TU Delft

Delft University of Technology

The Microbial World Magnified Through the Bright Lens of Comparative Genomics

Urhan, A.

DOI
10.4233/uuid:bc65313a-c2f1-42d5-ac43-42d247d4e531

Publication date
2024

Document Version
Final published version

Citation (APA)
Urhan, A. (2024). *The Microbial World Magnified Through the Bright Lens of Comparative Genomics*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:bc65313a-c2f1-42d5-ac43-42d247d4e531

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

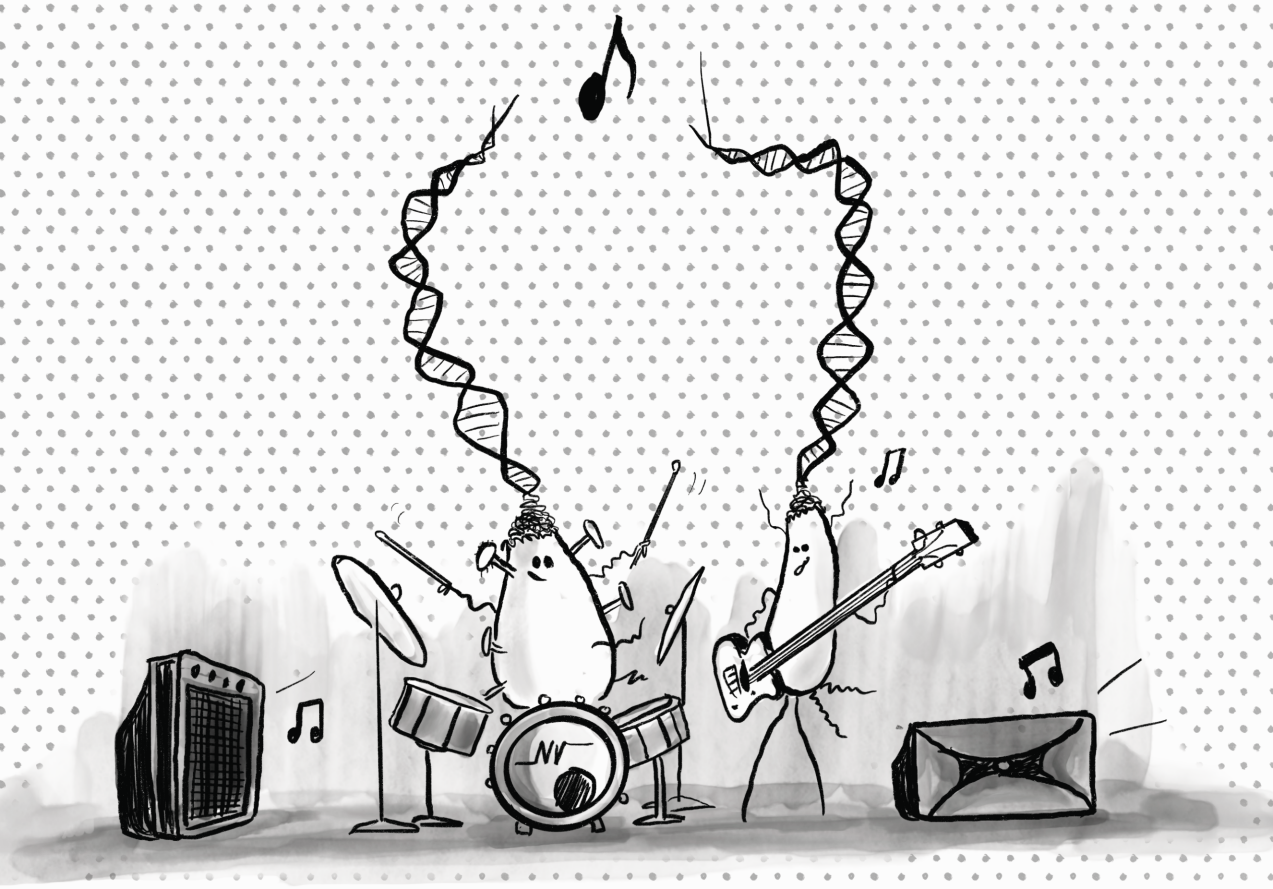
Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

The Microbial World

Magnified Through the Bright Lens of Comparative Genomics



Aysun Urhan

The Microbial World Magnified Through the Bright Lens of Comparative Genomics

The Microbial World Magnified Through the Bright Lens of Comparative Genomics

Dissertation

for the purpose of attaining the degree of doctor
at the Delft University of Technology,
by the authority of the Rector Magnificus, prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates,
to be defended publicly on
Monday, 21 October 2024, at 10:00.

by

Aysun URHAN

Master of Science in Chemical Engineering
Bogazici University, Istanbul, Turkey
born on 25 October 1994 in Istanbul, Turkey.

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. M.J.T. Reinders,	Delft University of Technology, *promotor*
Dr. T.E.P.M.F. Abeel,	Delft University of Technology, *promotor*
Independent members:	
Prof. dr. ir. M.K. de Kreuk,	Delft University of Technology
Dr. A. Özgür,	Bogazici University
Dr. A.C. Schürch,	UMC Utrecht
Prof. dr. M. Suarez Diez,	Wageningen University
Dr. G.F. Zeller,	Leiden University Medical Center
Prof. dr. M.M. de Weerdt,	Delft University of Technology, *reserve member*

TU Delft Delft University of Technology

Cover design by Narasimha Vedala | Diskrypt | https://tinyurl.com/The-Gaia-Project
Printed by

ISBN ...

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

To Dad, Mom and the music
guiding me from birth till the end.

Contents

Summary xi

Samenvatting xiii

Preface 1

1 Introduction 3
- 1.1 Microbes: wherever we may roam . . . 4
  - 1.1.1 Viruses . . . 5
  - 1.1.2 Bacteria . . . 9
- 1.2 Computational comparative genomics to magnify the microbial world . . . 14
- 1.3 Research objectives and contributions . . . 18

2 Emergence of novel SARS-CoV-2 variants in the Netherlands 31
- 2.1 Introduction . . . 32
- 2.2 Methods . . . 33
  - 2.2.1 Data retrieval and preprocessing, and multiple sequence alignment 33
  - 2.2.2 Sequence variation analysis . . . 34
  - 2.2.3 Phylogenetic tree construction . . . 34
- 2.3 Results . . . 34
  - 2.3.1 Distinct genetic patterns in the SARS-CoV-2 population emerge across the globe . . . 34
  - 2.3.2 Evolution of the SARS-CoV-2 genome and increased mutation frequency in hotspot regions . . . 37
  - 2.3.3 Population of SARS-CoV-2 is dominated by four mutations globally while emergence of locally distinct variants indicates local outbreaks . . . 39
  - 2.3.4 Introduction of COVID-19 in the Netherlands and local clusters with high genomic diversity . . . 41
  - 2.3.5 Novel mutations appear in the later phase of pandemic . . . 44
- 2.4 Discussion . . . 46
- 2.5 Conclusions . . . 49
- 2.6 Supplementary material . . . 49
  - 2.6.1 Supplementary text: Annotation of mutations further elucidate conserved regions and show a general preference for non-silent changes in the genome . . . 49

3 A comparative study of pan-genome methods for microbial organisms 63
- 3.1 Introduction . . . 64
- 3.2 Methods . . . 65
  - 3.2.1 Data preprocessing . . . 66

3.2.2 Tools . . . 66
3.2.3 Qualitative and quantitative assessment . . . 66
3.2.4 Replication of Yakkala *et al.* . . . 67
3.2.5 Combining methods . . . 67
3.3 Results and Discussion . . . 68
3.3.1 Qualitative and quantitative comparison . . . 68
3.3.2 Replication of Yakkala *et al.* . . . 71
3.3.3 Combining methods . . . 73
3.3.4 Structural variation in transposons carrying the $bla_{\text{OXA-23}}$ carbapenemase gene . . . 73
3.3.5 Exploring different plasmid structures . . . 75
3.4 Conclusions . . . 76
3.5 Supplementary Material . . . 80
3.5.1 Supplementary text . . . 80
3.5.2 Supplementary tables and figures . . . 83

4 SAFPred: Synteny-aware gene function prediction for bacteria using protein embeddings 87
4.1 Introduction . . . 88
4.2 Materials and methods . . . 89
4.2.1 Datasets . . . 89
4.2.2 Building the bacterial synteny database, SAFPredDB . . . 90
4.2.3 Comparison to published function prediction methods . . . 91
4.2.4 SAFPred algorithm . . . 93
4.2.5 SwissProt benchmark evaluation . . . 94
4.2.6 Applying SAFPred to a diverse set of enterococcal genomes, including detailed analysis of pore-forming toxins . . . 94
4.3 Results . . . 96
4.3.1 SAFPredDB: a database to leverage functional information from syntenic relationships across bacteria . . . 96
4.3.2 SAFPred outperforms other tools in function prediction for multiple bacterial species . . . 96
4.3.3 SAFPred surpasses existing tools for annotating distant homologs . . . 97
4.3.4 SAFPred provides more reliable predictions compared to other methods . . . 98
4.3.5 SAFPred identifies five potential novel pore-forming toxins among a diverse set of enterococcal genomes . . . 98
4.4 Discussion . . . 99
4.5 Supplementary Material . . . 102
4.5.1 Supplementary Text: SwissProt Dataset for Benchmarking . . . 102
4.5.2 Analysis of false positive enterococcal toxin predictions . . . 103
4.5.3 Supplementary Tables . . . 104
4.5.4 Supplementary Figures . . . 115

5 SAFPedDB: A comprehensive collection of bacterial operons and syntenic regions 135
5.1 Introduction . . . 136
5.2 Building SAFPredDB . . . 137
5.2.1 Data source . . . 137
5.2.2 Synteny model . . . 138
5.2.3 Vector representation of synteny . . . 138
5.2.4 Further refinement of SAFPredDB . . . 139
5.2.5 Assigning query points to syntenic regions. . . . 140
5.3 SAFPredDB is a comprehensive collection of bacterial operons and syntenic regions . . . 140
5.3.1 Synteny model in SAFPredDB approximates experimentally determined operons in ODB . . . 141
5.3.2 SAFPredDB predicts experimentally determined operons accurately 142
5.3.3 SAFPredDB annotations and region assignments . . . 144
5.4 Conclusion . . . 147

6 Intrinsic and extrinsic antimicrobial resistance in enterococcus genus 153
6.1 Introduction . . . 154
6.2 Materials and methods . . . 156
6.2.1 Enterococci data . . . 156
6.2.2 Data preprocessing . . . 156
6.2.3 Genome annotation . . . 156
6.2.4 Prediction of mobile genetic elements . . . 157
6.2.5 Prediction of antimicrobial resistance genes . . . 157
6.2.6 Assigning species labels and orthogroup clustering . . . 158
6.2.7 Pairwise syntenic distance between gene contexts . . . 158
6.3 Results and Discussion . . . 159
6.3.1 Most extensive view of the Enterococcaceae family . . . 159
6.3.2 Intrinsic resistance patterns in the *Enterococcus* phylogeny. . . . 161
6.3.3 Clade-specific AMR patterns are more prevalent than species-specific patterns . . . 163
6.3.4 Acquired resistance genes follow similar trajectories as they become intrinsic . . . 164
6.3.5 Patterns of intrinsic resistance correlates with subspecies differentiation in *E. faecium* . . . 167
6.3.6 Clade IV species exhibit distinct intrinsic resistance patterns . . . 168
6.4 Conclusion . . . 170
6.5 Supplementary material . . . 171
6.5.1 Supplementary Tables and Figures . . . 171

7 Discussion 179
7.1 With big data comes great insight, but also great responsibility . . . 180
7.2 You say eukaryote I say prokaryote: the perpetual need to catch up with eukaryotic genomics . . . 182
7.3 The isolating effects of isolate genomics . . . 184

Acknowledgments **189**

Curriculum Vitæ **193**

List of Publications **195**

Summary

We are witnessing an era of rapid technological advancements, which led to an explosion in the amount of genomic data collected. The field of comparative genomics, in parallel, is expanding at an unrepentant rate. Comparative genomics explores the similarities and differences in the genomes of various organisms, species or strains, and it is one of our most useful tools today for unraveling the complexities of microbial biology. However, despite growing interest in microbial genomics, there remains a significant gap in our understanding of microbial diversity and function. The microbial dark matter remains elusive, and we have a lot more to uncover.

This dissertation aims to leverage comparative genomics, and develop novel algorithms tailored for microbial genomes to enhance our understanding of microbial biology and address existing knowledge gaps. More specifically, it focuses on the representation of microbial diversity and the functional annotation of poorly characterized taxa. By harnessing large-scale genomic datasets, novel approaches and algorithms are designed to uncover hidden traits in microorganisms.

We begin our journey at the smallest scale with viruses; our study of SARS-CoV-2 genomes in the Netherlands during the COVID-19 pandemic showcases the power of genomic data to understand disease dynamics. The remainder of the dissertation concerns bacteria. We explore pangenome graphs to represent bacterial populations. As I discuss the limitations of current methods, I propose an ensemble approach to exploit graph representations for structural variant calling. This work sets the stage for future developments in pangenome graphs as a powerful framework to model bacterial populations and analyze their genetic makeup.

Following recent developments in algorithms for eukaryotes, I draw inspiration from natural language processing to predict gene functions in bacteria. I present SAFPred, a novel tool in which I integrate bacterial synteny into the predictive model, and demonstrate its use to identify variants of toxin genes in *Enterococcus*. The novelty of my approach lies partly in how I incorporated bacterial synteny into the function prediction algorithm. Thus, I also release our synteny database, SAFPredDB, that can facilitate various comparative genomic analyses in the future.

Our journey comes to an end in our study of the *Enterococcus* genus through the largest collection of genome assemblies. Here, I emphasize the importance of understanding microbial diversity and antibiotic resistance mechanisms once again, and note the power of large scale genomic analyses.

Overall, my main goal with this dissertation is to showcase the potential of comparative genomics in unraveling the mysteries of microbial life and addressing pressing global challenges in health, agriculture, and biotechnology. Through innovative methods and large-scale data analysis, my work, first and foremost, offers valuable insights into microbial biology and evolution, paving the way for future research in the field. And I hope it also encourages further exploration and appreciation of the mighty world of microbes.

Samenvatting

We leven in een tijd van snelle technologische ontwikkelen die hebben geleid tot een explosieve toename van de hoeveelheid genomische data. Parallel aan deze toename, is het veld van de vergelijkende genomica enorm aan het groeien aan het groeien. Dit veld bestudeert de gelijkenissen en verschillen tussen de genomen van verschillende organismes, soorten en stammen, en behoort tot de belangrijkste middelen voor het ontrafelen van de complexiteiten van de microbiële biologie. Ondanks de groeiende interesse in microbiële genomica, bestaat er een significant gat in onze kennis over microbiële diversiteit en functies. Deze "microbiële donkere materie" blijft moeilijk te doorgronden, en er is nog veel te ontdekken. Het doel van deze dissertatie is om vergelijkende genomica te gebruiken en nieuwe algoritmes te ontwikkelen bedoeld voor microbiële genomen, om ons begrip van microbiële biologie te versterken, en gaten in onze kennis hierover te vullen. In het bijzonder focust deze dissertatie zich op de representatie van microbiële diversiteit, en de annotatie van slecht gekarakteriseerde taxa. Door het gebruik van grootschalige genomische datasets zijn nieuwe aanpakken en algoritmes in staat om verborgen eigenschappen van micro-organismes te onthullen.

Onze eerste stap in deze reis begint op de kleinste schaal met virussen; onze studie van SARS-CoV-2 genomen in Nederland tijdens de COVID-19 pandemie laat zien hoe genomische data kan worden gebruikt om de dynamiek van een ziekte beter te begrijpen. De rest van deze dissertatie betreft bacteriën. We onderzoeken pan-genoom grafen om bacteriële populaties te representeren. Ik beschrijf de tekortkomingen van bestaande methodes, en ik stel een ensemble aanpak voor om graaf representaties te gebruiken voor het identificeren van structurele varianten. Dit werk is een stap richting toekomstige ontwikkelingen op het gebied van pan-genoom grafen die kunnen dienen als een doeltreffend kader voor het modelleren van bacteriële populaties, en het analyseren van hun genetische samenstelling.

In navolging van recente ontwikkelingen in algoritmes voor eukaryoten haal ik inspiratie uit natuurlijke taalverwerking om de functies van genen te voorspellen in bacteriën. Ik presenteer SAFPred, een nieuwe tool waarin ik bacteriële syntenie integreer in een voorspellend model, en ik demonstreer hoe dit model kan worden gebruikt om varianten van giftige genen in het *Enterococcus* geslacht te voorspellen. De innovatie van mijn aanpak zit hem deels in hoe ik bacteriële syntenie integreer in het algoritme dat functionaliteit voorspelt. Daarnaast stel ik onze syntenie database, SAFPredDB, beschikbaar zodat toekomstige vergelijkende genomische analyses kunnen worden gefaciliteerd.

Onze reis komt tot een einde met een studie naar het *Enterococcus* geslacht door middel van de grootste verzameling van genoomassemblages. Hier benadruk ik nogmaals hoe belangrijk het is om microbiële diversiteit en mechanismen van antibioticaresistentie te begrijpen, en wijs ik op de kracht van grootschalige genomische analyses.

Al met al is het doel van deze dissertatie om de potentie van vergelijkende genomica voor het ontrafelen van de mysteries van microbieel leven aan te tonen, en de dringende globale uitdagingen op het gebied van gezondheid, agricultuur en biotechnologie aan te

kaarten. Door middel van innovatieve methodes en grootschalige data-analyses biedt mijn werk allereerst belangrijke inzichten in microbiële biologie en evolutie, waarmee ik het fundament leg voor toekomstig onderzoek in dit vakgebied. Tenslotte hoop ik dat het verdere verkenning en waardering van de machtige wereld van microben aanmoedigt.

Preface

"When you light a candle, you also cast a shadow."

— Ursula K. Le Guin

The devil is in the detail, as people often say. Regardless of whether the anonymous author of this phrase had microorganisms in mind when she uttered these words, I can not help but connect it to my own work, which has been a five-year-long quest to uncover hidden traits in microorganisms, be it good or evil. Using the magnifying lens of comparative genomics, we can get a glimpse of the *inner workings* of our most fearful microbial adversaries, from the cause of the "white plague", *Mycobacterium tuberculosis*, to the agent of the "black death", *Yersinia pestis*. But life is not black or white. We drink their wine, we feast on their bread and we devour their cheese; microorganisms are an indispensable part of life on earth and understanding their biology is essential for addressing global challenges in health, agriculture, environmental sustainability, and biotechnology.

Each day, it is possible to sequence genomes more rapidly and cost-effectively, and we have seen a massive increase in the amount of DNA sequences available. To make sense of this trove of data, comparative genomics has advanced in parallel. However, a large fraction of the efforts have been devoted to eukaryotic organisms, while prokaryotes had to take a backseat. The Human Genome Project has been the fuel that ignited the growth in genomics, research on prokaryotic genomes and microorganisms[1] was overshadowed. Since eukaryotes, in particular humans, are of significant interest due to their relevance to human health and disease, human genomics has been prioritized, leading to an increased emphasis on eukaryotic genomics. Studying microbial genomes is crucial to understand microbial life, their biology, population dynamics, ecology as well as human health. Not to mention how rewarding it is; microbes are the most diverse and abundant forms of life on Earth, yet much of this diversity remains unexplored. Comparative genomics allows us to catalog and understand the genetic diversity within microbial communities, uncovering new species, lineages, and functions. By studying the evolutionary history of microbes, and understanding the mechanisms driving genome evolution, we gain insights into the evolutionary relationships between different organisms, revealing patterns of divergence, adaptation, and speciation over evolutionary time scales. In addition, microbial genomics is central to many biotechnological processes; we can analyze microbial genes that encode enzymes, pathways, and metabolic capabilities with potential biotechnological applications. Finally, by studying microbial genomes, we can understand the genetic basis of microbial pathogenicity, antibiotic resistance, and host-microbe interactions, leading to the development of new strategies for disease prevention, diagnosis, and treatment.

