THESIS FOR THE DEGREE OF LICENTIATE OF PHILOSOPHY

Data-Driven Methods for Surveillance and Diagnostics of Antibiotic-Resistant Bacteria

Juan Salvador Inda Díaz

UNIVERSITY OF GOTHENBURG

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

Data-Driven Methods for Surveillance and Diagnostics of Antibiotic-Resistant Bacteria
Juan Salvador Inda Díaz
Göteborg 2022

Division of Applied Mathematics and Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg
Sweden
Telephone +46 (0)317 72 53 15

Typeset with LaTeX
Printed by Stema Specialtryck AB, Borås, Sweden, 2022

Data-Driven Methods for Surveillance and Diagnostics of Antibiotic-Resistant Bacteria

Juan Salvador Inda Díaz

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

Abstract

Antimicrobial resistance is a rapidly growing challenge for the healthcare sector and multi-drug resistant bacterial infections are the cause of nearly a million deaths annually worldwide. Antibiotic resistance is conferred by either antibiotic resistance genes (ARGs) which can be acquired by horizontal gene transfer between bacterial cells or by mutations in pre-existing DNA. Large collections of ARGs are present in the bacterial communities hosted by humans and animals, and in the external environments. Most of these ARGs are uncharacterized and not well-studied. Furthermore, antimicrobial resistance has significantly hindered our ability to treat infections and novel diagnostic solutions are therefore needed to ensure efficient treatment. In paper I, the abundance and diversity of 24,074 ARGs of 17 classes were studied in metagenomic data. The majority of the ARGs were previously uncharacterized, of which several were commonly reoccurring and shared across the digestive system of humans and animals, suggesting that they are under strong selection pressures. The data-driven work in this paper showed that the analysis of all ARGs, including those that have previously not been described, is necessary to provide a comprehensive description of the resistance potential of bacterial communities. In paper II, an AI method for the prediction of bacterial susceptibility towards antibiotics is presented. The method is based on transformers and artificial neural networks and exploits the strong and highly non-trivial dependencies present in the resistance patterns of bacteria. The model was highly successful in predicting susceptibility for most antibiotics from the classes cephalosporins and quinolones but had a lower performance on penicillins and aminoglycosides. The AI-based methodology described in this paper may be used to improve the diagnostics chain of infectious diseases with the potential to reduce the morbidity and mortality of patients. This thesis provides methodologies for improved surveillance and diagnostics of antibiotic-resistant bacteria and, thereby, contributes to a more sustainable use of antibiotics.

Keywords: antibiotic resistance, metagenomics, data-driven diagnostics, transformers

List of publications

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

I. **Inda-Díaz, J.S.**, Johnning, A., Bengtsson-Palme, J., Parras Moltó, M., Kristiansson, E. (2022). The abundance and diversity of the latent resistome. Manuscript.

II. **Inda-Díaz, J.S.**, Johnning, A., Svensson, L., Parras-Moltó, M., Kristiansson, E. (2022). Antibiotic susceptibility prediction using transformers. Manuscript.

Additional papers not included in this thesis:

III. Diamanti, K., **Inda Díaz, J.S.**, Raine, A., Pan, G., Wadelius, C., and Cavalli, M. (2021) Single nucleus transcriptomics data integration recapitulates the major cell types in human liver. *Hepatol Res*, 51: 233– 238.

IV. Gustafsson, J., Robinson, J., **Inda-Díaz, J.S.**, Björnson, E., Jörnsten, R, Nielsen, J. (2020) DSAVE: Detection of misclassified cells in single-cell RNA-Seq data. *PLOS ONE*, 15(12): e0243360.

Author contributions

I. Participated in study design, implemented the analysis pipeline and retrieved and filtered the metagenomic data from MGnify and ENA. Performed the clustering of ARGs, the alignments between fARGene and ResFinder genes, and between genes and metagenomes. Estimated the abundance and diversity for all metagenomes, as well as the pan-resistome and core-resistomes for each environment. Carried out the PCA of the abundance and diversity of ARGs. Drew all the figures, drafted and edited the manuscript.

II. Participated in the study design, collected and parsed the data, and participated in conceptualizing the model's architecture. Also implemented, trained, and tested the model. Drew all the figures, drafted and edited the manuscript.

