Raphael Leplae, Anne O Summers, and Ariane Toussaint. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722–732, 2005.

[13] Christopher M Thomas and Kaare M Nielsen. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews microbiology*, 3(9):711–721, 2005.

[14] Kristin Hegstad, Haima Mylvaganam, Jessin Janice, Ellen Josefsen, Audun Sivertsen, and Dagfinn Skaare. Role of horizontal gene transfer in the development of multidrug resistance in haemophilus influenzae. *Msphere*, 5(1):e00969–19, 2020.

[15] Alfonso J Alanis. Resistance to antibiotics: are we in the post-antibiotic era? *Archives of medical research*, 36(6):697–705, 2005.

[16] Robert C Moellering Jr. Ndm-1—a cause for worldwide concern. *New England Journal of Medicine*, 363(25):2377–2379, 2010.

[17] Huiyan Ye, Yihui Li, Zhencui Li, Rongsui Gao, Han Zhang, Ronghui Wen, George F Gao, Qinghua Hu, and Youjun Feng. Diversified mcr-1-harbouring plasmid reservoirs confer resistance to colistin in human gut microbiota. *MBio*, 7(2):e00177–16, 2016.

[18] Manish Boolchandani, Alaric W D'Souza, and Gautam Dantas. Sequencing-based methods and resources to study antimicrobial resistance. *Nature Reviews Genetics*, 20(6):356–370, 2019.

[19] Valeria Bortolaia, Rolf S Kaas, Etienne Ruppe, Marilyn C Roberts, Stefan Schwarz, Vincent Cattoir, Alain Philippon, Rosa L Allesoe, Ana Rita Rebelo, Alfred Ferrer Florensa, et al. Resfinder 4.0 for predictions of phenotypes from genotypes. *Journal of Antimicrobial Chemotherapy*, 75(12): 3491–3500, 2020.

[20] Brian P Alcock, William Huynh, Romeo Chalil, Keaton W Smith, Amogelang R Raphenya, Mateusz A Wlodarski, Arman Edalatmand, Aaron Petkau, Sohaib A Syed, Kara K Tsang, et al. Card 2023: expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 2022.

[21] Lubna Maryam, Salman Sadullah Usmani, and Gajendra PS Raghava. Computational resources in the management of antibiotic resistance: speeding up drug discovery. *Drug Discovery Today*, 26(9):2138–2151, 2021.

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

[23] Fanny Berglund, Tobias Österlund, Fredrik Boulund, Nachiket P Marathe, DG Larsson, and Erik Kristiansson. Identification and reconstruction of novel antibiotic resistance genes from metagenomes. *Microbiome*, 7(1): 1–14, 2019.

[24] Fredrik Boulund, Fanny Berglund, Carl-Fredrik Flach, Johan Bengtsson-Palme, Nachiket P Marathe, DG Larsson, and Erik Kristiansson. Computational discovery and functional validation of novel fluoroquinolone resistance genes in public metagenomic data sets. *BMC genomics*, 18(1): 1–9, 2017.

[25] Fanny Berglund, Nachiket P Marathe, Tobias Österlund, Johan Bengtsson-Palme, Stathis Kotsakis, Carl-Fredrik Flach, DG Larsson, and Erik Kristiansson. Identification of 76 novel b1 metallo-β-lactamases through large-scale screening of genomic and metagenomic data. *Microbiome*, 5(1):1–13, 2017.

[26] Fanny Berglund, Maria-Elisabeth Böhm, Anton Martinsson, Stefan Ebmeyer, Tobias Österlund, Anna Johnning, DG Joakim Larsson, and Erik Kristiansson. Comprehensive screening of genomic and metagenomic data reveals a large diversity of tetracycline resistance genes. *Microbial genomics*, 6(11), 2020.

[27] Gustavo Arango-Argoty, Emily Garner, Amy Pruden, Lenwood S Heath, Peter Vikesland, and Liqing Zhang. Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6(1):1–15, 2018.

[28] Etienne Ruppé, Amine Ghozlane, Julien Tap, Nicolas Pons, Anne-Sophie Alvarez, Nicolas Maziers, Trinidad Cuesta, Sara Hernando-Amado, Irene Clares, Jose Luís Martínez, et al. Prediction of the intestinal resistome by a three-dimensional structure-based method. *Nature microbiology*, 4(1): 112–123, 2019.