[1] I have to note here that yeast, *Saccharomyces cerevisiae* in particular, is an exception, it is one of the most studied microorganisms. But it is an eukaryote.

In my research, I attempted to bridge the gap between eukaryotic and prokaryotic genomics; I applied novel methods and techniques developed for eukaryotes to prokaryotic organisms, and I designed algorithms and bioinformatics pipelines tailored specifically for bacterial genomes. I aimed to unlock the hidden potential of prokaryotic genomes and uncover novel insights into their population dynamics, evolutionary history, genetic diversity, and functional attributes.

1

Introduction

"For now we see through a glass, darkly;
but then face to face: now I know in part;
but then shall I know even as also I am known."

— 1 Corinthians 13:12, *King James Version*

1

Today, the amount of genomic data compounds every second, and the field of comparative genomics seems limitless as we flood the databases with more data every day. Comparative genomics allows us to delve into the intricacies of microbial biology on a scale never before imagined. Although our understanding of the microbial world has advanced immensely, we have come to realize the impressive amount of microbial diversity waiting to be cataloged and studied. Recognizing this untapped potential, this dissertation aims to broaden our understanding of the microbial world through comparative genomics and developing novel approaches and algorithms addressing the shortcomings of existing tools.

In this introductory chapter, I will start with an overview of microbial organisms, and their general characteristics before going into more detail on the particular organisms that I have studied, namely SARS-CoV-2, *Acinetobacter baumanii* and *Enterococcus*. After a brief description of comparative genomics and their use on microbial organisms, I will conclude with a summary of the remainder of this thesis, and explain the research questions and challenges addressed.

1.1 Microbes: wherever we may roam

Small yet mighty, the microbial world is invisible to the naked eye. From the depths of the seas and the damp dark corners of caves to the top of the highest mountains, microorganisms are everywhere. We do not have to venture out to such extremes to be aware of their profound influence, since microorganisms make up as much as half of the cellular content in a human body [1][1]. The hidden realm of the microbial world boasts a remarkable diversity, encompassing bacteria, viruses, fungi, protists, and archaea. Their collective impact on global ecology, human health, and biotechnology cannot be overstated.

Microorganisms span a wide range of forms, functions, and ecological niches. Bacteria can be found in diverse environments, such as soil, water, air, and extreme conditions as well as in symbiotic relationships with other living organisms [2]. Many plants rely on their symbiotic relationship with nitrogen-fixing bacteria to fix nitrogen from the soil, in exchange for carbon compounds. Nitrifying bacteria are a key player in the nitrogen cycle; they convert ammonia to nitrite and nitrate. Similarly, cellulolytic bacteria are essential in bioremediation as they break down plant cellulose in soil and digestive systems. Viruses infect and occupy a wide range of host organisms, including bacteria (bacteriophages), archaea, plants, animals, and humans [3]. There are many different types of viral genomes; the genetic content of a virus can be either DNA or RNA, and the genome can be of different lengths, structures as well as number of strands. Viruses also have an impact on global carbon and oxygen cycles as they play an essential role in ocean ecosystems by controlling algal blooms and nutrient cycling. Similar to bacteria and viruses, fungi and protists also inhabit diverse habitats, including soil, water, air, and symbiotic associations with plants, animals, and other microbes [2]. Fungi and protists have a large influence on soil and aquatic ecosystems; plants often form mutualistic relationships with fungi where fungi enhance nutrient uptake and plant growth. Similarly, predatory protists consume bacteria and other protists to regulate microbial populations in the ecosystem.

Microorganisms play a complex yet indispensable role in human life. Our microbiota is

[1] Contrary to the popular belief, we are not outnumbered by 10 to 1, neither in terms of the number of cells nor in weight: the current estimate is 38:30 in favor of microorganism [1].

a collection of trillions of microorganisms; gut, oral, skin, vaginal and urinary microbiota are some examples of microorganism communities within the human body. [1]. Microbiota are in close interaction with one another and the host, us, to maintain homeostasis, support our physiological functions, and protect us against pathogens. Disruption in these microbial communities often leads to various health conditions and diseases, such as inflammatory bowel diseases, allergies, autoimmune disorders, and mental health disorders [4].

Microorganisms are widely used in a wide range of applications in biotechnology: from industrial fermentation, biopharmaceutical and enzyme production to waste treatment[2] [5]. Bacteria and fungi are key components in industrial fermentation processes. *Saccharomyces cerevisiae* is the yeast species that ferments sugars to produce alcoholic beverages. Similarly, lactic acid bacteria ferment milk to produce many dairy products [7]. Biofuel production relies on a variety of yeast and bacteria species to convert biomass to biofuels by fermenting the sugar from corn, sugarcane or lignocellulosic materials [8].

1.1.1 Viruses

Following the recent COVID-19 outbreak, viruses have been put into the spotlight. Despite the newfound popularity, viruses are more widespread than one might realize since they have the ability to infect all life forms, from the smallest of all, microorganisms to animals and plants [9]. For the longest time, viruses had been incorrectly associated with diseases only, leading many to believe that viruses are an insignificant part of our environments. However, today, it is estimated that there are 10^{31} virus particles on earth, which is more than an order of magnitude more than the number of cells, outnumbering bacteria 10 to 1 [10]. Nestled comfortably between living and nonliving, viruses can replicate only in living cells, and they exist only as viral particles otherwise. These viral particles, or virions, contain genetic material in the form of DNA or RNA; a capsid, a protein coating for their genetic material, and sometimes a lipid outer layer. It is this dual nature of their *life* cycle that confuses many biologists who have struggled to put a label on viruses; failing to tick all boxes in the definition of a living organism, viruses are often referred to as "biological entities" [11].

Their abundance and diversity make it even more difficult to untangle the origins [12]. It is hypothesized that some viruses have evolved from fragments of DNA, such as plasmids, or more complex organisms, titled the *progressive* and *regressive* hypotheses, respectively. Being one of the biggest drivers of genetic diversity through horizontal gene transfer (HGT), viruses could have arisen from such genetic material that gained the ability to move not just within a genome, but also between cells, as proposed by the progressive hypothesis [13]. The regressive hypothesis, on the other hand, suggests viruses to have originated from organisms which lost their ability to replicate independently as the result of a prolonged parasitic relationship [14]. In contrast to the progressive or regressive hypotheses, both of which assume that cells existed before viruses, recently Koonin et al. have put forth the idea of an "ancient virus world" which predates all living organisms. Motivated by the advances in genomics and findings on virus-specific genes, the virus world hypothesis is more compatible with the genomic data currently available [15].

[2] I could go on for over and state millions of other industrial products we have to thank microbes for. But two of the most interesting and overlooked ones are insulin and enzymes in detergents [5, 6].

1

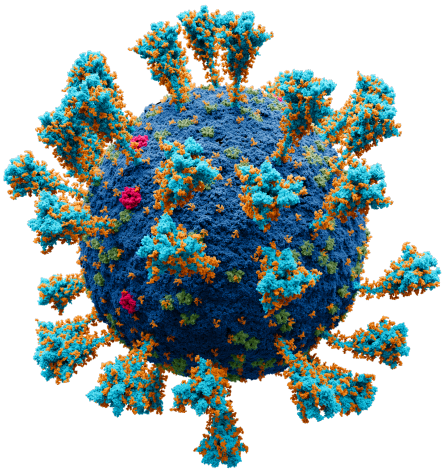


Figure 1.1: The most, if not the only, recognized virion structure to date: SARS-CoV-2 has embellished our websites, newspapers and social media feeds [22].

Regardless of the debates on their origins, viruses are unanimously recognized as the most genetically diverse entities on Earth, as they accompany every cell [10]. Thus, the characteristics of viral genomes, vary considerably but are the result of the differences in their genetic cycles, and their hosts [15]. The advances in genome sequencing have greatly expanded our understanding of viruses at the molecule level, initially with real-time PCR and whole-genome sequencing, and more recently through metagenomic sequencing which allowed us to study virus species in uncultured communities [12]. Viral genetic information can be encoded in either DNA or RNA, and the genome can be single or double-stranded regardless of the nucleic acid type. Virus genomes can differ further in their strandedness, *positive sense* or *negative sense* based on whether it is complementary to the viral messenger RNA (mRNA) or not, in addition to having a linear, circular, or segmented form [16]. Viral genomes can be as small as 2 kb in the *circovirus* genus [17], or as big as 1 Mbp in the *mimivirus*, which is known to be one of the largest virus genera [18].

There are several known mechanisms which can change the genetic material in viruses and lead to increased diversity, such as point mutation, recombination and rearrangement. Point mutations are changes at the level of an individual base in the DNA or RNA; they can be either *silent*, if the mutation does not affect protein encoding, or *non-silent* otherwise. On a larger scale, viral genomes can acquire or lose genes, or segments of genetic material, in addition to similar structural rearrangements within the genome itself [19]. If such genetic change confers evolutionary advantages, it can lead to an antigenic shift with possible implications for human health. For instance, the emergence of drug resistant HIV-1 species through point mutations, or the geminivirus species with a broader host range [20, 21].

SARS-CoV-2

In the last four years there were seemingly neverending periods where we have heard the name *SARS-CoV-2* almost daily, in addition to "2019 novel coronavirus", "human coron-

Table 1.1: Seven coronaviruses known to infect humans.

Genera	Strain	Discovery year
Alphacoronavirus	hCoV-229E	1966
	hCoV-NL63	2004
Betacoronavirus	hCoV-OC43	1967
	hCoV-HKU1	2005
	SARS-CoV	2003
	MERS-CoV	2012
	SARS-CoV-2	2020

avirus" and the "COVID-19 virus", latter of which was adopted to avoid connotations to the SARS outbreaks in 2003 [23]. SARS-CoV-2, a strain of the species severe-acute-respiratory-syndrome-related coronavirus (SARSr-CoV), is classified under the coronaviruses family within the virus taxonomy. Similar to the other members of the coronavirus family, SARS-CoV-2 is a single-stranded, positive-sense RNA virus that infects vertebrates [23]. Within the coronavirus family, SARS-CoV-2 is the seventh one observed to infect humans (Table 1.1) [24]. SARS-CoV-2 was shown to spread through airborne particles, and bind to the angiotensin-converting enzyme 2 (ACE2) to enter the human cells [25]. Interestingly, there were several reports on the transmission of SARS-CoV-2 from humans to animals, and back to humans from animals, for cats, dogs, and minks [26–28]. Today, COVID-19 Dashboard by the Center for Systems Science and Engineering at Johns Hopkins University reports 676,609,955 cases and 6,881,955 deaths from COVID-19 in total [29].

Although there is no clear consensus on the origins of SARS-CoV-2 yet, the most recent studies show bat coronaviruses to be its closest relatives, sharing 96.1% of its sequence at the whole-genome level [30]. In the absence of a confirmed intermediate host, it is not possible to decide whether the virus was transmitted directly from a bat to a human host or not. However, several studies have identified multiple sites of recombination on the SARS-CoV-2 genome, suggesting that it has possibly emerged from a recombination event, consistent with the general characteristics of the coronavirus family [23].

The interest in SARS-CoV-2 genomics was not only limited to uncovering its origins but also studying its evolution, epidemiological surveillance, as well as developing new methods to diagnose and treat the disease caused by the virus [31–35]. Thanks to these efforts, and building on top of the established work on coronaviruses, our understanding of SARS-CoV-2 biology has expanded rapidly. SARS-CoV-2 is placed within the subgenus *Sarbecovirus* (beta-CoV lineage B) [36]. Compared to other RNA virus families, coronaviruses have the largest genomes and SARS-CoV-2 is one of the largest members of the family with its 30,000 bp-long RNA sequence that mostly consists of protein-coding sequences [37]. The SARS-CoV-2 genome encodes 16 non-structural proteins: nsp1 to nsp11 in oRF1a, and nsp12 to nsp16 in ORF1b, in addition to 4 structural and 6 accessory proteins [38]. The polyproteins ORF1a and ORF1b occupy 70% of the entire genome. The structural proteins, S (spike), E (envelope), M (membrane) and N (nucleocapsid) are involved in virion formation, while the accessory proteins (3a, 6, 7a, 7b, 8 and 9b) have unknown function [31]. Among all proteins, though, the spike protein has been the only one to reach *the celebrity status*

since it is the protein that is directly involved in the attachment of the virus to the host cell membrane [38].

As part of its evolutionary trajectory, SARS-CoV-2 has gone through mutations in its genome that led to the emergence of variants. A variant of SARS-CoV-2 is a virus that has significantly diverged from its original genome through random mutations as part of its natural evolution. In general, RNA viruses evolve faster since the RNA molecule, due to its chemical properties, is less stable than DNA; the evolution of coronaviruses can be observed on timescales of as small as months. Having said that, there are external factors affecting the evolution rate of SARS-CoV-2, such as the number of active infections, host immunity and the reality of living in a world that has never been as interconnected as it is today. An advantageous mutation that increases a pathogen's fitness can spread the mutation and become dominant. However, this spread is directly influenced by the host mobility since the variant needs to be transmitted from one person to another to spread. For instance, one of the earliest mutations observed in SARS-CoV-2, the "D614G" mutation, was shown to increase transmissibility, and it rapidly became the established sequence. Nevertheless, given the randomness of evolution, some mutations, even if they do not offer any evolutionary advantage, can be passed on by luck [39].

The earliest variants detected were "A" and "B", the former being the ancestral type that resembled the closest SARS-CoV-2 relatives, the bat and pangolin coronavirus strains. The A type was mostly replaced by the B type which later produced, beta, gamma, delta and omicron as well as the B.1 variant [36]. While there is no set value on the number of mutations required to form a variant, the variants we observed later in the pandemic from the B type had diverged significantly more, carrying at least 11 non-synonymous mutations, compared to their predecessors [39]. In addition, these variants were labeled as "variants of concern" (VOC)[3] since the mutations led to increased transmission, virulence or a reduction in the vaccine efficacy. For instance, the alpha and delta variants were reported to be more transmissible and the omicron variant was more efficient at replicating within the host body [39].

Although there is still no clear consensus on what evolutionary events led to the formation of these VOCs, currently there are three possible explanations: (i) inadequate surveillance of SARS-CoV-2 genomes and their circulation in humans, (ii) circulation of SARS-CoV-2 in animals, and (iii) immunocompromised individuals carrying chronic SARS-CoV-2 infections. In practice, the answer is likely to be a combination of all three of these factors; since the first day of the pandemic, genomic surveillance was conducted successfully only in areas where genome sequencing could be afforded at large scale, leading to a severely imbalanced view of the SARS-CoV-2 evolutionary landscape. Similarly, while there have been reports of transmission from humans to minks and back from minks to humans, with evidence of the virus mutating within their mink hosts, the epidemiological surveillance of animals is not at the same level as it is of humans [28]. The third hypothesis, on immunocompromised individuals being a source of new variants, stems from the studies that show such individuals to shed viral particles for an extended period of time, and it is also supported by the overlap of certain mutations observed in VOCs and the variants in

[3] A term coined during the COVID-19 pandemic to designate SARS-CoV-2 variants of concern for public health; although every health organization maintains its own list of VOCs, according to the WHO, only the Omicron variant remains as a VOC.

immunocompromised hosts [40]. Nevertheless, all three routes outlined here lead to the emergence of VOCs, and they should all be given the utmost importance in our ongoing fight with not just COVID-19, but also future infectious diseases. While it is not possible to predict the next pandemic, we should keep our powder dry; it is not a matter of *if* but *when* it will arrive.

1.1.2 Bacteria

Bacteria are single-celled organisms, with a cell wall to protect the cell from internal pressure and external factors. Unlike other organisms, their cell wall contains peptidoglycan, a polysaccharide molecule that defines the shape of the bacterial cell and provides rigidity. Bacteria are classified into two main categories based on Gram staining; gram-positive and gram-negative. Gram-positive bacteria have a thick cell wall, which leaves a purple stain in the Gram stain. The cell wall in gram-negative bacteria, on the other hand, is thinner and leaves a pink stain. For both types, the genetic material is usually made up of the chromosome DNA and the plasmid DNA. While some bacteria can have multiple chromosomes, the chromosome DNA is in most cases circular [41]. The plasmid DNA is also circular; bacteria can carry different types of plasmids as well as multiple copies of the same plasmid sequence. Plasmid DNA forms a crucial part of the bacteria, and it is often included when we refer to a bacterial genome; we differentiate these two DNAs by chromosome and plasmid, when we can.

A typical bacterial genome is much smaller than a eukaryotic genome, the *E. coli* genome, for instance, is only 4.6 million base pairs (4.6 Mb) long compared to that of *Saccharomyces cerevisiae* (yeast) with 12.1 Mb and the human genome with 3.1 billion base pairs (3.1 Gb). Within the kingdom, symbiotic bacteria have the smallest genomes at 140 kilobases (140 kb), and some soil bacteria are almost 14 Mb long [42]. However, most of the genome sequences (75%) fall below 4 Mb [43]. On average, 90% of the genome consists of genes, or coding DNA; bacteria are considered to be gene-dense compared to other organisms. The *coding density*, the proportion of the coding DNA to non-coding DNA, can vary depending on the environment and ecology, although it remains high throughout the kingdom [43].

In addition to high gene density, a defining feature of bacterial genomes is how these genes are arranged on the chromosome. *Synteny*, first introduced by John H. Renwick in 1971, is a term derived from Greek, translating to "beads on a string". As the translation suggests, it refers to the unique arrangement of genes on the same DNA, and it is found across multiple organisms, conserved to a large extent [44]. While there is no strict definition of the term, syntenic regions in bacteria are specific structures in which the gene content as well as the order of these genes are consistent across multiple strains, species or genera. Although each instance of such genome structures can vary from one another at both the gene and sequence level, the non-random placement of genes in such a manner has been widely studied as it has evolutionary implications [45, 46]. This structural arrangement of genes can be functional; if the genes are localized in a syntenic region due to functional constraints, i.e. near the origin of a replication site, or a regulon which regulates the expression of all genes in the region. These functional regions are referred to as *operons*, which comprise genes that function in the same biological pathway [47]. Since genes in an operon can be directly linked to a specific pathway, it is possible to infer the

1

function of a gene simply based on its location on the genome to a reasonable extent [46].

The stability of smaller genome structures in bacteria, is in contrast to the overall nature of the genome. The bacterial genome is highly dynamic: it goes through gene loss and gain, in addition to recombination. The driving force behind this dynamicity is horizontal gene transfer (HGT). Genetic material of varying sizes is often transferred between bacteria and other organisms in their environment. Genes, small genome structures, including syntenic regions, as well as entire plasmids can be transferred in such manner. Plasmids play a crucial role in HGT; they can mobilize accessory genes, and genes that confer survival advantage in a given environment, such as antimicrobial resistance, virulence in addition to ability to metabolizes molecules [48]. Any genetic material that can be mobilized and transferred between bacteria are generally referred to as mobile genetic elements (MGE). MGEs include, insertion sequences (ISs), integrative and conjugation elements (ICEs), transposons and bacteriophages, all of which can be found in both the chromosomal DNA and the plasmid DNA in a bacteria [49].