Acknowledgments

I would like to express my gratitude to my supervisor Erik Kristiansson for his guidance and trust, that I deeply appreciate. Thank you Anna Johnning for being my co-supervisor, for your wonderful feedback and conversation, research related or not. Both of you have a joy for research that is contagious, and have been a great support and motivation during this time.

Thank you to all the people I have collaborated with at GU, Chalmers and Uppsala University. Johan Bengtsson-Palme for your formidable input on antibiotic resistance, Lennart Svensson for your inspirational deep learning lectures and discussions, and Marcos Parras for the time we shared in Gothenburg and the great collaboration. Thank you, José Sánchez, for your advice and all the work we have done together. Thank you Marcos Cavalli and Klev Dimitri, it has always been a pleasure to work with you.

Thank you Rebecka Jörnsten for welcoming me to the Department of Mathematical Sciences. Aila Särkkä and Marija Cvijovic, thank you for all the encouragement and the attention I have received from you. Thank you, Annika Lang and Serik Sagitov, for your valuable guidance.

Thank you to all my current colleagues at Mathematical Sciences for all the fun moments, foremost to Felix for being such a great friend and roommate. Special mention to Gabrijela, Barbara, Helga, Linnea, Malin, David and Oskar. Thank also you to my former colleagues, Fanny, Anna, Malin and Jonatan for showing us the way.

My time as a PhD student at Mathematical Sciences would not be the same without the help from all the technicians and administrators, you make the Department such an enjoyable working place.

Thank you to the present and former members of Erik Kristiansson's group, especially Fanny Berglund, Anna Johnning, Anna Rehammar, Astrid von Mentzer, Marcos Parras, David Lund, Mikael Gustavsson, Patrik Svedberg, and of course, Erik Kristiansson and Anna Johnning. You are a fantastic team.

Endless admiration, gratitude and love to my family.

Contents

1 Background

One of the most important advances in human health was the discovery of antibiotics and their introduction as a treatment for infectious diseases. During the first half of the 20th century, which is sometimes known as the golden age of antibiotics, several classes of antibiotics, with different targets and mechanisms, were discovered, massively produced and effectively employed to cure diseases of epidemic proportion (Mohr, 2016), having a great impact on life expectancy (Hutchings et al., 2019). Originally, antibiotics were chemical or biological compounds that were naturally secreted by fungi or bacteria, such as penicillin produced by the fungus *Penicillium*. More recently, antibiotics have also been synthetically produced (Hutchings et al., 2019), such as nalidixic acid, the first quinolone employed and introduced in the 1960's (Andersson and MacGowan, 2003). The rapid increase in the use of antibiotics was accompanied by the adaptation of bacteria to thrive in the presence of these compounds. Indeed, bacteria have developed and acquired a diverse set of molecular mechanisms to overcome antibiotics. The great development of antibiotics, was therefore, followed by the growing threat of antibiotic resistance. The main objectives in this thesis are the identification and localization of resistant mechanisms in the environment and the prediction of the resistance profiles in bacterial infections for personalized treatment.

1.1 Antibiotic resistance

Bacteria typically become resistant to antibiotics through mutations in their genome (Woodford and Ellington, 2007) and the acquisition of antibiotic resistance genes (ARGs) through the process of horizontal gene transfer (HGT)(Blair et al., 2015). Horizontal gene transfer is the ("lateral") flow of genetic material between organisms not conditioned to reproduction ("vertical") (Burmeister, 2015). Antibiotic resistance genes provide the bacteria with the tools to de-

grade, eject or modify antibiotics or to change their target in the cell, so they become less harmful to the host (Peterson and Kaur, 2018). Although ARGs are naturally present in microbial communities, the overuse of antibiotics by humans has created a selection pressure that enables the promotion of ARGs in pathogens, a process happening in both, external environmental and host-associated bacterial communities. In pathogenic bacteria, ARGs are most often found on mobile genetic elements (MGEs), e.g. transposons or conjugative elements such as plasmids (Rodríguez-Beltrán et al., 2021), which eases and facilitates HGT. Bacteria can be resistant to multiple antibiotics due to the accumulation of several ARGs or by having multi-drug ARGs (Nikaido, 2009). Several ARGs can be co-localized on one MGE, and a bacterium recruiting that MGE could become resistant to multiple antibiotics by acquiring a single MGE (Paulsen et al., 1996; Botts et al., 2017). Multi-drug resistance can also be caused by a combination of ARGs co-localized on one MGE, multiple MGEs carrying ARGs, multi-drug ARGs and chromosomal ARGs.