[29] Daniel H Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press, 2010.

[30] Mark Wilkinson, James O McInerney, Robert P Hirt, Peter G Foster, and T Martin Embley. Of clades and clans: terms for phylogenetic relationships in unrooted trees. *Trends in ecology & evolution*, 22(3):114–115, 2007.

[31] John P Huelsenbeck, Jonathan P Bollback, and Amy M Levine. Inferring the root of a phylogenetic tree. *Systematic biology*, 51(1):32–43, 2002.

[32] Paschalia Kapli, Ziheng Yang, and Maximilian J Telford. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, 21(7):428–444, 2020.

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

[34] Ziheng Yang and Bruce Rannala. Molecular phylogenetics: principles and practice. *Nature reviews genetics*, 13(5):303–314, 2012.

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

[36] E Michu. A short guide to phylogeny reconstruction. *Plant Soil and Environment*, 53(10):442, 2007.

[37] Joseph Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, 22(3):240–249, 1973.

[38] Mark E Siddall and Arnold G Kluge. Probabilism and phylogenetic inference. *Cladistics*, 13(4):313–336, 1997.

[39] John P Huelsenbeck, Fredrik Ronquist, Rasmus Nielsen, and Jonathan P Bollback. Bayesian inference of phylogeny and its impact on evolutionary biology. *science*, 294(5550):2310–2314, 2001.

[40] Mark Holder and Paul O Lewis. Phylogeny estimation: traditional and bayesian approaches. *Nature reviews genetics*, 4(4):275–284, 2003.

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

[42] Ross S McInnes, Gregory E McCallum, Lisa E Lamberte, and Willem van Schaik. Horizontal transfer of antibiotic resistance genes in the human gut microbiome. *Current opinion in microbiology*, 53:35–43, 2020.

[43] Matt Ravenhall, Nives Škunca, Florent Lassalle, and Christophe Dessimoz. Inferring horizontal gene transfer. *PLoS computational biology*, 11(5): e1004095, 2015.

[44] Jeffrey G Lawrence and Howard Ochman. Reconciling the many faces of lateral gene transfer. *TRENDS in Microbiology*, 10(1):1–4, 2002.

[45] Vincent Daubin, Emmanuelle Lerat, and Guy Perrière. The source of laterally transferred genes in bacterial genomes. *Genome biology*, 4(9):1–12, 2003.

[46] Falk Hildebrand, Axel Meyer, and Adam Eyre-Walker. Evidence of selection upon genomic gc-content in bacteria. *PLoS genetics*, 6(9):e1001107, 2010.

[47] Jeffrey G Lawrence and Howard Ochman. Molecular archaeology of the escherichia coli genome. *Proceedings of the National Academy of Sciences*, 95 (16):9413–9417, 1998.

[48] Diego Cortez, Patrick Forterre, and Simonetta Gribaldo. A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and orfans in archaeal and bacterial genomes. *Genome biology*, 10(6): 1–13, 2009.

[49] Shannon M Soucy, Jinling Huang, and Johann Peter Gogarten. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8): 472–482, 2015.

[50] Edward Susko. Tests for two trees using likelihood methods. *Molecular Biology and Evolution*, 31(4):1029–1039, 2014.

[51] David Schaller, Manuel Lafond, Peter F Stadler, Nicolas Wieseke, and Marc Hellmuth. Indirect identification of horizontal gene transfer. *Journal of Mathematical Biology*, 83(1):1–73, 2021.

[52] Christophe Dessimoz, Daniel Margadant, and Gaston H Gonnet. Dlight–lateral gene transfer detection using pairwise evolutionary distances in a statistical framework. In *Annual International Conference on Research in Computational Molecular Biology*, pages 315–330. Springer, 2008.

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

[54] Maria-Elisabeth Böhm, Mohammad Razavi, Nachiket P Marathe, Carl-Fredrik Flach, and DG Larsson. Discovery of a novel integron-borne aminoglycoside resistance gene present in clinical pathogens by screening environmental bacterial communities. *Microbiome*, 8(1):1–11, 2020.

[55] Michael N Alekshun and Stuart B Levy. Molecular mechanisms of antibacterial multidrug resistance. *Cell*, 128(6):1037–1050, 2007.