ESKAPE pathogens

Despite the advances in medicine and the development of novel therapies and antibiotics, millions of people die from bacterial infections. The biggest factor fueling this growing death rate is the emergence of highly virulent, antibiotic resistant pathogens, and the rapid spread of these pathogenic traits in hospitals. Since the first use of penicillin to treat bacterial infections in 1920, bacteria have shown an impressive ability to acquire resistance against new antibiotics introduced, amassing multiples of such resistance traits in some instances [50, 51]. Multi drug resistant (MDR) bacteria, in particular, were put into the spotlight by WHO in 2015 when they established the list of ESKAPE pathogens, a group of pathogenic bacteria that can readily escape many treatments we have currently available [52]. ESKAPE pathogens include *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter species*, while the acronym is sometimes extended to include *Escherichia coli* as well. ESKAPE pathogens, scoring at the top of WHO's list of 12 bacteria for which new treatments are needed urgently. ESKAPE pathogens are the major cause of nosocomial, hospital acquired infections worldwide, leading to increased mortality, infection severity and healthcare costs [53]. It is estimated that ESKAPE pathogens make up as much as 15.5% of hospital acquired infections [54].

The mechanisms of MDR varies widely from species to species, however, the ESKAPE pathogens commonly exhibit three types: reduction of the antibiotic molecule concentration, modification of the antibiotic target site, or the inactivation of the molecule [54]. In addition, biofilm formation plays an important role in the persistence of infections. All these mechanisms associated with MDR are naturally found in bacteria to aid in their survival in an environment where other microorganisms are also present and there is a need to compete. These survival tactics are ancient and they predate the human use of antibiotics. However, it is the recent use and abuse of antibiotics in healthcare settings, as well as the agricultural sector that led to the selective pressure for MDR to spread [55]. Another important yet often overlooked factor is the increased healthcare use in general. Following the improvements in medical technology, hospitals and medical care have become more accessible to a larger population. While it has been immensely successful in improving human health conditions as well as increasing average life expectancy, on the

other side of the coin, we are faced with increased hospital associated infections due to unnecessary hospital visits or prolonged hospital stays beyond needed [52].

It has been now recognized that an all-hand approach is required to address the threat posed by ESKAPE pathogens. This includes raising awareness, and increasing support for research to study these organisms and aid in developing new drugs. Raising awareness to mobilize stakeholders, such as the policymakers, healthcare providers, as well as the general public. We can increase the awareness of globally threatening consequences of infections, and advocate for responsible antibiotic use through educational campaigns and public health initiatives. In addition, healthcare providers should also be well equipped with the knowledge and skills to mitigate ESKAPE infections particularly. Moreover, research to advance our understanding of microbial pathogens and facilitate the discovery and development of new antimicrobial agents should be supported. By leveraging genomic data to identify drug targets, discover new antibiotics, elucidate resistance mechanisms, and guide precision medicine approaches, researchers can accelerate the development of effective antimicrobial therapies to address the global threat of antibiotic resistance.

Acinetobacter baumannii: a superstar pathogen

Acinetobacter baumannii[4] is a Gram-negative bacterium that belongs to the genus Acinetobacter [57]. Acinetobacter comprises a diverse group of Gram-negative bacteria characterized by their coccobacillary morphology and aerobic metabolism. Unlike other Acinetobacter species which are found in soil exclusively, *A baumannii* is often isolated from hospitals, although its natural habitat remains unknown [58]. *A. baumannii* can survive in diverse environmental conditions, including desiccation, nutrient limitation, and exposure to disinfectants. This resilience contributes to its widespread presence in healthcare settings and its role as an opportunistic pathogen. Being the "A" in the infamous group of "ESKAPE" pathogens, it is one of the most successful pathogens, exhibiting extensive MDR [52].

Among the shortest bacteria, *Acinetobacter baumannii*, has punched above its weight and emerged as a significant opportunistic pathogen in healthcare settings worldwide. Its clinical impact spans a wide range of infections, including pneumonia, bloodstream infections, urinary tract infections, and wound infections [57]. Particularly concerning is its propensity to cause infections in immunocompromised patients and those with prolonged hospitalizations, making it a formidable challenge for healthcare providers [59]. *A. baumannii* thrives in environments where patients are exposed to invasive procedures, such as intensive care units; prolonged hospitalization, mechanical ventilation, and immunocompromised status further elevate the risk of *A. baumannii* infections in ICU patients [58].

The *A. baumannii* genome ranges from 3 to 4 Mbp in length, and it often includes plasmid DNA in addition to its chromosome, in particular the MDR strains [57]. Its genome is highly plastic; it can rapidly adapt to changing environmental conditions as HGT plays a significant role in shaping the genomic diversity of *A. baumannii*. *A. baumannii* can acquire new traits through HGT of antibiotic resistance genes, virulence factors, and other genetic material that increase fitness [60].

[4] It was named after the bacteriologist Paul Baumann [56]. It must be an honor to have your *own species*, but at the same time I wonder if it is at all challenging to be associated with a notorious pathogen.

1

Antibiotic resistance in *A. baumannii* is usually caused by genes which encode enzymes that modify or degrade antibiotic molecules, efflux pumps that actively remove antibiotics from the bacterial cell, and alterations in antibiotic targets that reduce their binding affinity. *A. baumannii* is intrinsically resistant to many antibiotics including beta lactams, aminoglycosides, fluoroquinolones, and trimethoprim-sulfamethoxazole. This intrinsic resistance is due to the impermeable outer membrane, efflux pumps, and enzymatic inactivation mechanisms [61]. In addition, *A. baumannii* can acquire resistance from its environment, and the most clinically significant acquisition is usually plasmid-mediated. Resistance genes are often carried on plasmids, and plasmids can harbor multiple resistance genes; thus MDR can rapidly spread through such plasmids among bacterial populations. Several studies had identified specific genetic loci and genomic regions associated with MDR on both the chromosome and plasmid in *A. baumannii*, such as resistance islands and genomic islands [60]. These regions often contain clusters of resistance genes and other MGE, such as integrons, transposons and IS, which aid in mobilizing these genomic regions and spreading antibiotic resistance within bacterial population [60].

Carbapenem resistance in *A. baumannii* is the most concerning trait since carbapenems are often considered the last line of defense against MDR Gram-negative bacteria [62]. Resistance to carbapenems in *A. baumannii* is primarily mediated by the production of carbapenemases, enzymes that hydrolyze carbapenem antibiotics and render them ineffective [63]. Carbapenemases of clinical significance in *A. baumannii* include Acinetobacter-derived cephalosporinases (ADCs) and OXA-51, often found within the context of IS, and it is usually transferred by HGT [64]. In addition to carbapenemases, *A. baumannii* can exhibit carbapenem resistance through decreased permeability of the outer membrane; the effectiveness of carbapenems are greatly reduced when the entry of antibiotic molecules is limited. Similarly, efflux pump systems, such as AdeABC, AdeIJK, AbaR and AdeFGH, actively remove carbapenems from the bacterial cell, further decreasing the concentration of antibiotic molecule [61]. Similarly, mutations in the penicillin-binding proteins, the targets of carbapenem antibiotics, can also reduce their binding affinity and efficacy [65].

A. baumannii has recently made the news during the COVID-19 pandemic when patients with severe COVID-19 that required ICU and prolonged hospitalization, forming the breeding ground for carbapenem resistant *A. baumannii* infections [66]. This was compounded by the increased antibiotic usage in hospital settings; antibiotics, including carbapenems were commonly prescribed to treat bacterial co-infections or secondary infections in COVID-19 patients. Since hospitalized patients with severe COVID-19 were often at increased risk of acquiring infections due to prolonged hospitalization, invasive procedures, and compromised immune function, it put further strain on the healthcare system which was already under the pressure of a global pandemic. This event demonstrated the imminence of MDR bacteria outbreaks.

Enterococcus: a creature of survival

The enterococcus genus is a gram-positive, low-GC lactic acid bacteria within the taxonomic class, Bacilli. Bacilli is one of the 11 classes within the phylum Bacillota, previously named Firmicutes[5], meaning "tough skin" in Latin, which remains the best description of

[5]Firmicutes phylum was renamed to Bacillota in 2021; three years have passed already but as it is the case with many things in science, the renaming holds its controversial status. In this thesis, I will follow the nomenclature

enteroccus. The Enterococcaceae family within the lactic acid bacteria includes the closest relatives of *Enterococcus*: *Melissococcus*, *Pilibacter*, *Tetragenococcus*, and *Vagococcus*. The tough skinned enterococci stand out with their resistance and ubiquity; they can withstand various harsh environmental conditions, and they are one of the most abundant species in the gut microbiota across a wide range of terrestrial animals. They have established their presence in the gut microbiomes of practically all animals on land: from insects, snails, and reptiles to mammals, birds, and fish. They can also be found in water, soil, fermented food, and plants. This innate hardiness contributes to their persistence in hospital settings, and they have been frequently identified as one of the leading causes of antibiotic resistant bacteria in hospitals as well as hospital acquired infections in humans. *Enterococcus spp.*, is the first "E" in the infamous "ESKAPE" pathogens, which WHO has established as the priority list of MDR pathogens to investigate [52].

Pathogenic enterococci have emerged as one of the most clinically relevant bacteria in the last 25 years, causing bacteremia, urinary tract infections, meningitis, and endocarditis and several other infectious diseases. *E. faecium* and *E. faecalis* are two species that have been associated the most with infections globally and they are the most commonly found species in the genus [67]. *E. cecorum*, *E. gallinarum*, and *E. durans* have been isolated frequently from animal sources, while *E. casseliflavus* and *E. mundtii* are found in plants and soil [48].

The genetic variation in enterococci is highly affected by its environment; as an indispensable member of the animal gut microbiome, their genotype is shaped by the host-microbe interactions. Some of the most diverse enterococci are found in arthropods even though we have explored only the tip of the iceberg of their diversity [48]. Similarly, enteroccal species in wild environments such as soil and water sources differ from those isolated from fermented food or dairy [68, 69]. The enterococcal genome size ranges from 2Mb to 5Mb, and it contains 3000 genes on average. Recently, Schwartzman et al. reported 417 single copy core (SCC) genes in the enterococcus pangenome, and found 1336 genes were shared by more than 80% of the species. Enterococci consists of four major clades, which were initially introduced by Lebreton et al., then verified and expanded with novel species and increased support for the branching [48].

As enterococci are quite versatile and adaptive in nature, they can readily acquire new traits through gene gain, and their genome is subject to HGT events, increasing their fitness to survive in their niche, perhaps even more so than most other bacteria. As one of the leading causes of the spread of antibiotic resistance, this trait has allowed enterococci to acquire resistance mechanisms against commonly used antibiotics and become one of the most threatening MDR bacteria. Both vancomycin and tetracycline resistance were acquired in the 1980s, until which point these drugs were established as a treatment for enterococcal infections [71]. Vancomycin resistance is often carried by transposons, and it is found in an operon; there are multiple such operons detected in enterococci, *vanA* operon, which includes the VanA resistance gene, is the most common one in hospital associated infections [72].

In addition to acquired resistance, enterococcus is known to be intrinsically resistant to aminoglycosides, beta lactams, fluoroquinolones and lincosamides [73]. Intrinsic resistance in enterococci is particularly alarming given their propensity to exchange genetic material

in NCBI and use Bacillota, unless I am referring to data retrieved before the renaming.

with their environment, and it can lead to the spread of antibiotic resistance. Intrinsic resistance, as it predates human use of antibiotics, plays an often overlooked role in antibiotic resistance; enterococcal species from undersampled, wild environments, such as insect guts or soil, harbor functional repertoire that is yet to be mined. For instance, most enterococci are intrinsically susceptible to vancomycin and they rapidly acquired resistance through the prolonged use of these antibiotics for treatment. However, *E. gallinarum*, an undersampled species, is intrinsically resistant to vancomycin and it was put forth as a potential pathogen since the resistance genotype could be transmitted although it was observed in only limited quantities in clinical settings [74]. Later on, both *E. faecalis* and *E. faecium* species were found to carry the *E. gallinarum* specific vancomycin operons [75]. Thus, sampling new species to discover intrinsic resistance genes, before they can make their way into hospitals or even cause infections has the potential to contribute greatly to combat the spread of antibiotic resistance.

Enterococci possess a large pool of virulence factors that can enhance their ability to colonize a host, bypass the host's immune system and establish infections [71]. Surface adhesins, extracellular enzymes, cytolysin and quorum sensing systems are among some enterococcal virulence traits in addition to biofilm formation [76]. For instance, cytolysins are toxin proteins that were shown to increase death in infected patients five folds and the extracellular surface protein gene mediates cell adhesion and evasion of host immunity [71, 77]. Their highly adaptive, hardy nature, and their propensity to exchange virulence traits and antibiotic resistance with their environment, in addition to their ubiquity in land animals as well as the environment, put enterococci forth as one of the most important bacterial species to investigate.

1.2 Computational comparative genomics to magnify the microbial world

The first complete nucleotide sequence we obtained was an RNA virus; the complete genome of a bacteriophage was sequenced at Ghent University in 1976. Growing in number since then, we have more than 3 million complete nucleotide sequences on NCBI Virus database today. Similarly, bacterial genome sequencing is also well established, going back almost 30 years ago when the first complete genome sequence of *Haemophilus influenzae* was obtained[78]. As the world of genomics moved from Sanger shotgun sequencing to high-throughout next generation sequencing (NGS), and then single molecule long read sequencing, the field of microbiology has followed along. Today, the vast majority of what we know about microbial genomes comes from this ongoing "roller-coaster ride" [79]. The new millennium brought in more genomes, more genes and more problems. NGS became well established in 2005, and we were able to produce millions of reads in a short time for a fraction of the price of older Sanger technologies. As we processed these reads and assembled them into genomes with high coverage, we were faced with the stark reality that not every *E. coli* genome is identical to one another, and the databases grew in size every day [80].

This pivotal moment brought to attention the need for comparative genomics at a large scale, as well as a growing interest in microbial genome sequencing. To store, share and analyze such massive amounts of sequencing data, we need more advanced computational

Table 1.2: Number of taxonomy nodes on the NCBI Taxonomy database (statistics retrieved on 13 April, 2024 from the NCBI Taxonomy database): our databases are still highly skewed towards eukaryotes, almost 95% of all nodes on the NCBI Taxonomy database are from eukaryotes, the prokaryotic taxonomy remains unexplored [84].

Ranks:	Higher taxa	Genus	Species	Total
Archaea	705	299	994	1,998
Bacteria	6,823	5,275	26,345	39,408
Eukaryota	69,631	100,683	540,982	749,738
Viruses	2,266	2,795	5,976	11,831
All taxa	79,456	109,053	574,283	802,993

methods. The field of comparative genomics, as it emerged with the broader subject of genomics, involves studying genomes in comparison to others, typically across species or closely related individuals within a species. The primary goal of comparative genomics is to identify similarities and differences in genomic content, organization, and function to gain insights into evolutionary relationships, genetic variation, and biological diversity [81]. Comparative genomics allows researchers to infer evolutionary histories, identify conserved genomic regions, and discover genes or genomic elements associated with specific traits or phenotypes [82]. The field comprises an arsenal of tools and various analytical approaches, including sequence alignment, phylogenetic analysis, synteny analysis, and functional annotation, to interpret genomic data and understand the underlying biological mechanisms shaping genome evolution [83]. The title of my thesis emphasizes the power of comparative genomics and the added value of mining the ever expanding trove of genomic data as a whole. With an emphasis on improving computational methods *specifically* to use on microbial genomes, comparative genomics has been a central component of microbial studies.

Comparative genomics for microbial genomes both helped fuel the growth of genome data, as well as benefited from the established databases. Today, we have several databases available for microbial genomes, although they are still lagging behind those of eukaryotes in terms of both size and number. The main databases are the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ), although NCBI has taken over the latter two in content. In addition to these, there are "meta databases" that retrieve microbe sequences and compile them in a single database to aid bacteria-specific analyses. The Genome Taxonomy Database (GTDB) and Genomes OnLine Database (GOLD) are such meta databases tailored to microbes. Finally, the Global Initiative on Sharing All Influenza Data (GISAID) has received the most attention among all microbial databases during the COVID-19 pandemic.

GTDB was established in 2018 to catalog the wealth of sequencing data on bacterial and archaeal organisms [85]. GTDB provides a standardized, phylogeny-based approach to taxonomic classification. The GTDB taxonomy uses genomic and metagenomic data from Genbank and NCBI RefSeq databases, which then go through a quality check. Their latest release (08-RS214 from 28th April 2023) contains 402709 genomes in total which comprises 394932 bacterial genomes with 80789 species and 7777 archaeal genomes with 4416 species. Similarly, GOLD, an online resource that keeps track of sequencing projects,

1

compiles information about genomic data available on public databases [86]. GOLD reports 433761 bacterial genomes, 27222 of which are type strains. Note the difference between the number of species in these two databases: GTDB holds 80789 species compared to 27222 in GOLD. The GTDB species are essentially clusters of genomes represented by a single genome, which is picked from genomes of the type strains if available. Only 14826 (17.4%) of these genomes are isolate type strains, and the remainder are cluster representatives, mostly obtained from metagenome-assembled genomes (more than 75%). This discrepancy highlights the benefits of harnessing data from metagenome-assembled genomes (MAGs) where uncultured bacteria can provide invaluable information about the *microbial dark world*. Finally, GISAID, to this day, is the biggest resource of SARS-CoV-2 genome sequences with over 15 million genome sequences in its inventory [87]. GISAID is a science initiative established in 2008; upon renaming itself from the Global Initiative on Sharing Avian Influenza Data, it has set the ground for rapid sequencing and sharing of data related to pathogens of global concern [88]. The catalog of GISAID includes a wide range of data, from genome sequences, metadata specific to a virus species to clinically relevant data and epidemiological surveys. With its emphasis on open science, GISAID is one of the strongest driving forces behind the relentless work on influenza viruses. During the COVID-19 pandemic, we observed it accelerate the research on COVID-19 globally; having access to epidemiological and genomic data from all around the world has led to a better understanding of the SARS-CoV-2 evolution and the emerging variants [88].

The wide spread adoption of comparative genomics has transformed our understanding of microbial genomics, with many applications to study their evolution, diversity and genotype; as we became more familiar with the characteristics of microbial genomes, we learned to hone our tools and methods in our arsenal. As I described before, the dynamic nature of bacterial genome rewards them with an immense genomic diversity. For that reason, bacterial evolution can be both more intriguing and tricky to investigate. We use phylogenetic trees to study the evolution of an organism; the phylogeny is represented with a tree data structure, where the branches are *speciation* events. New species form when the genome has diverged sufficiently from a parent node in the tree. Since the major driving force behind this dynamicity is HGT, it is often futile, if not incorrect, to build a bacterial family tree [89]. Bacterial evolution is best described in the form of a network instead of a tree, although they found only limited applications in the field [90]. There are fewer methods and tools developed to accommodate a network-based phylogeny, and it can be difficult to interpret a network structure for evolutionary insight since a tree is a more *intuitive* way to understand the evolution of a species for us humans. Keeping the caveats of a tree structure, as well as the dynamicity of a bacterial genome, phylogenetic trees are useful to study bacteria and most commonly used.

For many bacteria, evolution at genome scale can be incompatible with the evolution at the scale of a gene, leading to ambiguous species and further complicating the taxonomy [89]. Thus, it is also useful to look at the evolution of a specific gene, as opposed to the whole species. Similar to species taxonomy, we can group genes into clusters to relate one gene to another in terms of its evolutionary history and/or its function. A *gene family*, is a cluster of *homologous genes* often formed based on sequence similarity. Homolog genes can be either *orthologous* or *paralogous*, the former necessitating a shared common ancestor whereas the latter may be the result of a gene duplication. Identifying orthologous

genes is one of the core aspects of studying evolutionary relationships, gene function, and genome evolution across different microbial species. We can infer gene function, conserved pathways and biological processes by studying gene families, as well as annotate novel genes [91, 92].