The origins of most of the well-characterized and clinically relevant ARGs identified to this date are unknown and only a small fraction of them has their origin in human pathogens and commensals (Ebmeyer et al., 2021b). This suggests that the vast majority of ARGs originates in environmental bacteria (Allen et al., 2010; Bengtsson-Palme et al., 2017). In order to assess the threat that ARGs pose for humans, we need to investigate where ARGs are found and how they are transferred into pathogens. We should be aware that it is difficult, if not impossible, to have a compilation of all ARGs present in nature. This is due to the large biodiversity with over 10^{12} microbial species estimated on earth (Locey and Lennon, 2016) compared to only 1.5 million species that have been whole-genome sequenced (Schloss et al., 2016). Furthermore, it is unfeasible to do functional studies of all known bacterial sequences, and new genes conferring resistance can arise at any time, e.g. through mutations conferring resistance.

Public ARG databases, such as ResFinder (Zankari et al., 2012) and CARD (McArthur et al., 2013), contain collections of ARGs and often their associated phenotype. ResFinder focuses on mobile ARGs that have been detected within MGE, many of which are clinically relevant and found in pathogens. In addition, computational methods have been developed and employed recently to predict ARGs from sequence data available in public repositories (Berglund et al., 2019; Arango-Argoty et al., 2018; Ruppé et al., 2019), creating a vast list of predicted ARGs. One of them, fARGene, uses optimized Hidden Markov Models, has been shown to have a high sensitivity and specificity, and has contributed to a significant repertoire of resistant genes (Berglund et al., 2019, 2017; Lund et al., 2021; Berglund et al., 2020; Boulund et al., 2017). Most of these genes are latently present in bacterial communities but do not yet constitute a

clinical problem.

1.2 Diagnostics

The recovery from a bacterial infection depends on a prompt prescription of an antibiotic that is efficient against the infecting bacteria. Any delay in the treatment can have a negative impact on the morbidity and mortality of bacterial infections (Friedman et al., 2016). Furthermore, the lag in administering a correct treatment for an infection is also linked to high societal and economical costs (European Centre for Disease Prevention and Control and European Medicines Agency, 2009). Although antibiotics have been used as the main treatment for bacterial infections since the 20th century, their effectiveness has been compromised by the increasing antibiotic resistance in bacteria.

It is common practice to assess the antibiotic susceptibility of the bacteria causing the infection prior to the prescription of a drug. The results from these diagnostic tests provide medical doctors with crucial information for the selection of antibiotic(s) to treat an infection. Diagnostics is thus both important to cure infections but also to limit the spread of pathogens and reduce the consumption of antibiotics.

Antibiotic susceptibility testing is a diagnostic method that aim to detect resistance and quantify the degree of susceptibility to antimicrobial agents (Jorgensen and Ferraro, 1998). The testing is most often performed using cultivation-based techniques, such as disc diffusion, gradient diffusion and broth micro-dilutions tests(Reller et al., 2009). In those tests, the bacterium is grown in the presence of different antibiotics at different concentrations. The bacterium is classified as resistant, intermediate or susceptible to an antibiotic depending on its growth under the different conditions. Although cultivation-based techniques are accurate, they are also time-consuming and depend on the growth rate of each bacterium. Often, antibiotics are tested sequentially in order to reduce costs and save resources. Consequently, if an appropriate antibiotic is not found in an early stage, the correct treatment could be significantly delayed. The increment of antibiotic resistance and multi-drug resistant bacteria in the later years, together with the duration of antibiotic susceptibility testing and the limited time and resources pose a growing challenge for the healthcare. Therefore, innovative solutions that are able to provide diagnostic information in the early stages are needed.

1.3 Metagenomics

Metagenomics is the high-throughput sequencing of the genetic material present in an environment and is employed to identify the taxonomic and functional composition of microbial communities (Fricke et al., 2011). In addition, metagenomics allows us to study bacterial populations directly from their source, without the need to isolate individual bacteria or create cultures: all genetic material is potentially sequenced at once (Schloss and Handelsman, 2005).