Another important aspect of comparative genomics is the study of gene content and genome organization. Due to the gene-dense nature of the genome, the presence of stable genome structures, as well as its plasticity, it is useful to view bacterial genomes in a modular way instead of nucleotide sequences. This view lends itself naturally to studying bacterial evolution at different levels, such as the gene or a genomic region, and thus analyzing the genome organization based on these genomic units. Identifying genes or genome structures that are present or absent is one of the most trivial ways to compare genome organization across species. Differences in genome organization between species or strains can indicate nice adaptation, evolution, or HGT [48]. Since HGT is one of the main sources of virulence and antibiotic resistance in pathogens, understanding HGT is important to detect the emergence of novel pathogens or traits [75].

Since some species can vary as much as 50%, many researchers adopted the view of a species as a collection of genomes as opposed to a single linear sequence. Coined by Tettelin et al. first in 2005, pangenome is a collective way to represent this collection and describe the shared genetic material as well as the variance. Although it was initially proposed to study species, a pangenome can be constructed at any taxonomic level, such as strains, genus and phyla. Pangenome can be broken down into a *core genome* and an *accessory genome* at a coarse level. The core genome is conserved in at least 99% of the genomes, and the accessory genome is broken further into soft-core (95-99%), shell (15-95%) and cloud (less than 15%)[6]. Within the view of a modular bacterial genome, where the smallest genetic unit is a gene, bacterial pangenomes are usually constructed from genes directly [95]. Bacterial pangenomics has found several applications, and it is a growing field as we sequence more genomes, discover new species and new genotypes to study [96–98].

More recently, comparative genomics has been brought to public attention during COVID-19 where it was proven instrumental in advancing our understanding of SARS-CoV-2 biology, transmission dynamics, and host interactions, with important implications for public health, disease surveillance, and the development of countermeasures against COVID-19 [88]. Many researchers combined novel comparative genomics tools with traditional epidemiological methods for genomic epidemiology [33]. This approach allowed us to track the evolution of SARS-CoV-2 over time and across different geographic regions; by identifying mutations, deletions, and insertions in the viral genome, we understood their potential impact on viral transmissibility, virulence, and immune evasion [34, 99]. In addition, beginning with the rapid sequencing of the first SARS-CoV-2 genome, comparative genomics has informed the design and development of vaccines and antiviral therapeutics for COVID-19 [35]. Understanding the genetic diversity of SARS-CoV-2 and its variants helped researchers anticipate potential escape mutations and design vaccines that induce broad and durable immune responses [24]. Finally, the compounding effects of comparative genomics were also demonstrated during the COVID-19 pandemic where the insights

[6]There is no established consensus on the definition of these partitions, however, in the context of this thesis I will follow the soft-core threshold described in [94].

gained previously from SARS outbreaks in 2003 were informative [23]. Thus, the recent work on SARS-CoV-2 will prove beneficial in the future as newer pathogens emerge, and are faced with another pandemic. In a world which grows more interconnected as time passes, it is only *natural* to assume that the next pandemic is right around the corner.

1.3 Research objectives and contributions

Despite all the impressive methodological advances in microbial genomics that gave us valuable insight into microbial biology, ecology and evolution, several gaps and challenges remain. Since microbial genomics is of clinical concern, most of the sequencing efforts and subsequent analyses are case-specific; they are limited in their scope, and often confined to a single setting, such as a hospital outbreak, or they focus on an individual organism with a particular pathogenic trait. We have significantly more to gain from adopting a wider view in our work and exploiting the growing size of microbial genomic data. For instance, the extensively diverse functional repertoire of microbial organisms can only be unraveled through mining large genomics datasets. Especially for less studied organisms and poorly characterized microbial taxa such as uncultivated bacteria, or the microorganisms from distant ecologies, functional annotation remains a challenge [100]. The aptly named microbial dark matter, the uncultured microbial populations, make up more than 80% of the taxa and it is out there to be discovered [101]. Similarly, we have a limited understanding of the origins and evolution of pathogenic traits in microbial organisms. While there are several systems and methods established for genomic surveillance of pathogens, we have yet to discover the ecological niche and environmental conditions where many antimicrobial resistance genes originated and further spread out to infect humans.

Fascinated by the remarkable diversity of microbial organisms, and encouraged by the developments in the field of computational comparative genomics, this thesis is a journey embellished with algorithms in the realm of microbial genomics. My goal is to bridge these knowledge gaps in microbial genomics, to illuminate the microbial world, and to unravel the mysteries of microbial life, through the magnifying glass of comparative genomics.

In Chapter 2, we begin at the smallest scale, with viruses. During the COVID-19 pandemic, we witnessed the exponential growth of SARS-CoV-2 genomes and we were motivated to mine the trove of data accumulating to understand the introduction, population dynamics and the ongoing evolution of SARS-CoV-2 genomes in the Netherlands. Several studies investigated hospital outbreaks, case studies, or COVID-19 in animal farms in the Netherlands, but there was no work analyzing SARS-CoV-2 genomes broadly in the whole country. Thus, we amassed the largest collection of SARS-CoV-2 genomes in the Netherlands, which allowed us to gain insight into the local dynamics of COVID-19, and track its spread through monitoring variants. Our findings are valuable to initiate region-specific measures to control outbreaks, design viral treatments, or develop vaccines. Our work demonstrated the power of harnessing large scale genomics data using suitable methods.

In Chapter 3, we are welcomed into the magnificent kingdom of bacteria. Amazed by the population diversity, our work was motivated by the pitfalls of inadequate representations that fail to account for the bacterial population diversity. In particular, the use of a single linear reference sequence to represent the bacterial species, *A. baumannii*. This issue had been highlighted in several studies in the literature, leading to the emergence of pangenome

graphs to study bacterial populations, and the consequent development of methods to build such pangenome structures [96, 98, 102]. Our goal in this study was to explore these methods in depth and analyze their features since there was a lack of knowledge of their practical use and applicability. In this work, we provided an overview of the bacterial pangenomics landscape, and practical guidelines on how to use pangenomes to study any bacteria. We also performed a small comparative study on a collection of *A. baumanni* strains, where we developed an ensemble method to build pangenome graphs. Our ensemble approach was proven effective for structural variant calling as we detected a novel plasmid structure carrying MDR. Our findings show the clinical significance of comparative genomics in studying bacteria.

Chapter 4 continues to develop novel methods tailored to bacteria. This time we were influenced by the seemingly distant field of natural language processing (NLP). Recognizing the glaring similarities between a genome sequence and sentences in human language, many researchers adopted recent methods from NLP to study genomes and develop protein language models. These models were incredibly powerful in studying gene sequences since they allowed us to extract secondary and tertiary structures from the linear sequence alone, and go beyond the sequence similarity [103]. However, there have been limited applications on microbial organisms [104]. Given the extensive functional diversity of bacteria and the fact that we know the function of only a small fraction of the bacterial genes, we hypothesized that we had a lot to gain from using protein language models to understand the diverse functional repertoire of bacteria. Thus, we developed SAFPred, a novel synteny-aware gene function prediction tool based on protein embeddings from state-of-the-art protein language models. Our work is novel in our emphasis on being bacteria-specific in both the development and evaluation of our method; our method exploits bacterial synteny in combination with protein embeddings representation, and we assessed the predictive performance on benchmarks tailored to bacteria. SAFPred outperformed conventional sequence-based bacterial genome annotation pipelines as well as a state-of-the-art deep learning method. We also demonstrated SAFPred's performance by predicting potential novel toxin genes in *Enterococcus* species, which could have clinical implications. Chapter 4 is the climax of this thesis where we clearly showed the power of designing algorithms tailored to bacteria.

Chapter 5 was initially an appendix to Chapter 4, which later spawned its own chapter due to its added value independent of Chapter 4. SAFPred, presented in Chapter 4, exploits bacterial synteny using SAFPredDB, a bacterial synteny database we have built. There are only a handful of bacterial operon databases, most of which are limited in their scope or out of date. To address the need to catalog bacterial synteny across a wide range of species, we presented SAFPredDB, a comprehensive collection of bacterial operons and syntenic regions found across the entire bacterial kingdom. In this chapter we explained the purely computational, bottom-up approach we designed to build SAFPredDB which is based on a synteny model we proposed, and we demonstrated the validity of our approach by comparing it to existing databases. SAFPredDB deserved its own chapter as it is not only a repository for SAFPred, but it is a valuable resource of genomic information, facilitating comparative genomic analyses, evolutionary studies, and functional genomics research in bacteria.

Chapter 6 is the culmination of the collective effort within microbial comparative

1

genomics research. Within the genus *Enterococcus*, the two most common pathogens, *E. faecium* and *E. faecalis*, have been studied extensively, however, there is relatively less effort in analyzing the genus in its entirety, including those species isolated from rare ecologies and geographies. Our goal in this chapter is to understand the incredibly complex and rich world of the *tough-skinned Enterococcus* through comparative genomics, taking a holistic view of the genus. We curated the largest, most diverse collection of *Enterococcus* genomes publicly available. Our *Enterococcus* collection allowed us to clarify species boundaries within the genus, and expand the known species labels to accommodate the recently discovered species, in addition to increasing the support for the known clade definitions. Our approach forms a solid foundation for future investigations into *Enterococcus* diversity and evolution. We further mined the Enterococcus collection to explore AMR traits unconstrained by any specific sampling location, setting, drug class, or species boundaries. We described our systematic approach to identify AMR, distinguish between intrinsic and acquired resistance, and illuminate the evolutionary trajectories of resistance traits within *Enterococcus* populations. Our work contributed to our understanding of the mechanisms and evolution of AMR in bacteria in general.

References

[1] Ron Sender, Shai Fuchs, and Ron Milo. Revised estimates for the number of human and bacteria cells in the body. *PLoS biology*, 14(8):e1002533, 2016. doi: 10.1371/journal.pbio.1002533.

[2] Eric E Allen and Jillian F Banfield. Community genomics in microbial ecology and evolution. *Nature reviews microbiology*, 3(6):489–498, 2005. doi: 10.1038/ismej.2009.88.

[3] Edward F DeLong. The microbial ocean from genomes to biomes. *Nature*, 459(7244):200–206, 2009. doi: 10.1038/nature08059.

[4] Andrina Rutsch, Johan B Kantsjö, and Francesca Ronchi. The gut-brain axis: how microbiota and host inflammasome influence brain physiology and pathology. *Frontiers in immunology*, 11:604179, 2020. doi: 10.3389/fimmu.2020.604179.

[5] Wolfgang Landgraf and Juergen Sandow. Recombinant human insulins–clinical efficacy and safety in diabetes therapy. *European endocrinology*, 12(1):12, 2016. doi: 10.17925/ee.2016.12.01.12.

[6] Francois Niyongabo Niyonzima and Sunil More. Detergent-compatible proteases: microbial production, properties, and stain removal analysis. *Preparative Biochemistry and Biotechnology*, 45(3):233–258, 2015. doi: 10.1080/10826068.2014.907183.

[7] G Licitra, M Caccamo, and S Lortal. Chapter 9-artisanal products made with raw milk. *Raw Milk*, pages 175–221, 2019. doi: 10.1016/B978-0-12-810530-6.00009-2.

[8] Anjani Devi Chintagunta, Gaetano Zuccaro, Mahesh Kumar, SP Jeevan Kumar, Vijay Kumar Garlapati, Pablo D Postemsky, NS Sampath Kumar, Anuj K Chandel, and

Jesus Simal-Gandara. Biodiesel production from lignocellulosic biomass using oleaginous microbes: Prospects for integrated biofuel production. *Frontiers in Microbiology*, 12:658284, 2021. doi: 10.3389/fmicb.2021.658284.

[9] Robert A Edwards and Forest Rohwer. Viral metagenomics. *Nature Reviews Microbiology*, 3(6):504–510, 2005. doi: 10.1038/nrmicro1163.

[10] Mya Breitbart and Forest Rohwer. Here a virus, there a virus, everywhere the same virus? *Trends in microbiology*, 13(6):278–284, 2005. doi: 10.1016/j.tim.2005.04.003.

[11] Eugene V Koonin and Petro Starokadomskyy. Are viruses alive? the replicator paradigm sheds decisive light on an old but misguided question. *Studies in history and philosophy of science part C: Studies in history and philosophy of biological and biomedical sciences*, 59:125–134, 2016. doi: 10.1016/j.shpsc.2016.02.016.

[12] Moïra B Dion, Frank Oechslin, and Sylvain Moineau. Phage diversity, genomics and phylogeny. *Nature Reviews Microbiology*, 18(3):125–138, 2020. doi: 10.1038/s41579-019-0311-5.

[13] Patrick Forterre. The origin of viruses and their possible roles in major evolutionary transitions. *Virus research*, 117(1):5–16, 2006. doi: 10.1016/j.virusres.2006.01.010.

[14] Didier Raoult and Patrick Forterre. Redefining viruses: lessons from mimivirus. *Nature Reviews Microbiology*, 6(4):315–319, 2008. doi: 10.1038/nrmicro1858.

[15] Eugene V Koonin, Tatiana G Senkevich, and Valerian V Dolja. The ancient virus world and evolution of cells. *Biology direct*, 1:1–27, 2006. doi: 10.1186/1745-6150-1-29.

[16] Alan J Cann. *Principles of molecular virology (standard edition)*. Academic press, 2001. doi: 10.1016/C2014-0-01081-7.

[17] Xing-xiao Zhang, Shao-ning Liu, Zhi-jing Xie, Yi-bo Kong, and Shi-jin Jiang. Complete genome sequence analysis of duck circovirus strains from cherry valley duck. *Virologica Sinica*, 27:154–164, 2012. doi: 10.1007/s12250-012-3214-4.

[18] Elodie Ghedin and Jean-Michel Claverie. Mimivirus relatives in the sargasso sea. *Virology Journal*, 2:1–6, 2005. doi: 10.1186/1743-422X-2-62.

[19] John Holland and Esteban Domingo. Origin and evolution of viruses. *Virus genes*, 16:13–21, 1998. doi: 10.1023/A:1007989407305.

[20] Francisco J Morales. History and current distribution of begomoviruses in latin america. *Advances in virus research*, 67:127–162, 2006. doi: 10.1016/S0065-3527(06)67004-8.

[21] Karin J Metzner, Stefano G Giulieri, Stefanie A Knoepfel, Pia Rauch, Philippe Burgisser, Sabine Yerly, Huldrych F Gunthard, and Matthias Cavassini. Minority quasispecies of drug-resistant hiv-1 that lead to early therapy failure in treatment-naive and-adherent patients. *Clinical infectious diseases*, 48(2):239–247, 2009. doi: 10.1086/595703.

[22] Alexey Solodovnikov and Valeria Arkhipova. Scientifically accurate atomic model of the external structure of the severe acute respiratory syndrome coronavirus 2 (sars-cov-2), a strain (genetic variant) of the coronavirus that caused coronavirus disease (covid-19), first identified in wuhan, china, during december 2019, 2021. URL https://commons.wikimedia.org/wiki/File:Coronavirus._SARS-CoV-2.png. File: Coronavirus. SARS-CoV-2.png.

[23] The species severe acute respiratory syndrome-related coronavirus: classifying 2019-ncov and naming it sars-cov-2. *Nature microbiology*, 5(4):536–544, 2020. doi: 10.1038/s41564-020-0695-z.

[24] Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. A novel coronavirus from patients with pneumonia in china, 2019. *New England journal of medicine*, 382 (8):727–733, 2020. doi: 10.1056/NEJMoa2001017.

[25] Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, et al. Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *cell*, 181(2):271–280, 2020. doi: 10.1016/j.cell.2020.02.052.

[26] Peter J Halfmann, Masato Hatta, Shiho Chiba, Tadashi Maemura, Shufang Fan, Makoto Takeda, Noriko Kinoshita, Shin-ichiro Hattori, Yuko Sakai-Tagawa, Kiyoko Iwatsuki-Horimoto, et al. Transmission of sars-cov-2 in domestic cats. *New England Journal of Medicine*, 383(6):592–594, 2020. doi: 10.1056/NEJMc2013400.

[27] Thomas HC Sit, Christopher J Brackman, Sin Ming Ip, Karina WS Tam, Pierra YT Law, Esther MW To, Veronica YT Yu, Leslie D Sims, Dominic NC Tsang, Daniel KW Chu, et al. Infection of dogs with sars-cov-2. *Nature*, 586(7831):776–778, 2020. doi: 10.1038/s41586-020-2334-5.

[28] Bas B Oude Munnink, Reina S Sikkema, David F Nieuwenhuijse, Robert Jan Molenaar, Emmanuelle Munger, Richard Molenkamp, Arco Van Der Spek, Paulien Tolsma, Ariene Rietveld, Miranda Brouwer, et al. Transmission of sars-cov-2 on mink farms between humans and mink and back to humans. *Science*, 371(6525):172–177, 2021. doi: 10.1126/science.abe5901.

[29] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020. doi: 10.1080/10826068.2014.907183.

[30] Sarah Temmam, Khamsing Vongphayloth, Eduard Baquero, Sandie Munier, Massimiliano Bonomi, Béatrice Regnault, Bounsavane Douangboubpha, Yasaman Karami, Delphine Chrétien, Daosavanh Sanamxay, et al. Bat coronaviruses related to sars-cov-2 and infectious for human cells. *Nature*, 604(7905):330–336, 2022. doi: 10.1038/s41586-022-04532-4.

[31] C. Wang, Z. Liu, Z. Chen, X. Huang, M. Xu, and T. He. The establishment of reference sequence for sars-cov-2 and variation analysis. *J Med Virol*, 2020. doi: 10.1002/jmv.25762.

[32] Oude Munnink BB, Nieuwenhuijse DF, Stein M, O.'Toole Á, Haverkate M, and Mollers M. Rapid sars-cov-2 whole-genome sequencing and analysis for informed public health decision-making in the netherlands. *Nat Med*, 2020. doi: 10.1038/s41591-020-0997-y.

[33] Y. Tang, T.D.A. Serdan, L.N. Masi, S. Tang, R. Gorjao, and S.M. Hirabara. Epidemiology of covid-19 in brazil: using a mathematical model to estimate the outbreak peak and temporal evolution. *Emerg Microbes Infect*, 9:1453–6, 2020. doi: 10.1080/22221751.2020.1785337.

[34] Aysun Urhan and Thomas Abeel. Emergence of novel sars-cov-2 variants in the netherlands. *Scientific reports*, 11(1):6625, 2021. doi: 10.1038/s41598-021-85363-7.

[35] Thanh Le T, Andreadakis Z, Kumar A, Gómez Román R, Tollefsen S, and Saville M. The covid-19 vaccine development landscape. *Nature reviews. Drug discovery*, 19:305–6, 2020.

[36] Devika Singh and Soojin V Yi. On the origin and evolution of sars-cov-2. *Experimental & molecular medicine*, 53(4):537–547, 2021. doi: 10.1038/s12276-021-00604-z.

[37] Dongwan Kim, Joo-Yeon Lee, Jeong-Sun Yang, Jun Won Kim, V Narry Kim, and Hyeshik Chang. The architecture of sars-cov-2 transcriptome. *Cell*, 181(4):914–921, 2020. doi: 10.1016/j.cell.2020.04.011.

[38] Ayslan Castro Brant, Wei Tian, Vladimir Majerciak, Wei Yang, and Zhi-Ming Zheng. Sars-cov-2: from its discovery to genome structure, transcription, and replication. *Cell & bioscience*, 11:1–17, 2021. doi: 10.1186/s13578-021-00643-z.