The information contained in metagenomic sequences enables us to study the ARGs present specific environments (Garmendia et al., 2012). Using alignment tools, e.g. BLAST (Altschul et al., 1990) and DIAMOND (Buchfink et al., 2015), we can search for DNA sequences of ARGs within metagenomes. Thereby, we can establish the diversity of ARGs carried by bacteria in a sample, i.e. how many different genes are found in a sample, and their abundance, i.e. how much genetic material from each gene is found. Using public metagenomic repositories, such as the European Nucleotide Archive (Harrison et al., 2021) and MGnify (Mitchell et al., 2019), we can investigate the resistomes, such as the collection of ARGs (Kim and Cha, 2021; Allen et al., 2010), of environmental and host-associated microbiomes. Moreover, for each environment (marine water, soil, human gut, etc.) we can identify its pan-resistomes, i.e. the collection of all ARGs present in any metagenome from the investigated environment, and its core-resistome, i.e. ARGs commonly present in the metagenomes of that environment. The abundance and diversity, together with the pan- and core-resistome form a valuable source of information for studies on the origin of ARGs and the selection pressures governing the spread of ARGs across environments and between environmental, commensal and pathogenic bacteria.

The ease and costs linked to next-generation sequencing have promoted a significant increase of data in the last years (Keegan et al., 2016). The challenge has therefore shifted to finding efficient ways to store and access the data, as well as developing computational and analytical tools to explore it. For reference, 150TB of data containing 22,272 metagenomes and 4×10^{11} reads served as the initial data set for the first paper in this thesis. The complexity of the data is not limited to storage needs, the data is high-dimensional, sparse, and contains large technical and biological variability. Nevertheless, a wide range of tools and techniques have been developed to work with genomic and metagenomic data. These include among others, alignment tools, normalization techniques, and a mix of unsupervised and supervised methods that all together provide reliable statistically valid results.

2 Summary of results

This chapter provides a summary of the overall aims and findings from the two papers included in the thesis, to facilitate the understanding of their contribution to the research field.

2.1 Paper I

In paper I, we analyzed the abundance and diversity of antibiotic resistance genes in different microbial environments. We considered well-characterized and clinically relevant genes ("established" ARGs) as well as putative resistance genes that are novel and have not yet been thoroughly characterized ("latent" ARGs). The established genes have been more thoroughly studied while the latent genes have so far been overlooked in literature. The resistome we built composed of both established and latents ARGs of different gene classes is, therefore, more comprehensive. Additionally, we described the pan- and core-resistomes of external and host-associated microbial environments depicting the selection pressures present in each environment that contribute to the fixation of established and latent ARGs in these communities. Furthermore, we found associations of latent ARGs between the resistomes of the environments and evaluated the risk of their transmission to pathogens, and thus, their potential impact on human health. Previous resistome studies have only focused on established ARGs, whilst we focused also on the latent genes, providing new insights into the presence of ARGs in nature.

Our results, based on the analysis of 10,744 metagenomic samples from 20 environmental types, included the abundance and diversity of 572 established and 23,502 latent ARGs of 17 gene classes (six aminoglycoside, five β-lactam, two macrolide, one quinolone and three tetracycline resistance gene classes). We showed that latent ARGs indeed constitute a large part of the resistome.

The pan-resistomes of both external and host-associated environments were composed mainly of latent genes (91%-98% and 71%-90%, respectively). Additionally, 40%-73% of the core-resistomes of host-associated environments were latent genes. The pan-resistomes of external environments were on average larger than the host-associated environments, representing a wider diversity and reservoir of ARGs.

We found that the size of the core-resistome of external environments was significantly smaller than host-associated environments. This suggests that the selection pressures in external environments may not be sufficiently strong to select for specific gene variants to be fixated in these environments. Another explanation is the between-sample variability, which is higher in the external environments, and where taxonomic distribution and environmental factors may limit the promotion of specific resistance genes variants. Interestingly, the core-resistomes in human and animal digestive systems and wastewater were extensive, had a large degree of similarity, indicating that these environments have common selection pressures that are acting on individual gene variants.

Furthermore, the abundance and diversity of each ARG class differed between environments, and the composition of the pan-resistome contrasted significantly from that of the core-resistome, suggesting that selection pressures are acting differently on each ARG class. Finally, our analysis suggested that the wastewater bacterial communities are hot-spots for the mobilization of latent ARGs, as they satisfy both criteria to make them high-risk environments: a large and diverse pan-resistome of latent ARGs as well as a high abundance of established mobile ARGs suggesting the presence of mobile genetic elements necessary for mobilization.

Our results indicate that the latent part of the resistome is present in both external and host-associated environments. Latent ARGs dominated the pan-resistomes, and several of the latent ARGs were commonly reoccurring in human- and animal-associated metagenomes. Thus, we argue that latent ARGs should play a more relevant role in future resistome studies.

2.2 Paper II

Diagnostics of bacteria before treatment, e.g. susceptibility testing, have been implemented in hospitals due to, among others, the increment of antibiotic resistance in pathogens. Although diagnostics are vital to finding the optimal treatment, current methods can take several days, time that can be crucial for the patient. Moreover, when no diagnostics are available, treatment is based on

medical doctors' educated guesses, often failing to achieve the desired effect. Faster and precise diagnostic methods are needed in order to reduce morbidity and mortality of patients, to prevent the spread of bacterial infections, and lower the societal costs related to them. In paper II, we build and train a deep learning model to do predictions of susceptibility and resistance of bacterial isolates based on patient data and partial diagnostics information.

We investigated strong and highly non-trivial resistant dependencies present in 9,224,373 antibiotic susceptibility tests done between 2013 and 2017 on 261,378 *Escherichia coli* isolates from 30 European countries retrieved from The European Surveillance System. The susceptibility tests corresponded to a minimum of seven and a maximum of sixteen antibiotics of four classes (five penicillins, five quinolones, four cephalosporins and two aminoglycosides). For each bacterial isolate, the age, country and gender of the patient where the isolate was taken from and the date were also available in the data. A made-up example of the final information for one isolate is the sentence: "SV 30 M 2013_01 LVX_R AMC_S CAZ_S AMP_S CIP_S CTX_S GEN_S TZP_R", representing a bacterium isolated at a hospital from Sweden (SV), from a 30 years old male patient in January of 2013. The isolate was resistant to levofloxacin (LVX) and piperacillin/tazobactam (TZP), and susceptible to amoxicillin/clavulanic acid (AMC), ceftazidime (CAZ), ampicillin (AMP), ciprofloxacin (CIP), cefotaxime (CTX) and gentamicin (GEN).

The deep learning model presented is a combination of transformers and artificial neural networks. We trained the model using the resistance profiles of six out of sixteen different antibiotics, age, country and gender of the patient and date, as input to predict the resistance profile of the antibiotics that were masked and not used as input. We evaluated the performance of the model by determining the major error (ME) rates, i.e. the proportion of true susceptible isolates that are predicted as being resistant, and the very major error (VME) rates, i.e. the proportion of true resistant isolates that are predicted as being susceptible. We obtained ME and VME rates for each antibiotic and different number of input information in the test data set. We analyzed the performance of the model both when used to predict masked data and when used as an autoencoder to reconstruct the input data. Furthermore, we examined the distribution of the decision scores for the correct predictions, the MEs and the VMEs.

The model showed high performance for the prediction of resistance (low VME rates) and susceptibility (low ME rates) for cephalosporins and the majority of quinolones, but a lower performance for penicillins and aminoglycosides. In general, the model showed a better performance when predicting susceptible compared to resistant bacterial isolates, i.e. lower ME rates compared to

VME rates. Similar results were observed when the model was evaluated as an autoencoder. Although this is ongoing research, we have shown that artificial intelligence models can be used to predict diagnostic tests results. The implementation of this model can provide information about resistance to the physician in an earlier phase and suggest proper treatments, thereby, benefiting the life quality of the patients and improving the healthcare system.

3 Future work

For each of the latent antibiotic resistance genes present in the core-resistomes from paper I, we would like to determine if these genes are already present on mobile genetic elements. In order to achieve that, we intent to annotate their genetic context using GEnView Ebmeyer et al. (2021a) and several databases containing sequence information of commonly occurring MGEs.

The development and evaluation of the model proposed in paper II is an ongoing work and planned actions include a more systematic evaluation of the model's topology and hyper-parameters. Examples of that are the network and embedding sizes, and implementing different embeddings for the patient data and the isolate susceptibility data. We would also like to implement individual losses for each antibiotic, allowing us to weight the major errors and the very major errors separately for each antibiotic. Furthermore, we would also like to use a variable number of antibiotics as input to train the model. Additionally, we would like to implement conformal prediction as a measure of certainty for the predictions Vovk et al. (2005); Papadopoulos (2008).