[39] Peter V Markov, Mahan Ghafari, Martin Beer, Katrina Lythgoe, Peter Simmonds, Nikolaos I Stilianakis, and Aris Katzourakis. The evolution of sars-cov-2. *Nature Reviews Microbiology*, 21(6):361–379, 2023. doi: 10.1038/s41579-023-00878-2.

[40] Bina Choi, Manish C Choudhary, James Regan, Jeffrey A Sparks, Robert F Padera, Xueting Qiu, Isaac H Solomon, Hsiao-Hsuan Kuo, Julie Boucau, Kathryn Bowman, et al. Persistence and evolution of sars-cov-2 in an immunocompromised host. *New England Journal of Medicine*, 383(23):2291–2293, 2020. doi: 10.1056/NEJMc2031364.

[41] Zuoshuang Xiang, Wenjie Zheng, and Yongqun He. Bbp: Brucella genome annotation with literature mining and curation. *BMC bioinformatics*, 7:1–14, 2006. doi: 10.1186/1471-2105-7-347.

[42] Yun-Juan Chang, Miriam Land, Loren Hauser, Olga Chertkov, Tijana Glavina Del Rio, Matt Nolan, Alex Copeland, Hope Tice, Jan-Fang Cheng, Susan Lucas, et al. Non-contiguous finished genome sequence and contextual data of the filamentous soil bacterium ktedonobacter racemifer type strain (sosp1-21 t). *Standards in genomic sciences*, 5:97–111, 2011. doi: 10.4056/sigs.2114901.

[43] Alejandro Rodríguez-Gijón, Julia K Nuy, Maliheh Mehrshad, Moritz Buck, Frederik Schulz, Tanja Woyke, and Sarahi L Garcia. A genomic perspective across earth's microbiomes reveals that genome size in archaea and bacteria is linked to ecosystem type and trophic strategy. *Frontiers in microbiology*, 12:761869, 2022. doi: 10.3389/fmicb.2021.761869.

[44] N Stein. Synteny (syntenic genes). In *Brenner's Encyclopedia of Genetics*, pages 623–626. Elsevier, 2013. doi: 10.1016/B978-0-12-374984-0.01508-4.

[45] Alex N Salazar and Thomas Abeel. Approximate, simultaneous comparison of microbial genome architectures via syntenic anchoring of quiver representations. *Bioinformatics*, 34(17):i732–i742, 2018. doi: 10.1093/bioinformatics/bty614.

[46] Aysun Urhan, Bianca-Maria Cosma, Ashlee M Earl, Abigail L Manson, and Thomas Abeel. Safpred: Synteny-aware gene function prediction for bacteria using protein embeddings. *Bioinformatics*, 40(6):btae328, 2024. doi: 10.1093/bioinformatics/btae328.

[47] Yanbin Yin, Han Zhang, Victor Olman, and Ying Xu. Genomic arrangement of bacterial operons is constrained by biological pathways encoded in the genome. *Proceedings of the National Academy of Sciences*, 107(14):6310–6315, 2010. doi: 10.1073/pnas.0911237107.

[48] Julia A Schwartzman, Francois Lebreton, Rauf Salamzade, Terrance Shea, Melissa J Martin, Katharina Schaufler, Aysun Urhan, Thomas Abeel, Ilana LBC Camargo, Bruna F Sgardioli, et al. Global diversity of enterococci and description of 18 previously unknown species. *Proceedings of the National Academy of Sciences*, 121(10):e2310852121, 2024. doi: 10.1073/pnas.2310852121.

[49] Jerónimo Rodríguez-Beltrán, Javier DelaFuente, Ricardo Leon-Sampedro, R Craig MacLean, and Alvaro San Millan. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nature Reviews Microbiology*, 19(6):347–359, 2021. doi: 10.1038/s41579-020-00497-1.

[50] Keira A Cohen, Thomas Abeel, Abigail Manson McGuire, Christopher A Desjardins, Vanisha Munsamy, Terrance P Shea, Bruce J Walker, Nonkqubela Bantubani, Deepak V Almeida, Lucia Alvarado, et al. Evolution of extensively drug-resistant tuberculosis over four decades: whole genome sequencing and dating analysis of mycobacterium tuberculosis isolates from kwazulu-natal. *PLoS medicine*, 12(9):e1001880, 2015. doi: 10.1371/journal.pmed.1001880.

[51] Elizabeth J Klemm, Vanessa K Wong, and Gordon Dougan. Emergence of dominant multidrug-resistant bacterial clades: Lessons from history and whole-genome sequencing. *Proceedings of the National Academy of Sciences*, 115(51):12872–12877, 2018. doi: 10.1073/pnas.1717162115.

[52] Evelina Tacconelli, Elena Carrara, Alessia Savoldi, Stephan Harbarth, Marc Mendelson, Dominique L Monnet, Céline Pulcini, Gunnar Kahlmeter, Jan Kluytmans, Yehuda Carmeli, et al. Discovery, research, and development of new antibiotics: the who

priority list of antibiotic-resistant bacteria and tuberculosis. *The Lancet infectious diseases*, 18(3):318–327, 2018. doi: 10.1016/S1473-3099(17)30753-3.

[53] Xuemei Zhen, Cecilia Stålsby Lundborg, Xueshan Sun, Xiaoqian Hu, and Hengjin Dong. Economic burden of antibiotic resistance in eskape organisms: a systematic review. *Antimicrobial Resistance & Infection Control*, 8:1–23, 2019. doi: 10.1186/s13756-019-0590-7.

[54] Jack N Pendleton, Sean P Gorman, and Brendan F Gilmore. Clinical relevance of the eskape pathogens. *Expert review of anti-infective therapy*, 11(3):297–308, 2013. doi: 10.1586/eri.13.12.

[55] François Lebreton, Willem van Schaik, Abigail Manson McGuire, Paul Godfrey, Allison Griggs, Varun Mazumdar, Jukka Corander, Lu Cheng, Sakina Saif, Sarah Young, et al. Emergence of epidemic multidrug-resistant enterococcus faecium from animal and commensal strains. *MBio*, 4(4):10–1128, 2013. doi: 10.1128/mbio.00534-13.

[56] Paul Baumann. Isolation of acinetobacter from soil and water. *Journal of bacteriology*, 96(1):39–42, 1968. doi: 10.1128/jb.96.1.39-42.1968.

[57] Luísa CS Antunes, Paolo Visca, and Kevin J Towner. Acinetobacter baumannii: evolution of a global pathogen. *Pathogens and disease*, 71(3):292–301, 2014. doi: 10.1111/2049-632X.12125.

[58] Aoife Howard, Michael O'Donoghue, Audrey Feeney, and Roy D Sleator. Acinetobacter baumannii: an emerging opportunistic pathogen. *Virulence*, 3(3):243–250, 2012. doi: 10.4161/viru.19700.

[59] Wenwen Huo, Lindsay M Busch, Juan Hernandez-Bird, Efrat Hamami, Christopher W Marshall, Edward Geisinger, Vaughn S Cooper, Tim van Opijnen, Jason W Rosch, and Ralph R Isberg. Immunosuppression broadens evolutionary pathways to drug resistance and treatment failure during acinetobacter baumannii pneumonia in mice. *Nature Microbiology*, 7(6):796–809, 2022. doi: 10.1038/s41564-022-01126-8.

[60] Udomluk Leungtongkam, Rapee Thummeepak, Kannipa Tasanapak, and Sutthirat Sitthisak. Acquisition and transfer of antibiotic resistance genes in association with conjugative plasmid or class 1 integrons of acinetobacter baumannii. *PLoS One*, 13 (12):e0208468, 2018. doi: 10.1371/journal.pone.0208468.

[61] Ioannis Kyriakidis, Eleni Vasileiou, Zoi Dorothea Pana, and Athanasios Tragiannidis. Acinetobacter baumannii antibiotic resistance mechanisms. *Pathogens*, 10(3):373, 2021. doi: 10.3390/pathogens10030373.

[62] Krisztina M Papp-Wallace, Andrea Endimiani, Magdalena A Taracila, and Robert A Bonomo. Carbapenems: past, present, and future. *Antimicrobial agents and chemotherapy*, 55(11):4943–4960, 2011. doi: 10.1128/aac.00296-11.

[63] Bruno Périchon, Sylvie Goussard, Violaine Walewski, Lenka Krizova, Gustavo Cerqueira, Cheryl Murphy, Michael Feldgarden, Jennifer Wortman, Dominique

Clermont, Alexandr Nemec, et al. Identification of 50 class d β-lactamases and 65 acinetobacter-derived cephalosporinases in acinetobacter spp. *Antimicrobial agents and chemotherapy*, 58(2):936–949, 2014. doi: 10.1128/aac.01261-13.

[64] Paul G Higgins, Francisco J Pérez-Llarena, Esther Zander, Ana Fernández, Germán Bou, and Harald Seifert. Oxa-235, a novel class d β-lactamase involved in resistance to carbapenems in acinetobacter baumannii. *Antimicrobial agents and chemotherapy*, 57(5):2121–2126, 2013. doi: 10.1128/aac.02413-12.

[65] Sébastien Coyne, Patrice Courvalin, and Bruno Périchon. Efflux-mediated antibiotic resistance in acinetobacter spp. *Antimicrobial agents and chemotherapy*, 55(3):947–953, 2011. doi: 10.1128/aac.01388-10.

[66] Arta Karruli, Filomena Boccia, Massimo Gagliardi, Fabian Patauner, Maria Paola Ursi, Pino Sommese, Rosanna De Rosa, Patrizia Murino, Giuseppe Ruocco, Antonio Corcione, et al. Multidrug-resistant infections and outcome of critically ill patients with coronavirus disease 2019: a single center experience. *Microbial Drug Resistance*, 27(9):1167–1175, 2021. doi: 10.1089/mdr.2020.0489.

[67] Terence Lee, Stanley Pang, Sam Abraham, and Geoffrey W Coombs. Antimicrobial-resistant cc17 enterococcus faecium: The past, the present and the future. *Journal of global antimicrobial resistance*, 16:36–47, 2019. doi: 10.1016/j.jgar.2018.08.016.

[68] Nabil Ben Omar, Araceli Castro, Rosario Lucas, Hikmate Abriouel, Nuha MK Yousif, Charles MAP Franz, Wilhelm H Holzapfel, Pérez-Pulido Rubén, Magdalena Martínez-Canãmero, and Antonio Gálvez. Functional and safety aspects of enterococci isolated from different spanish foods. *Systematic and Applied Microbiology*, 27(1):118–130, 2004. doi: 10.1078/0723-2020-00248.

[69] Rahat Zaheer, Shaun R Cook, Ruth Barbieri, Noriko Goji, Andrew Cameron, Aaron Petkau, Rodrigo Ortega Polo, Lisa Tymensen, Courtney Stamm, Jiming Song, et al. Surveillance of enterococcus spp. reveals distinct species and antimicrobial resistance diversity across a one-health continuum. *Scientific reports*, 10(1):3937, 2020. doi: 10.1038/s41598-020-61002-5.

[70] François Lebreton, Abigail L Manson, Jose T Saavedra, Timothy J Straub, Ashlee M Earl, and Michael S Gilmore. Tracing the enterococci from paleozoic origins to the hospital. *Cell*, 169(5):849–861, 2017. doi: http://dx.doi.org/10.1016/j.cell.2017.04.027.

[71] Ana M Guzman Prieto, Willem van Schaik, Malbert RC Rogers, Teresa M Coque, Fernando Baquero, Jukka Corander, and Rob JL Willems. Global emergence and dissemination of enterococci as nosocomial pathogens: attack of the clones? *Frontiers in microbiology*, 7:198281, 2016. doi: 10.3389/fmicb.2016.00788.

[72] Mónica García-Solache and Louis B Rice. The enterococcus: a model of adaptability to its environment. *Clinical microbiology reviews*, 32(2):10–1128, 2019. doi: 10.1128%2FCMR.00058-18.

[73] Brian L Hollenbeck and Louis B Rice. Intrinsic and acquired resistance mechanisms in enterococcus. *Virulence*, 3(5):421–569, 2012. doi: 10.4161/viru.21282.

[74] Kathryn L Ruoff, Lorena De La Maza, Margaret J Murtagh, Jean D Spargo, and Mary J Ferraro. Species identities of enterococci isolated from clinical specimens. *Journal of Clinical Microbiology*, 28(3):435–437, 1990. doi: 10.1128/jcm.28.3.435-437.1990.

[75] Masateru Nishiyama, Atsushi Iguchi, and Yoshihiro Suzuki. Identification of enterococcus faecium and enterococcus faecalis as vanc-type vancomycin-resistant enterococci (vre) from sewage and river water in the provincial city of miyazaki, japan. *Journal of Environmental Science and Health, Part A*, 50(1):16–25, 2015. doi: 10.1080/10934529.2015.964599.

[76] Wioleta Chajęcka-Wierzchowska, Anna Zadernowska, and Łucja Łaniewska-Trokenheim. Virulence factors of enterococcus spp. presented in food. *LWT*, 75: 670–676, 2017. doi: 10.1016/j.lwt.2016.10.026.

[77] Mark M Huycke, Carol A Spiegel, and Michael S Gilmore. Bacteremia caused by hemolytic, high-level gentamicin-resistant enterococcus faecalis. *Antimicrobial agents and chemotherapy*, 35(8):1626–1634, 1991. doi: 10.1128/aac.35.8.1626.

[78] Robert D Fleischmann, Mark D Adams, Owen White, Rebecca A Clayton, Ewen F Kirkness, Anthony R Kerlavage, Carol J Bult, Jean-Francois Tomb, Brian A Dougherty, Joseph M Merrick, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *science*, 269(5223):496–512, 1995. doi: 10.1126/science.7542800.

[79] Nicholas J Loman and Mark J Pallen. Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology*, 13(12):787–794, 2015. doi: 10.1038/nrmicro3565.

[80] Tetsuya Hayashi, Kozo Makino, Makoto Ohnishi, Ken Kurokawa, Kazuo Ishii, Katsushi Yokoyama, Chang-Gyun Han, Eiichi Ohtsubo, Keisuke Nakayama, Takahiro Murata, et al. Complete genome sequence of enterohemorrhagic eschelichia coli o157: H7 and genomic comparison with a laboratory strain k-12. *DNA research*, 8(1): 11–22, 2001. doi: 10.1093/dnares/8.1.11.

[81] Jessica Alföldi and Kerstin Lindblad-Toh. Comparative genomics as a tool to understand evolution and disease. *Genome research*, 23(7):1063–1068, 2013. doi: 10.1101/gr.157503.113.

[82] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009. doi: 10.1073/pnas.0903103106.

[83] Sharon R Grossman, Kristian G Andersen, Ilya Shlyakhter, Shervin Tabrizi, Sarah Winnicki, Angela Yen, Daniel J Park, Dustin Griesemer, Elinor K Karlsson, Sunny H Wong, et al. Identifying recent adaptations in large-scale genomic data. *Cell*, 152(4): 703–713, 2013. doi: 10.1016/j.cell.2013.01.035.

[84] Conrad L Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, et al. Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020:baaa062, 2020. doi: 10.1093/database/baaa062.

[85] Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. Gtdb: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794, 09 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab776. URL 10.1093/nar/gkab776.

[86] Supratim Mukherjee, Dimitri Stamatis, Cindy Tianqing Li, Galina Ovchinnikova, Jon Bertsch, Jagadish Chandrabose Sundaramurthi, Mahathi Kandimalla, Paul A Nicolopoulos, Alessandro Favognano, I-Min A Chen, et al. Twenty-five years of genomes online database (gold): data updates and new features in v. 9. *Nucleic acids research*, 51(D1):D957–D963, 2023. doi: 10.1093/nar/gkac974.

[87] Yuelong Shu and John McCauley. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13):30494, 2017. doi: 10.2807/1560-7917.ES.2017.22.13.30494.

[88] Stefan Elbe and Gemma Buckland-Merrett. Data, disease and diplomacy: Gisaid's innovative contribution to global health. *Global challenges*, 1(1):33–46, 2017. doi: 10.1002/gch2.1018.

[89] Adam C Retchless and Jeffrey G Lawrence. Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. *Proceedings of the National Academy of Sciences*, 107(25):11453–11458, 2010. doi: 10.1073/pnas.1001291107.

[90] Eduardo Corel, Philippe Lopez, Raphaël Méheust, and Eric Bapteste. Network-thinking: graphs to analyze microbial complexity and evolution. *Trends in Microbiology*, 24(3):224–237, 2016. doi: 10.1016/j.tim.2015.12.003.

[91] Carlos A Ruiz-Perez, Roth E Conrad, and Konstantinos T Konstantinidis. Microbeannotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes. *BMC bioinformatics*, 22:1–16, 2021. doi: 10.1186/s12859-020-03940-5.

[92] Yibo Dong, Chang Li, Kami Kim, Liwang Cui, and Xiaoming Liu. Genome annotation of disease-causing microorganisms. *Briefings in Bioinformatics*, 22(2):845–854, 2021. doi: 10.1093/bib/bbab004.

[93] Hervé Tettelin, Vega Masignani, Michael J Cieslewicz, Claudio Donati, Duccio Medini, Naomi L Ward, Samuel V Angiuoli, Jonathan Crabtree, Amanda L Jones, A Scott Durkin, et al. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, 102(39):13950–13955, 2005. doi: 10.1073/pnas.0506758102.

[94] Bruno Contreras-Moreira and Pablo Vinuesa. Get_homologues, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and environmental microbiology*, 79(24):7696–7701, 2013. doi: 10.1128/AEM.02411-13.

[95] Aysun Urhan and Thomas Abeel. A comparative study of pan-genome methods for microbial organisms: Acinetobacter baumannii pan-genome reveals structural variation in antimicrobial resistance-carrying plasmids. *Microbial Genomics*, 7(11): 000690, 2021. doi: 10.1099/mgen.0.000690.

[96] Gerry Tonkin-Hill, Neil MacAlasdair, Christopher Ruis, Aaron Weimann, Gal Horesh, John A Lees, Rebecca A Gladstone, Stephanie Lo, Christopher Beaudoin, R Andres Floto, et al. Producing polished prokaryotic pangenomes with the panaroo pipeline. *Genome biology*, 21:1–21, 2020. doi: 10.1186/s13059-020-02090-4.

[97] S.C. Bayliss, H.A. Thorpe, N.M. Coyle, S.K. Sheppard, and E.J. Feil. Pirate: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *GigaScience*, 8(598391), 2019. doi: 10.1093/gigascience/giz119.

[98] Guillaume Gautreau, Adelme Bazin, Mathieu Gachet, Rémi Planel, Laura Burlot, Mathieu Dubois, Amandine Perrin, Claudine Médigue, Alexandra Calteau, Stéphane Cruveiller, et al. Ppanggolin: depicting microbial diversity via a partitioned pangenome graph. *PLoS computational biology*, 16(3):e1007732, 2020. doi: 10.1371/journal.pcbi.1007732.

[99] Kai Kupferschmidt. New coronavirus variants could cause more reinfections, require updated vaccines. *Science*, 10, 2021. doi: 10.1126/science.abg6028.

[100] Yannick Mahlich, Chengsheng Zhu, Henri Chung, Pavan K Velaga, M Clara De Paolis Kaluza, Predrag Radivojac, Iddo Friedberg, and Yana Bromberg. Learning from the unknown: exploring the range of bacterial functionality. *Nucleic Acids Research*, 51(19):10162–10175, 2023. doi: 10.1093/nar/gkad757.

[101] Stephen Nayfach, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, David Paez-Espino, I-Min Chen, Marcel Huntemann, et al. A genomic catalog of earth's microbiomes. *Nature biotechnology*, 39(4):499–509, 2021. doi: 10.1038/s41587-020-0718-6.

[102] Computational pan-genomics: status, promises and challenges. *Briefings in bioinformatics*, 19(1):118–135, 2018. doi: 10.1093/bib/bbw089.

[103] Konstantin Weißenow, Michael Heinzinger, and Burkhard Rost. Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure*, 2022. doi: 10.1016/j.str.2022.05.001.

[104] Danielle Miller, Adi Stern, and David Burstein. Deciphering microbial gene function using natural language processing. *Nature Communications*, 13(1):5731, 2022. doi: 10.1038/s41467-022-33397-4.

2

Emergence of novel SARS-CoV-2 variants in the Netherlands

"If you want the present to be different from the past, study the past."

— Baruch Spinoza

This chapter is based on Aysun Urhan and Thomas Abeel. Emergence of novel sars-cov-2 variants in the Netherlands. *Scientific reports*, 11(1):6625, 2021. [1].

Abstract

Coronavirus disease 2019 (COVID-19) has emerged in December 2019 when the first case was reported in Wuhan, China and turned into a pandemic with 27 million (September 9th) cases. Currently, there are over 95,000 complete genome sequences of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus causing COVID-19, in public databases, accompanying a growing number of studies. Nevertheless, there is still much to learn about the viral population variation when the virus is evolving as it continues to spread. We have analyzed SARS-CoV-2 genomes to identify the most variant sites, as well as the stable, conserved ones in samples collected in the Netherlands until June 2020. We identified the most frequent mutations in different geographies. We also performed a phylogenetic study focused on the Netherlands to detect novel variants emerging in the late stages of the pandemic and forming local clusters. We investigated the S and N proteins on SARS-CoV-2 genomes in the Netherlands and found the most variant and stable sites to guide development of diagnostics assays and vaccines. We observed that while the SARS-CoV-2 genome has accumulated mutations, diverging from reference sequence, the variation landscape is dominated by four mutations globally, suggesting the current reference does not represent the virus samples circulating currently. In addition, we detected novel variants of SARS-CoV-2 almost unique to the Netherlands that form localized clusters and region-specific sub-populations indicating community spread. We explored SARS-CoV-2 variants in the Netherlands until June 2020 within a global context; our results provide insight into the viral population diversity for localized efforts in tracking the transmission of COVID-19, as well as sequenced-based approaches in diagnostics and therapeutics. We emphasize that little diversity is observed globally in recent samples despite the increased number of mutations relative to the established reference sequence. We suggest sequence-based analyses should opt for a consensus representation to adequately cover the genomic variation observed to speed up diagnostics and vaccine design.

2.1 Introduction

In late December 2019, officials had reported the first case of coronavirus disease 2019 (COVID-19) in China, caused by a novel type of coronavirus named severe acute respiratory syndrome coronavirus 2, (SARS-CoV-2) [2]. COVID-19 has consequently led to the global pandemic we are going through at the moment; according a situation report released by the World Health Organization (September 9th) there are 27.4 million cases and almost 900,000 deaths in total [3]. SARS-CoV-2 has been placed under the betacoronavirus genus, closest relatives being bat and pangolin coronaviruses [4, 5].

Despite having major commonalities with recent outbreaks of betacoronaviruses, SARS in 2002 and Middle East respiratory syndrome (MERS) in 2012, it is unprecedented not only in its ease of spread but also in the collective effort of several international scientists to investigate and understand the biology of the disease and the virus causing it since the day the first complete SARS-CoV-2 genome sequence had been published [5–8]. Early studies on the SARS-CoV-2 genome has shown its closest relative, in terms of sequence identity, to be the bat coronavirus RaTF13 with over 93.1% match in the spike (S) protein and >96% sequence identity overall [6, 9]. Immediately a reference sequence had been established [10], paving the way for the exponential growth in both the number and the scale of studies

on the SARS-CoV-2 genome [11–16].

At the moment, the GISAID database has established the SARS-CoV-2 population consists of six major clades: G, GH, GR, L, S and V [17]. There is a growing number of studies on the genetic variability of SARS-CoV-2 relative to the reference genome [18–21]. From previous viral outbreaks, it is known that as part of the natural evolution of a virus, subpopulations of clades that can affect the severity of a disease emerge and alter the trajectory of a pandemic [22]. It has been reported that while the two major structural proteins, S and nucleocapsid (N) protein are rich in sites of episodic selection, ORF3a and ORF8 had also been shown to carry a lot of mutations [23].

In this study, we investigate the genetic variability of SARS-CoV-2 genomes in the Netherlands until mid-2020, in the context of global viral population with a particular focus on the later stages of the first wave of the pandemic (from early April to the end of May). We have identified the most variant proteins in the SARS-CoV-2 genome, as well as the most frequent mutations in the Netherlands that also showed high dominance in the rest of the world. We found relatively conserved regions in the S and N proteins of SARS-CoV-2, as well as frequent mutations on the target regions of some RT-qPCR diagnostic tests. Tracing the viral genome since its first introduction into the Netherlands, we detected novel mutations unique to the Netherlands, and local clusters of distinct viral sub-populations emerging in different provinces. Our work provides valuable insights into the regional variance of SARS-CoV-2 populations in the Netherlands that would prove beneficial for localized efforts in tracking routes of transmission through genetic variation, primer/probe design in RT-qPCR tests targeting viral sub-populations. We recommend that emergent variants are examined when developing sequence-based diagnostics, vaccines or therapeutics against COVID-19. In order to do so, genomic surveillance needs to continue at a sufficiently high level throughout the course of the pandemic.

2.2 Methods

Our study of SARS-COV-2 genomes in the Netherlands consists of three main steps: data retrieval, preprocessing and multiple sequence alignment, phylogenetic tree construction and sequence variation analysis. We have also analyzed the global phylogenetic tree of SARS-COV-2 genomes using additional metadata on patients and travel history.

2.2.1 Data retrieval and preprocessing, and multiple sequence alignment

Complete, high quality (number of undetermined bases less than 1% of the whole sequence) genome sequences of SARS-COV-2 that were isolated from human hosts only were obtained from GISAID, NCBI and China's National Genomics Data Center (NGDC) on June 13th [17, 24, 25]. The dataset contained 29,503 sequences with unique identifiers in total, including the Wuhan-Hu-1 reference sequence (accession ID NC_045512.2). The "Collection date" field was also extracted for all sequences, and it is referred to as "date" throughout this work. The acknowledgment table for GISAID sequences can be found in Supplementary file 2 and the full list of sequence identifiers for NCBI and NGDC records are provided in Supplementary file 3.

All sequences were aligned against the Wuhan-Hu-1 reference using MAFFT (v7.46)

with the FFT-NS-fragment option, and the alignment was filtered to remove identical sequences to obtain 24,365 non-redundant genomes [26].

2.2.2 Sequence variation analysis

In order to determine mutations, the filtered multiple sequence alignment was trimmed to remove gaps from the Wuhan-Hu-1 reference (accession ID NC_045512.2) and used as input to the coronapp web application to obtain nucleotide variations [27]. Next, the trimmed alignment was used to cluster genomes according to the nomenclature on GISAID website. We assigned all 29,503 sequences to one of the clades S, L, V, G, GH and GR.

We retrieved primer/probe sequence sets released by US CDC, WHO, Institut Pasteur and China CDC, and identified mutations which overlap with these sequences [28–31].

2.2.3 Phylogenetic tree construction

The maximum likelihood phylogenetic tree for the samples in the Netherlands (1338 genomes in total) was built using IQ-TREE (v2.05) with GTR model, allowing to collapse non-zero branches, and ultrafast bootstrap with 1000 replicates [32]. A time tree was also constructed for dating branches in IQ-TREE (v2.05) and the final tree was rooted at the ancestral node of S clades in the tree using ETE Toolkit (v3.1.1) [33]. ETE was also used for visualizing tree. Collection date and region (within the Netherlands) fields of each sequence record (if available) were retrieved, and utilized to infer the spread of variants within the Netherlands.

2.3 Results

The global SARS-CoV-2 dataset was filtered considering only the sequence quality, hence we observe a large discrepancy in the distribution of genomes across different countries. Initially, most sequencing effort was concentrated in China and other countries where the outbreak had begun. However, at the time of data retrieval (June 13th) the dataset is dominated by samples from the UK, the USA and Australia (Table 2.1 and Fig. 2.1).

Since we have not corrected for sampling differences, in this section, we will provide a view of the current situation of pandemic mainly in Europe, focusing on the Netherlands, where most of the viral genomes are available today (Fig. 2.1). While initially many genome sequences were generated, by April virtually no sequences were determined.

2.3.1 Distinct genetic patterns in the SARS-CoV-2 population emerge across the globe

In order to get a broad overview of the viral diversity throughout the pandemic, we monitored changes in proportion of clades in time using the clade definitions proposed by the GISAID database [17]. We observe the distribution of different clades in the Netherlands resemble that of Australia where the first samples are genetically diverse and there is no dominating variant (subplots in Fig. 2.2, see Fig. S2.3 for absolute number of genomes). A similar pattern is seen in other European counties such as the UK and Belgium, while the USA, Canada and Denmark have distinct trajectories with GH clade dominating the population (Figs. S2.4 and S2.5). Also note that clade S has gradually faded out despite its

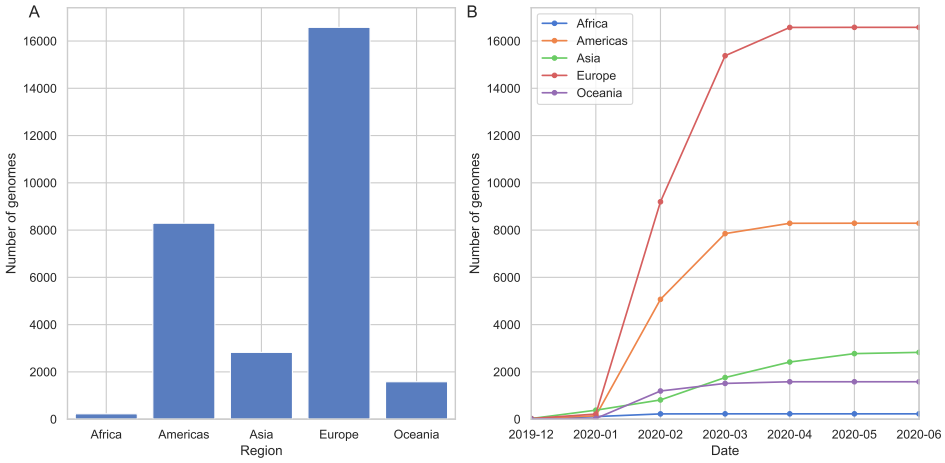


Figure 2.1: Distribution of SARS-CoV-2 genomes across five continents: (A) total number of genomes is shown on y-axis and the regions in x-axis, (B) the change in number of genomes collected over the course of pandemic, x-axis shows the collection date. See (B) for colors of each continent.

Table 2.1: 20 countries with the largest number of genomes in the dataset.

Country	Number of genomes
The UK	9641
The USA	7294
Australia	1398
The Netherlands	1338
Spain	886
India	710
China	651
Belgium	645
Denmark	581
Canada	560
Portugal	500
Iceland	481
France	376
Sweden	353
Switzerland	314
Singapore	285
Austria	247
Russia	218
Germany	209
Luxembourg	192

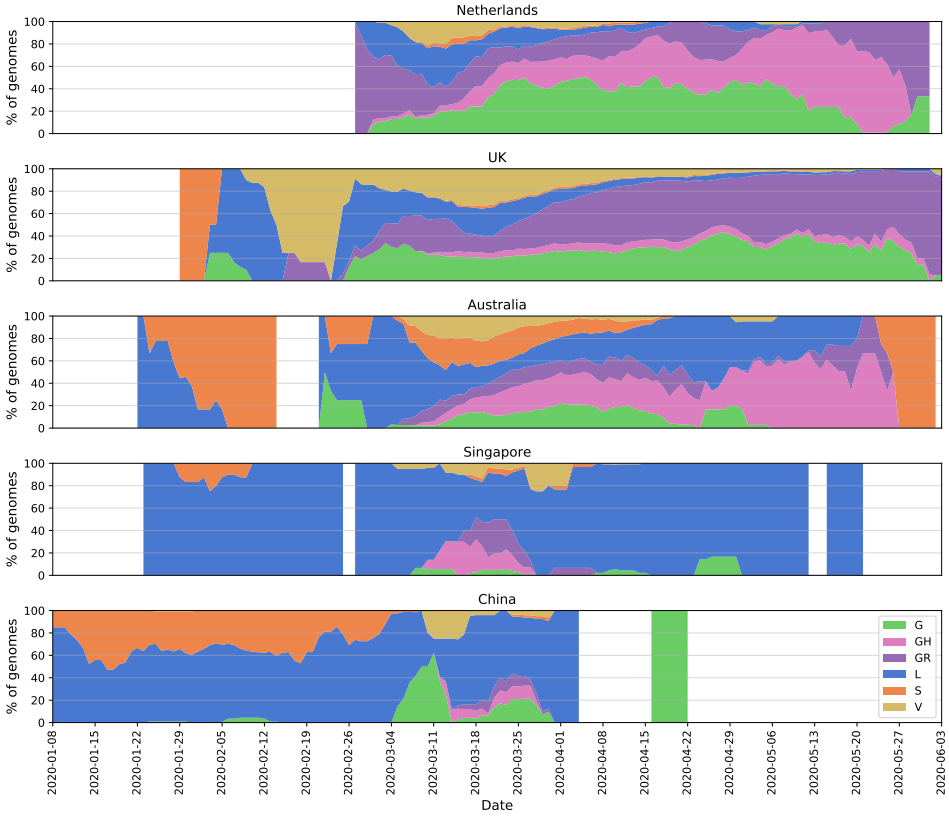


Figure 2.2: Distribution of SARS-CoV-2 clades in a selection among the 12 most sampled countries in comparison to the Netherlands: y-axis shows a 7-day moving average of the relative abundance of the six clades, and x-axis shows the collection date. (See the legend for clade names and colors) Intervals with fewer than one genome per day were discarded. See Fig. S2.3 for absolute number of genomes.

high prevalence before April in several countries, this is particularly noticeable in Australia, China (Fig. 2.2), the USA, Spain and Canada (Figs. S2.4 and S2.5).

Viral diversity can be observed more clearly when put into context with less diverse populations in other countries where the outbreak had begun the earliest. For instance, China, Singapore and Italy had experienced the outbreak the earliest in the world, and there are only few of the major clades circulating (Fig. 2.2, Italy not shown due to small sample size, see Figs. S2.3 and S2.4 for other countries). China had opted for possibly the most severe restrictions; similarly in Singapore, the initial cases of COVID-19 had been followed up with strict precautions, preventing both the spread and new introduction of the virus. While it is tricky to formulate any clear hypothesis since there has not been any public data submission from these countries since April, it is certainly interesting to see the contrast between them and countries where COVID-19 arrived at relatively late stages of the pandemic, such as the Netherlands, the UK and Australia. However, more data is

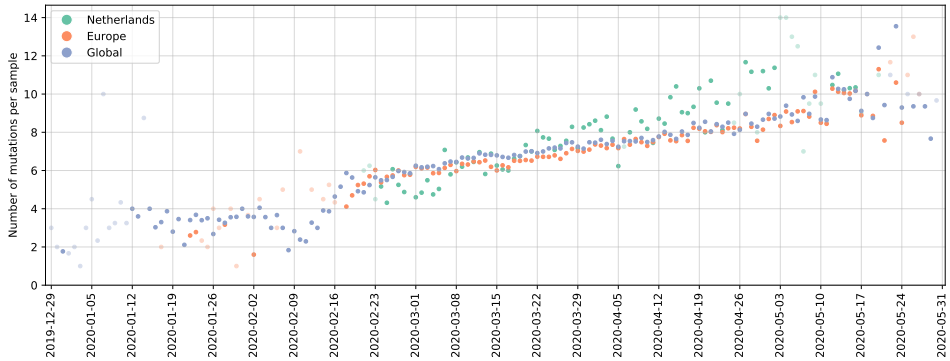


Figure 2.3: Number of mutations per sample per day over the course of pandemic in the Netherlands (green), Europe (orange) and globally (blue): each point is the average number of mutations observed in samples collected on the same date in the Netherlands (green), Europe (orange) and globally (blue), x-axis shows collection date. Data points corresponding to days with fewer than five samples are colored transparently to indicate uncertainty.

needed to form a better understanding of the population structures.

2.3.2 Evolution of the SARS-CoV-2 genome and increased mutation frequency in hotspot regions

To assess the mutational landscape and its impact as the pandemic progressed, we investigated dominant mutations across time in the viral population. It is essential to monitor these changes in the SARS-CoV-2 genome to identify conserved sites relevant for designing therapeutics and vaccines, as well as to study the viral evolution during a pandemic. Currently, each new sample has on average around ten mutation sites in total compared to the Wuhan-Hu-1 reference (accession ID NC_045512.2) in the Netherlands where the trajectory has been in parallel with those in Europe and the world (Fig. 2.3 shows number of mutations per sample each day from December 2019 to June 2020). Clearly showing a divergence away from the original reference.

In particular, the S and N proteins have both been reported as the most variant proteins in the SARS-CoV-2 genome [23, 34]. S:D614G and N:RG203KR amino acid changes comprise a large fraction of the mutation in these regions (Fig. 2.4); former being one of the mutation that defines G, GR and GH clades. Apart from carrying the majority of mutations observed in the populations, both proteins play an important role in RT-qPCR based diagnostic tests as well as vaccine and drug development [35]. The S protein has been investigated in great detail for its significance in binding to the host cell and a potential target for COVID-19 treatment and vaccine design [36–38]. In a recent study, prior information on the SARS-CoV S and N proteins, and their known epitopes were combined to identify regions in the SARS-CoV-2 genome that could potentially serve as epitopes for B-cells and T-cells [39]. In Fig. 2.4 we have highlighted the predicted epitope regions from [39] with mutations using red rectangles on the x-axis. The authors confirmed that the most abundant mutations in these regions, S:D614G in particular, should be taken into account for vaccine design and development of treatments. We also note that the prevalence of

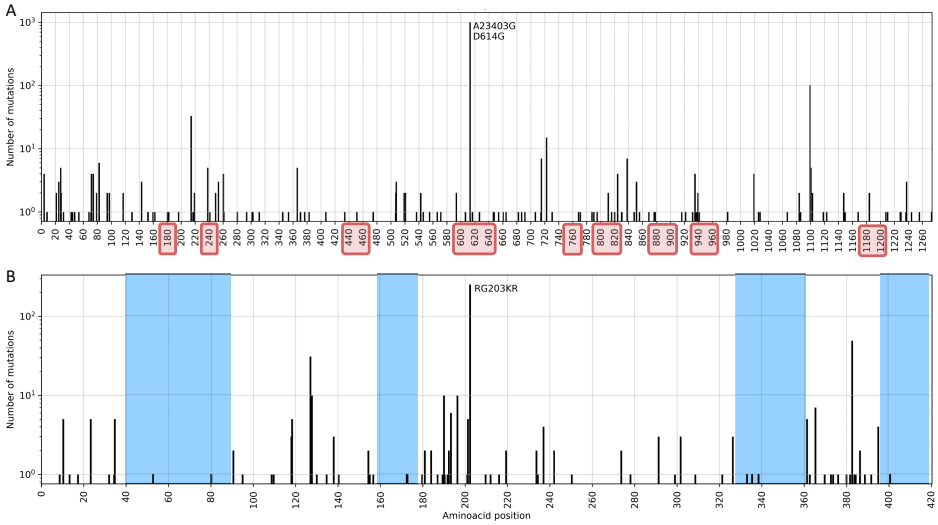


Figure 2.4: Total number of nucleotide mutations in the S (A) and N (B) proteins in samples from the Netherlands are displayed on y-axis; predicted epitope regions from [39] are shown with red rectangles on x-axis, and conserved sites (free of mutations) on N protein are shaded in blue in (B). X-axis tick marks are labelled with the corresponding amino-acid position to complement the mutation annotations.