Bibliography

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

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Andersson, M. I. and MacGowan, A. P. (2003). Development of the quinolones. *J Antimicrob Chemother*, 51 Suppl 1:1–11.

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

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

Berglund, F., Böhm, M.-E., Martinsson, A., Ebmeyer, S., Österlund, T., Johnning, A., Larsson, D. G. J., and Kristiansson, E. (2020). Comprehensive screening of genomic and metagenomic data reveals a large diversity of tetracycline resistance genes. *Microbial Genomics*, 6(11).

Berglund, F., Marathe, N. P., Österlund, T., Bengtsson-Palme, J., Kotsakis, S., Flach, C.-F., Larsson, D. J., and Kristiansson, E. (2017). Identification of 76 novel b1 metallo-β-lactamases through large-scale screening of genomic and metagenomic data. *Microbiome*, 5(1):1–13.

Berglund, F., Österlund, T., Boulund, F., Marathe, N. P., Larsson, D. G. J., and Kristiansson, E. (2019). Identification and reconstruction of novel antibiotic resistance genes from metagenomes. *Microbiome*, 7(1):52.

Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O., and Piddock, L. J. V. (2015). Molecular mechanisms of antibiotic resistance. *Nature Reviews Microbiology*, 13(1):42–51.

Botts, R. T., Apffel, B. A., Walters, C. J., Davidson, K. E., Echols, R. S., Geiger, M. R., Guzman, V. L., Haase, V. S., Montana, M. A., La Chat, C. A., Mielke, J. A., Mullen, K. L., Virtue, C. C., Brown, C. J., Top, E. M., and Cummings, D. E. (2017). Characterization of four multidrug resistance plasmids captured from the sediments of an urban coastal wetland. *Frontiers in microbiology*, 8:1922–1922. Publisher: Frontiers Media S.A.

Boulund, F., Berglund, F., Flach, C.-F., Bengtsson-Palme, J., Marathe, N. P., Larsson, D. J., and Kristiansson, E. (2017). Computational discovery and functional validation of novel fluoroquinolone resistance genes in public metagenomic data sets. *BMC genomics*, 18(1):1–9.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60.

Burmeister, A. R. (2015). Horizontal Gene Transfer. *Evol Med Public Health*, 2015(1):193–194.

Ebmeyer, S., Coertze, R. D., Berglund, F., Kristiansson, E., and Larsson, D. G. J. (2021a). GEnView: a gene-centric, phylogeny-based comparative genomics pipeline for bacterial genomes and plasmids. *Bioinformatics*.

Ebmeyer, S., Kristiansson, E., and Larsson, D. J. (2021b). A framework for identifying the recent origins of mobile antibiotic resistance genes. *Communications Biology*, 4(1):1–10.

European Centre for Disease Prevention and Control and European Medicines Agency (2009). *The bacterial challenge : time to react : a call to narrow the gap between multidrug-resistant bacteria in the EU and the development of new antibacterial agents*. Publications Office.

Fricke, W. F., Cebula, T. A., and Ravel, J. (2011). Chapter 28 - genomics. In Budowle, B., Schutzer, S. E., Breeze, R. G., Keim, P. S., and Morse, S. A., editors, *Microbial Forensics (Second Edition)*, pages 479–492. Academic Press, San Diego, second edition edition.

Friedman, N. D., Temkin, E., and Carmeli, Y. (2016). The negative impact of antibiotic resistance. *Clin Microbiol Infect*, 22(5):416–422.

Garmendia, L., Hernandez, A., Sanchez, M. B., and Martinez, J. L. (2012). Metagenomics and antibiotics. *Clin Microbiol Infect*, 18 Suppl 4:27–31.

Harrison, P. W., Ahamed, A., Aslam, R., Alako, B. T. F., Burgin, J., Buso, N., Courtot, M., Fan, J., Gupta, D., Haseeb, M., Holt, S., Ibrahim, T., Ivanov, E., Jayathilaka, S., Balavenkataraman Kadhirvelu, V., Kumar, M., Lopez, R., Kay, S., Leinonen, R., Liu, X., O'Cathail, C., Pakseresht, A., Park, Y., Pesant, S., Rahman, N., Rajan, J., Sokolov, A., Vijayaraja, S., Waheed, Z., Zyoud, A., Burdett, T., and Cochrane, G. (2021). The European Nucleotide Archive in 2020. *Nucleic Acids Res*, 49(D1):D82–D85.