S:D614G variant has steadily increased over the course of the pandemic: it is observed in all the sequences sampled recently in the world (Fig. S2.6).

In order to determine the appropriate primers to use when diagnosing patients with RT-PCR tests or when designing novel primer/probe sequences, variations in the nucleotide sequence should be considered since it plays a crucial role in achieving accurate tests [10, 20]. We have identified mutations on target regions of primer/probe assay sets most used in the Netherlands. We found that assay sequences published by US CDC had fewer than ten genomes with mutations, and those from WHO had fewer than 18 out of 1338 genomes. We have also checked the assay sets from China CDC and Institut Pasteur, even though they are not in use in the Netherlands to our knowledge. We report 73 genomes (5%) with mutations on ORF1ab and 771 genomes (57%) with mutations on N protein for the sets released by China CDC, and fewer than six genomes for Institut Pasteur. Without being too specific, amino-acid positions from 40 to 90, 160–170, 330–360 and 400–420 on N protein appear to be relatively conserved sites, free of any mutations and could potentially be utilized as primer sequences (blue shaded regions in Fig. 2.4). The N protein is recommended as a screening assay by the WHO as well, and is utilized in many countries other than the Netherlands [29]. Further investigation of the location and frequency of mutations indicate the existence of conserved regions and show a general preference for non-silent changes in the genome (see “Supplemental Text”).

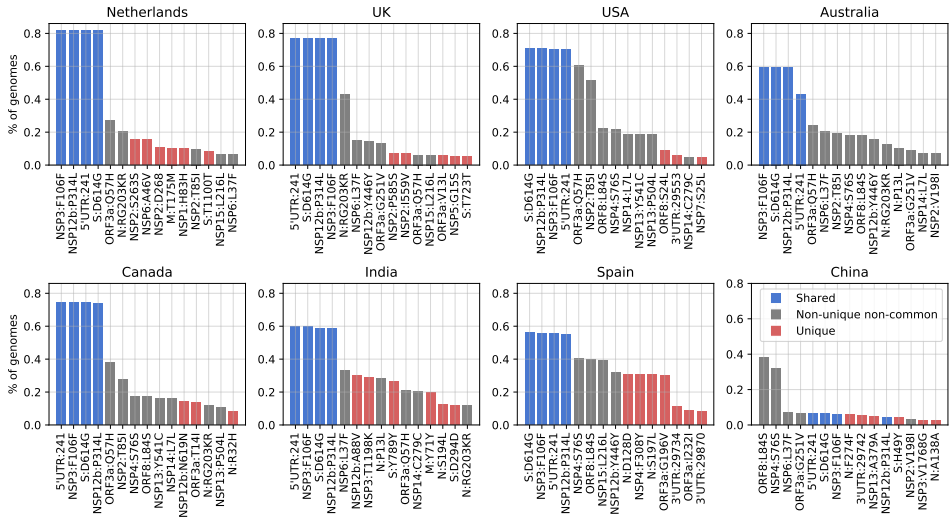


Figure 2.5: Total frequency (% of genomes) of the top 15 mutations in the most-sampled countries in our dataset: x-axes are the top 15 mutations and y-axes show the frequency (number of mutations per sample) of mutations. Blue bars are mutations shared across all these countries in the top 15, while the red bars are unique to that one country in the top 15 and the gray bars are non-unique and non-common variants where the mutation is observed in more than one country.

2.3.3 Population of SARS-CoV-2 is dominated by four mutations globally while emergence of locally distinct variants indicates local outbreaks

To study the global SARS-CoV-2 population and viral diversity in more detail, and observe the mutational landscape in the Netherlands within a global context, we have identified the most abundant mutations in our dataset. In addition to S:D614G and N:RG203KR, several other mutations, NSP12b:P314L, NSP3:F106F and 5'UTR:241 in particular, appear to dominate the most frequent mutations in the world; Fig. 2.5 shows the 15 most dominant SNPs in some of the most-sampled countries in our dataset. Due to over-representation of few European countries, it is difficult to comment on the geographical dominance of any mutations. However, four mutations, S:D614G, NSP12b:P314L, NSP3:F106F and 5'UTR:241 (blue bars in Fig. 2.5) are established within the global collection genomes, except for China where these mutations have very low frequencies.

While we observe a diverse mutational landscape in Australia, India and Spain, the viral population in China has remained relatively homogenous and with very few variants compared to the Wuhan-Hu-1 reference. The most frequent mutation is ORF8:L84S, which defines the S clade that appears to be fading out even though it had been circulating since the beginning of the pandemic along with the L clade. Recently, a possible link between two mutations, ORF8:L84S and NSP4:S76S, has been suggested, we also observed they co-occur several times outside of Europe; in China, the USA, Australia and Canada [40]. While keeping in mind that we do not have any sequences collected after April from China, we note a few region-specific mutations: first one being ORF8:L84S, which is more frequent

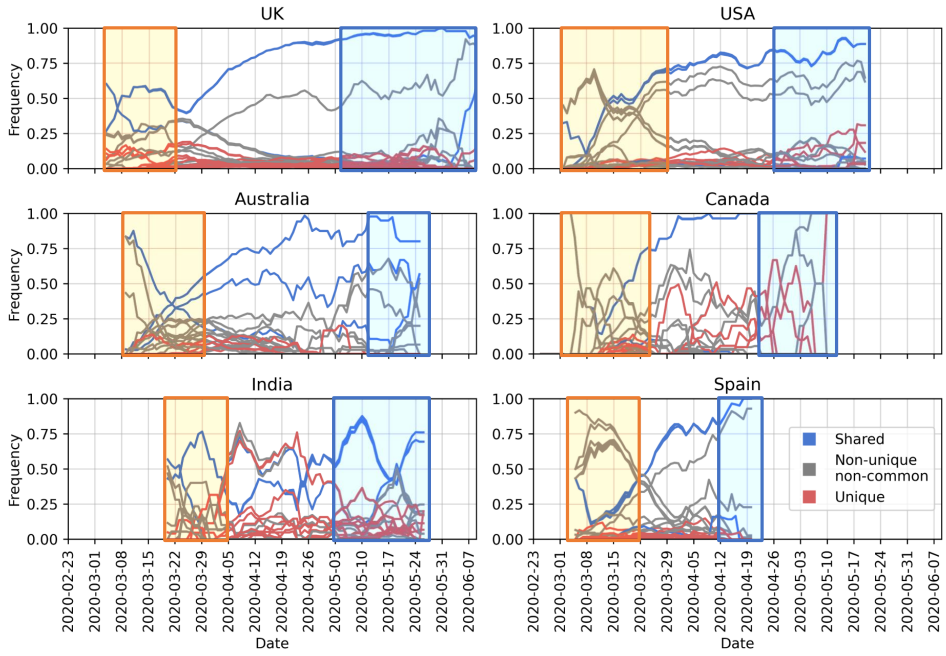


Figure 2.6: Change in frequency of the top 15 mutations in the most-sampled countries in our dataset: y-axes show mutation frequency (number of mutations per sample) averaged over a period of 7 days where periods with fewer than one sample per day were removed and x-axes show the collection date. Line colors were kept consistent with Fig. 2.5: blue lines are mutations shared across all these countries in the top 15, while the red lines are unique to that one country in the top 15 and the gray lines are non-unique and non-common variants where the mutation is observed in more than one country. Areas highlighted in yellow and blue as, mentioned in text, to indicate pre-lockdown and post-lockdown.

in the USA and China and, second is NSP6:L37F which is frequent in in Australia and the USA.

Considering the fluctuations in rate of sequencing, and over-representation of samples from the USA, the UK and Europe in general, it is difficult to comment on the geographical spread. Nevertheless, when we look into the frequency of the top four mutations, S:D614G, NSP12b:P314L, NSP3:F106F and 5'UTR:241 over the course of pandemic, we see a steady increase of their abundance in the viral population, regardless of the date of introduction in each country (Fig. 2.6).

A common pattern emerges in how shared and rarer mutations change in frequency in time: in the early phase of the pandemic, the viral population is diverse with relatively few mutations shared among all countries (areas highlighted in yellow in Fig. 2.6: first 2 weeks of March in the Netherlands, the UK, Australia, Canada and Spain [41–45], also late March in the USA and India [46, 47]). From mid-March to end of April when strict measures against travel were imposed universally, frequency of shared mutations increase more rapidly. As the pandemic progresses, the four most abundant mutations shared across each country (blue lines in Fig. 2.6) become well-established as part of the viral genome. In May,

however, restrictions on domestic travel were slowly eased [48–52], which we presume allowed for regional transmission, leading to again an increase in unique/rare mutations (areas highlighted in blue in Fig. 2.6) as they spread and form local clusters of variants. In addition, for most of the countries, number of sequences peaked in March or April, and has been on decline since then, except for India (Fig. S2.7). Hence it does not appear to be driving the changes in the frequency of rare/unique. Abundance of these unique variants suggests community-driven spread, which can be elaborated by monitoring such variants to detect super-spreading events.

To assess the impact of lockdown attempts to control the pandemic on the viral diversity we investigated Dutch viral samples in detail. It is non-trivial to relate lockdown status to the viral population diversity across countries; while all measures to control COVID-19 have been reported throughout the pandemic, it is highly likely that there are both national and regional differences in their implementation as well as their impact on human behavior, especially in federal governments such as the USA, Australia and Canada, where regional governments play an influential role. For that reason, we focus on the Netherlands to understand this pattern better: we have placed the major milestones in national response against COVID-19 in the Netherlands along with the mutation frequencies below in Fig. 2.7A in comparison to number of sequences collected in Fig. 2.7B. We observe local/rare mutations to increase in frequency around the same time as restrictions are relaxed. One other explanation for the increase in frequency of rare mutations could be the gradual expansion of testing and sequencing capacity. Testing in the Netherlands was almost exclusively available to healthcare workers due to limited capacity until May [53]. It is conceivable that testing different groups of individuals has made it possible to collect more diverse samples of the virus.

2.3.4 Introduction of COVID-19 in the Netherlands and local clusters with high genomic diversity

Next, we examined the Dutch phylogenetic tree to better understand the dynamics of COVID-19 in the Netherlands: from its introduction in the earliest samples to its further spread through localized infection clusters. We have identified multiple points of introduction in different provinces via highly diverse samples of virus. As the pandemic progresses, we see deeper branching in the tree with unique, localized mutations as well as similar patterns of evolution emerge in separate locations. While the virus population carries an increased number of mutations in general, these mutations are localized in their own clusters with little genomic diversity.

We observe two separate sections on the radial tree in Fig. 2.8, representing the diversity of introduction to the Netherlands in terms of both the viral genome and location. First, at the top, starting from around 12 o'clock to 3 o'clock consists of some of the earliest samples from early March of V, L and S clades (denoted with a blue arch and text "Early March"). This is further broken down into four sections numbered from 1 to 4 where the second section is the early outbreak in Noord-Brabant in parallel with the first case reports [11]. However, the remaining sections are mixed in location and date as we encounter samples isolated from Limburg, Zuid-Holland, Gelderland and Utrecht, also from later into the pandemic in late March and early April.

The second point of introduction is from 4 to 6 o'clock on the tree, denoted as "late

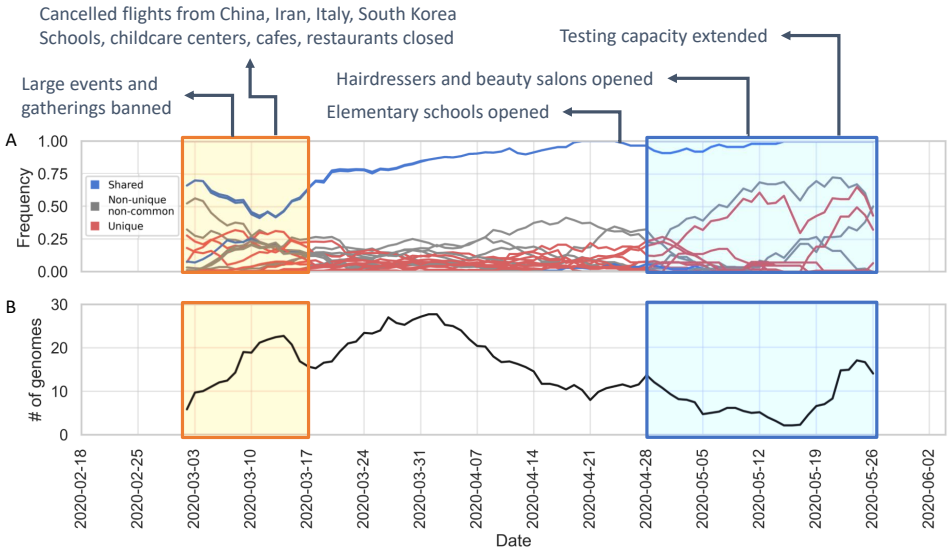


Figure 2.7: (A) Change in frequency of the top 15 mutations in the Netherlands, averaged over a period of 7 days and removed periods with less than one sample per day. Each line represents the abundance of a specific mutation over time. Line colors were kept consistent with Figs. 2.5 and 2.6: blue lines are mutations shared across all these countries in the top 15, while the red lines are unique to that one country in the top 15 and the gray lines are non-unique and non-common variants where the mutation is observed in more than one country. Areas highlighted in yellow and blue indicate pre-lockdown and post-lockdown respectively. Major milestones in the national response against COVID-19 are annotated at the top. (B) Number of submitted genome sequences in the Netherlands, averaged over a period of 7 days and removed periods with less dan one sample.

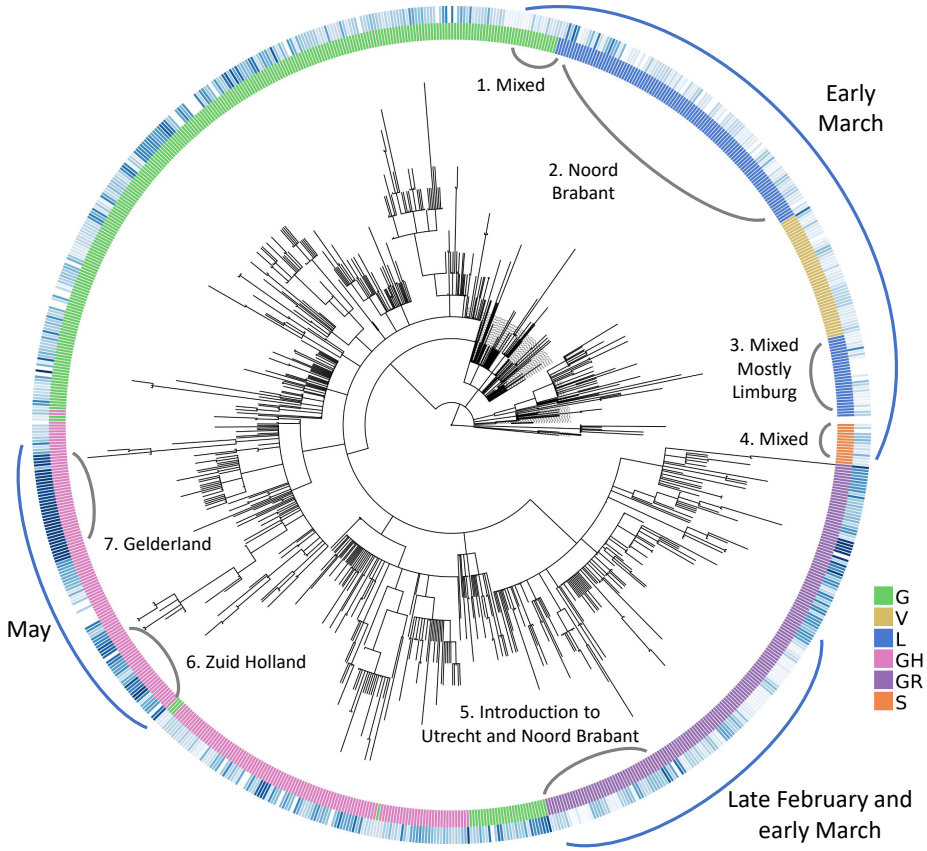


Figure 2.8: Radial representation of the Dutch phylogenetic tree: inner circle colored w.r.t. the assigned clades (see legend for clade names), outer circle is color-coded according to the sample collection date (if available), where the darker shade of blue represents more recent samples. Major points discussed in the text have been indicated with blue arches on the outer circle, along with more detailed information (numbered in clockwise direction) in gray arches on the inner circle.

February and early March" with a blue arch. This section differs from the first one in that we observe only samples of G and GR clades, both of which are dominant in the Europe while absent in China. The earliest SARS-CoV-2 genome in our dataset with full sample collection date (accession ID EPI_IS_454750, collected on February 27) is also located in this section and it was first isolated in Utrecht (also see Supplementary file 4, rectangular Dutch tree annotated with GISAID clade assignments, collection dates and within-Netherlands location).

Recall the clade distribution over time in the Netherlands (Figs. 2.2 and 2.3) showed an initial phase of high diversity with L and GR dominating the dataset, also supported by the phylogenetic analysis. As part of the Dutch initiative to investigate transmission of COVID-19 in the Netherlands, Munnink et al. had conducted a detailed analysis on the earlier samples with patient data [11]. More recently, Sikkema et al. have published their findings on COVID-19 infection in health-care workers in early March [54]. Their studies suggest multiple introductions from Italy and Switzerland, as well as localized community transmissions in super-spreading events in late February and early March. We also note early samples from the Netherlands scattered among samples from outside the Netherlands, mostly Europe, collected around the same time in global phylogenetic tree (Supplementary file 5). In addition, the authors note the diversity of early strains even for patients with similar travel histories, also in parallel with our observations in our study. In addition to Noord-Brabant, Munnink et al. had detected local clusters in Zuid-Holland and Utrecht.

2.3.5 Novel mutations appear in the later phase of pandemic

To explore local transmission clusters, we analyzed mutations that appeared after the initial pandemic response in the Netherlands. Munnink et al. have stated three phases of response to pandemic in the Netherlands in their study; (1) before the first case was reported, (2) from the first reported case to the start of screening of healthcare workers and (3) the period from the introduction of stricter measures along with events and large gatherings of people being banned until March 15th when the most strict phase of lockdown had begun as retail and catering industries were closed, as well as schools and childcare centers [41]. Since March 15th, the spread of COVID-19 has been very limited due to more stringent measures on travel and widely adopted practice of social distancing. For this reason, it is particularly interesting to investigate the deeper branching in Fig. 2.8 with later samples around 8–9 o'clock (denoted with a blue arch and the text "May").

Below in Fig. 2.9, we have zoomed into the two "May" regions from Fig. 2.8 (numbered 3 and 4 in Fig. 2.9) as well as the remaining deep branches (numbered 1 and 2 in Fig. 2.9). To simplify, we have indicated the absence/presence of a mutation with a circle where the branch ends. Additional information about sample collection date and its location are also displayed aligned to the leaves, if available and in the case of duplicate sequences separated with a semicolon. Dates are expressed in format month-day. The large squares next to the leaf names are color-coded clade assignments, colors have been kept consistent throughout our study in Figs. 2.2, 2.8 and 2.9.