Hutchings, M. I., Truman, A. W., and Wilkinson, B. (2019). Antibiotics: past, present and future. *Current Opinion in Microbiology*, 51:72–80. Antimicrobials.

Jorgensen, J. H. and Ferraro, M. J. (1998). Antimicrobial susceptibility testing: general principles and contemporary practices. *Clin Infect Dis*, 26(4):973–980.

Keegan, K. P., Glass, E. M., and Meyer, F. (2016). MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. *Methods Mol Biol*, 1399:207–233.

Kim, D.-W. and Cha, C.-J. (2021). Antibiotic resistome from the one-health perspective: understanding and controlling antimicrobial resistance transmission. *Experimental & Molecular Medicine*, 53(3):301–309.

Locey, K. J. and Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A*, 113(21):5970–5975.

Lund, D., Kieffer, N., Parras-Moltó, M., Ebmeyer, S., Berglund, F., Johnning, A., Larsson, J. D., and Kristiansson, E. (2021). Large-scale characterization of the macrolide resistome reveals high diversity and several new pathogen-associated genes. *Submitted*.

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., Bhullar, K., Canova, M. J., De Pascale, G., Ejim, L., Kalan, L., King, A. M., Koteva, K., Morar, M., Mulvey, M. R., O'Brien, J. S., Pawlowski, A. C., Piddock, L. J. V., Spanogiannopoulos, P., Sutherland, A. D., Tang, I., Taylor, P. L., Thaker, M., Wang, W., Yan, M., Yu, T., and Wright, G. D. (2013). The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy*, 57(7):3348–3357. Edition: 2013/05/06 Publisher: American Society for Microbiology.

Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., and Finn, R. D. (2019). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research*, 48(D1):D570–D578.

Mohr, K. I. (2016). History of Antibiotics Research. *Curr Top Microbiol Immunol*, 398:237–272.

Nikaido, H. (2009). Multidrug resistance in bacteria. *Annual review of biochemistry*, 78:119–146.

Papadopoulos, H. (2008). Inductive conformal prediction: Theory and application to neural networks. In *Tools in Artificial Intelligence*, chapter 18. IntechOpen, Rijeka.

Paulsen, I. T., Brown, M. H., and Skurray, R. A. (1996). Proton-dependent multidrug efflux systems. *Microbiological Reviews*, 60(4):575–608.

Peterson, E. and Kaur, P. (2018). Antibiotic resistance mechanisms in bacteria: Relationships between resistance determinants of antibiotic producers, environmental bacteria, and clinical pathogens. *Frontiers in Microbiology*, 9:2928.

Reller, L. B., Weinstein, M., Jorgensen, J. H., and Ferraro, M. J. (2009). Antimicrobial Susceptibility Testing: A Review of General Principles and Contemporary Practices. *Clinical Infectious Diseases*, 49(11):1749–1755.

Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R. C., and San Millán, Á. (2021). Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nature Reviews Microbiology*, pages 1–13.

Ruppé, E., Ghozlane, A., Tap, J., Pons, N., Alvarez, A.-S., Maziers, N., Cuesta, T., Hernando-Amado, S., Clares, I., Martínez, J. L., Coque, T. M., Baquero, F., Lanza, V. F., Máiz, L., Goulenok, T., de Lastours, V., Amor, N., Fantin, B., Wieder, I., Andremont, A., van Schaik, W., Rogers, M., Zhang, X., Willems, R. J. L., de Brevern, A. G., Batto, J.-M., Blottière, H. M., Léonard, P., Léjard, V., Letur, A., Levenez, F., Weiszer, K., Haimet, F., Doré, J., Kennedy, S. P., and Ehrlich, S. D. (2019). Prediction of the intestinal resistome by a three-dimensional structure-based method. *Nature Microbiology*, 4(1):112–123.

Schloss, P. D., Girard, R. A., Martin, T., Edwards, J., and Thrash, J. C. (2016). Status of the Archaeal and Bacterial Census: an Update. *mBio*, 7(3).

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

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg.

Woodford, N. and Ellington, M. J. (2007). The emergence of antibiotic resistance by mutation. *Clin Microbiol Infect*, 13(1):5–18.

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