We have identified four mutations all of which have emerged after March 15th and have led to deeper branching on the phylogenetic tree and are either unique to the Netherlands or very rarely observed in the rest of the world: N:P383P, NSP14:D390D, NSP14:S374A and ORF7a:F87F. These rare mutations could be further utilized to track local transmissions of

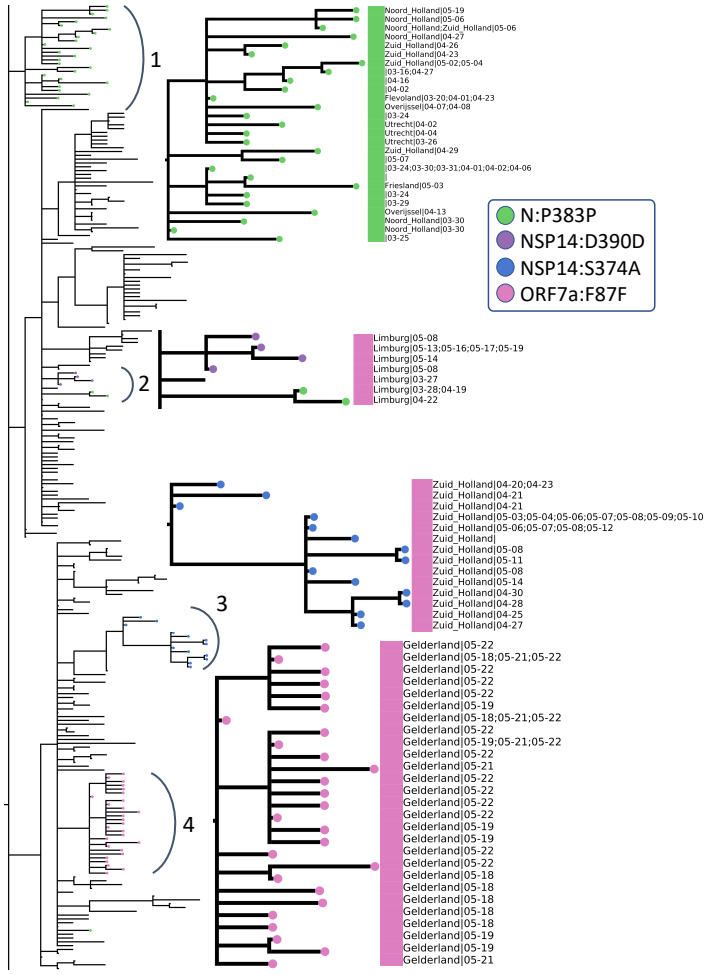


Figure 2.9: Zoomed-in view of rectangular representation of the Dutch phylogenetic tree: three regions of focus are numbered next to the corresponding arch. Newer, unique mutations that define deep branching in the tree are drawn in circles and the common mutations within Europe are rectangle (see legend for mutation annotations). Assigned clades are indicated with large rectangle aligned next to the leaves (pink: GH and green: G) and additional information about sequences (location and sequence collection date) are displayed next to the clade color, if available.

disease within the Netherlands.

N:P383P (green circles), a silent mutation on N protein is fairly unique to the Netherlands; it is present in less than five sequences in many European countries, including the most well-sampled ones Denmark and the UK, as well as the USA and Canada. Considering the sample size, it is surely intriguing that this mutation has been observed only in the Netherlands in such abundance. This mutation is also one of the oldest circulating ones since its first occurrence was in a sequence from the Switzerland on February 27th. However, we observe it for the first time in the Netherlands 2 weeks later on March 16th (province unknown). Later on, the same mutation has appeared in multiple provinces, Noord Holland, Zuid Holland, Flevoland, Utrecht and Limburg, in 50 sequences in total. Moreover, we observe it in two separate branching events in the phylogenetic tree in different provinces of the Netherlands; several provinces in arc number 1 and only in Limburg in arc number 3. In a recent study, this mutation had been detected as one of several homoplasies on the SARS-CoV-2 genome [55]. Since the Limburg branching contains only three sequences carrying the mutation, it is difficult to comment whether it is convergent or not. Given that branching with number 1 contains several provinces; it is also likely that this is a consequence of relaxations in domestic travel restrictions, rather than convergent evolution.

The second mutation, NSP14:D390D (purple circles), is tricky to interpret because it is present in only nine genomes, seven of which had been sequenced in the Netherlands and the remaining two in the UK. It has first appeared in the UK on March 24th, during strict lockdown conditions, and it has emerged in the Netherlands in May. We hypothesize this is a small cluster of variants genomes, localized in Limburg only and it has not found the chance to spread outside of the province yet.

NSP14:S374A (blue circles) is the only non-silent mutation in this list, and is very unique to Zuid Holland; it is present in 35 genomes in total, all collected in Zuid-Holland region within 3 weeks. Similar to NSP14:D390D, it is highly likely to be a small, contained cluster of individuals.

ORF7a:F87F (pink circles) is also incredibly rare since it was observed only in Gelderland in the Netherlands from late April to early May, and less than five times in any other country. It occurs in only one sequence from Canada in April 13th, twice in the USA in late March and four times in the UK in mid-April.

2.4 Discussion

In this work, we retrieved 29,503 complete, high quality SARS-CoV-2 from publicly available databases to explore the viral population diversity In the Netherlands, within a global context. Considering the rapid increase in public data and research on this subject, our work is among the more comprehensive ones to lend insight into the genetic variation of SARS-CoV-2 in the later stages of the pandemic in April and early May.

As a consequence of the natural evolution of a virus, SARS-CoV-2 genome has been diverging from the initial reference sequence Wuhan-Hu-1 established based on viral samples from Wuhan, China. The six major clades designated by GISAID had varying distributions in different regions, at different points of time through the course of pandemic. We demonstrated that in most countries, viral population goes through an initial phase of high diversity followed by a decline in genetic variety in which the population is

comprised of mostly G, GR or GH clades (Fig. 2.2). With increased ease of travel, COVID-19 was able to spread rapidly across the world and several studies had reported multiple introductions of a diverse viral population into many countries outside of China that lends itself to a more homogeneous population diverged from the Wuhan-Hu-1 reference [56–58]. Interesting, we have also observed that China and Singapore, both of which are countries that experienced the outbreak the earliest, harbor a markedly different viral population that remains mostly homogeneous with L being the dominant clade that also includes the Wuhan-Hu-1 reference (Fig. 2.2). Note that this could also be the artifact of the dramatic decline in number of sequences from China, where we do not have any sequence collected after April.

The S and N proteins in SARS-CoV-2 genome has received much attention; both have been reported as the most variant proteins [23, 34] and are also significant in RT-qPCR based diagnostic tests as well as vaccine and drug development [35]. We have identified the most variant sites on the S and N proteins in sequences from the Netherlands (Fig. 2.4). Koyama et al. had noted the effect of these variants on sequence-based vaccine and therapeutics against COVID-19 [39]. Following their discussion, we highlight their predicted epitope regions derived from SARS and the mutations we detected on the S and N proteins in Fig. 2.4. In addition, Kim et al. discussed variations on SARS-CoV-2 genes targeted by diagnostic assays in [21], and Vanaerschot et al. observed a mutation on N gene decrease the sensitivity of SARS-CoV-2 detection [59]. More recently, it was reported that a novel variant, first detected in the UK, denoted B.1.1.7, could lead to false negative results in diagnostic tests targeting the S gene [60].

Similarly, we analyzed primer/probe sequences currently in use in the Netherlands for diagnostics targeting S and N genes (Fig. 2.4); we found a mutation on N protein, RG203KR (in 57% of the genomes) overlapping with the target region of China CDC diagnostic test. While there were no major mutations on target regions of tests released by US CDC, WHO or Institut Pasteur in our dataset, emerging variants should be monitored routinely to ensure the reliability of diagnostics. In our studies, we find amino-acid positions from 40 to 90, 160–170, 330–360 and 400–420 on N protein could potentially be utilized as targets (blue shaded regions in Fig. 2.4). Even though RT-qPCR tests contain primer/probe sets targeting multiple genes, according to the recent WHO guidelines, a single target could be used as well, particularly in areas where COVID-19 has spread widely. Hence, it is recommended that primer/probe binding sites are investigated for mismatches [61].

When we observed the global landscape of variants, we found four mutations, S:D614G, N:RG203KR, NSP3:F106F and 5'UTR:241, are not only the most frequent ones, but also have been steadily increasing in the frequency outside of China since the beginning of pandemic. The 614G variant has been reported to exhibit increased transmissibility in human cells and animal models61, as well as phylodynamic studies [62], although there are currently no known effects on the disease trajectory or clinical outcome [63]. Volz et al. also report two mutations, S:D614G and N:RG203KR, to be linked [62]. Some studies have suggested certain linked mutations which poses a different question on its own [63]. We also reported the increase in frequency of these shared mutations, regardless of the date of introduction (Fig. 2.6). On one hand, the abundance of these mutations might suggest that viral genome has converged to a new variant, different than the Wuhan-Hu-1 reference. On the other hand, since most of the viral sequences are from diagnostic tests performed

on hospitalized patients at the moment, we are looking at only a small portion of the whole virus population in humans and we do not know clearly whether milder, or even asymptomatic cases of COVID-19 also carry these mutations or not. To our knowledge, studies have not found any significant correlation between these specific mutations and the COVID-19 disease in patients [63]. Nevertheless, it is surely interesting consider that these four mutations, linked to one another, might also influence the infection in the human host.

With our phylogenetic study in the Netherlands, we confirmed multiple introductions in distinct provinces as well as the population diversity in the initial samples. We found sequences collected from late February to early March in Noord Brabant, Limburg, Utrecht as well as Zuid Holland spread around the tree indicating genetically very diverse strains (Fig. 2.8). We also detected emerging local clusters, defined by four mutations, N:P383P, NSP14:D390D, NSP14:S374A and ORF7a:F87F, all of which are either entirely unique to the Netherlands or very rarely observed elsewhere (Fig. 2.9). N:P383P had occurred at two distinct sections in different regions, we presume this is likely a domestic travel event rather than a convergent mutation. We note the detection and monitoring of such unique mutations could be utilized for tracking the spread of virus and identifying possible routes of transmission during the outbreak. In addition, our findings are in line with previous studies in the Netherlands by Munnink et al. and Sikkema et al.; they had also observed sequence diversity in the earliest days of the outbreak as well as community transmission [11, 54].

The single most prominent pattern that we encountered in our study was that despite the continual increase in number of mutations in the genome, diverging away from the Wuhan-Hu-1 reference, there is little diversity in the new variants as we enter the later stages of the first wave of the pandemic. This suggests the current SARS-CoV-2 reference genome should be re-evaluated, perhaps replaced with a new one that represents the viral population more accurately. Further work is required to investigate implications of an inadequate reference in sequence-based analyses as well as develop alternative models. Having a good quality reference sequence is crucial in sequence-based analyses; we expect better read mapping and variant calling would improve phylogenetic studies and clade designations, and allow for reliable detection of transmission clusters and emerging variants. Improved variant detection would enable design of more accurate diagnostic assays. We assert this line of research will continue to supplement the global effort to fight COVID-19.

The major limitation of our study is the biased sampling of SARS-CoV-2 sequences. Despite our efforts to combine all genome sequences publicly available up to date, due to imbalanced sampling and dramatic changes in the frequency of genome sequencing, our dataset is over-represented by samples from the Europe and the USA and there are several gaps in time since the beginning of pandemic. In addition, most of the viral sequencing today is performed on hospitalized patients. These issues could be circumvented to some extend by stratified sampling or controlled sequencing efforts with random samples collected from individuals. Nevertheless, our findings are significant to understand the SARS-CoV-2 genome and both its national and global population diversity.

2.5 Conclusions

In this study, we have analyzed 29,503 SARS-CoV-2 genomes retrieved from public databases to investigate genetic diversity in viral population as the pandemic progresses, with a focus on the Netherlands in particular. Our dataset contained 1338 genomes from the Netherlands, most of them sequenced in April and early May. We assert our work provides valuable information on the genetic diversity of SARS-CoV-2 and its local dynamics in the Netherlands for tracking the transmission of COVID-19, as well as localized, region-specific efforts in DNA-based therapeutic or vaccine development against COVID-19, and primer/probe design in RT-qPCR tests. Our work demonstrates the use of genomics in guiding diagnostics and outbreak investigation at a limited scale. In order to fully realize the potential of genomic epidemiology, we need routine sequencing of viral DNA established in parallel with COVID-19 testing. We emphasize the little diversity observed globally in recent samples despite the increased number of mutations relative to the established reference sequence, suggesting the current reference may not be representative of the population; potential implications of an inadequate reference on downstream analyses should be investigated.

2.6 Supplementary material

2.6.1 Supplementary text: Annotation of mutations further elucidate conserved regions and show a general preference for non-silent changes in the genome

We have characterized and annotated point mutations in the SARS-CoV-2 genomes sampled within the Netherlands. A large portion of these mutations are found in the S and N proteins (subplots A and B in Figure S2.1, number of unique mutations and the total number of mutations at the top of each bar); overall NSP3, NSP12b, S and N proteins carry a majority of the mutations (protein names in orange color in Figure S2.1). To classify the mutations, we follow a similar nomenclature to coronapp's: SNPs leading to a change in the amino acid sequence are non-silent, SNPs with no amino acid change are silent and SNP stop denotes SNPs where a stop codon is introduced.

Most variant sites, in terms of total number of unique mutations, also favor non-silent SNPs. More than half of unique mutations in NSP1, NSP5, NSP6, NSP7, NSP8 and NSP9 are non-silent (orange parts of bars in Figure S2.1A). According to a recent study analyzing SARS-CoV-2 proteins in terms of their codon usage, S, N, NSP7 and NSP8 proteins might play a critical role in adaptation to their host since they prefer a smaller, more "optimized" set of codons compared to the rest of the proteins [64].

Since there are only few dominant mutations in each protein, the percent breakdown becomes more skewed when the total number of mutations is considered in Figure S2.1B. More than 80% of total nucleotide substitutions in NSP3 are silent, whereas substitutions do cause a change in the amino acid sequence in more than 60% of the instances on the other proteins (compare blue and orange bars in Figure S2.1B). However, we have not observed a significant change in the relative frequency of silent and non-silent SNPs over time (Figure S2.2).

In terms of total number of mutations, NSP7, NSP10, ORF6, ORF7b and ORF10 appear to be the least variable proteins in the SARS-CoV-2 genome (protein names in blue color

Table S2.1: Percent breakdown of unique and total mutations observed in the Netherlands on different proteins of the SARS-CoV-2 genome, total number of unique and total mutations in each protein are also reported. Throughout this report, the term "nonsilent" refers to nucleotide changes accompanied with an amino acid change, whereas "silent" mutations are nucleotide changes with no change in the amino acid sequence.

	Unique mutations					Total mutations				
Protein	SNP nonsilent	SNP silent	SNP stop	Deletion	Total #	SNP nonsilent	SNP silent	SNP stop	Deletion	Total #
NSP1	29.2	58.3	0	12.5	24	6.4	90.3	0	3.4	236
NSP2	62	36.6	0	1.4	71	34.5	44.3	0	21.2	707
NSP3	67.9	30.9	0.6	0.6	165	19	80.9	0.1	0.1	1542
NSP4	56.8	40.9	0	2.3	44	61.9	37.5	0	0.6	160
NSP5	50	50	0	0	32	54.9	45.1	0	0	113
NSP6	54.2	45.8	0	0	24	90.3	9.7	0	0	424
NSP7	37.5	62.5	0	0	8	27.3	72.7	0	0	11
NSP8	47.6	47.6	0	4.8	21	12.7	85.8	0	1.5	134
NSP9	33.3	66.7	0	0	12	37.6	62.4	0	0	85
NSP10	66.7	33.3	0	0	9	76.9	23.1	0	0	13
NSP12a	100	0	0	0	1	100	0	0	0	73
NSP12b	58.9	41.1	0	0	56	87.3	12.7	0	0	1312
NSP13	62.5	37.5	0	0	64	67.5	32.5	0	0	326
NSP14	58	42	0	0	50	47.2	52.8	0	0	218
NSP15	65.8	31.6	0	2.6	38	37.6	62	0	0.4	271
NSP16	42.9	57.1	0	0	21	85	15	0	0	107
S	61.4	35.7	0.7	2.1	140	85.7	13.9	0.1	0.3	1471
ORF3a	70.4	24.1	3.7	1.9	54	94.1	5	0.8	0.2	623
E	66.7	33.3	0	0	9	72.3	27.7	0	0	47
M	35.7	64.3	0	0	28	85.2	14.8	0	0	244
ORF6	60	30	10	0	10	41.2	52.9	5.9	0	17
ORF7a	63.6	18.2	13.6	4.5	22	19.6	75.7	3.7	0.9	107
ORF7b	57.1	28.6	14.3	0	7	50	41.7	8.3	0	12
ORF8	64.7	23.5	5.9	5.9	17	75.5	20.4	2	2	49
N	69.5	30.5	0	0	82	82	18	0	0	523
ORF10	57.1	28.6	14.3	0	7	33.3	61.1	5.6	0	18

in Figure S2.1). ORF10, in particular is not only conserved, but it is also rather unique; there are currently no homologs of ORF10 on NCBI [65]. While studies have shown it is possible for ORF10 to be encoded in pangolin and bat viruses [66], ORF10 is mostly unique to SARS-CoV-2 and could be used as a specific marker.

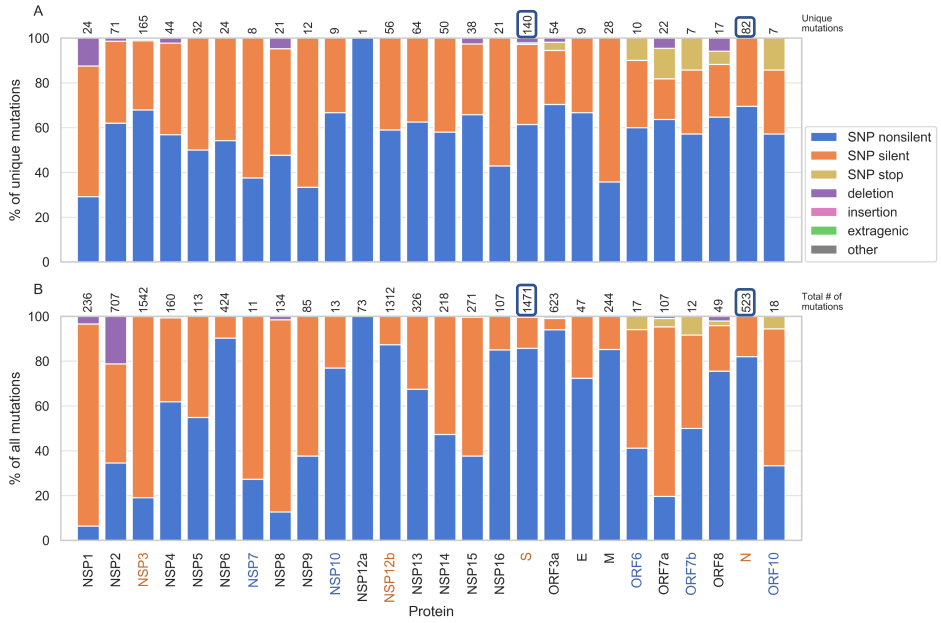


Figure S2.1: Percent breakdown of unique (A) and total (B) mutations observed in the Netherlands on different proteins of the SARS-CoV-2 genome, total number of unique and total mutations in each protein are placed at the top of the bars. Throughout this report, the term “nonsilent” refers to nucleotide changes accompanied with an amino acid change, whereas “silent” mutations are nucleotide changes with no change in the amino acid sequence.

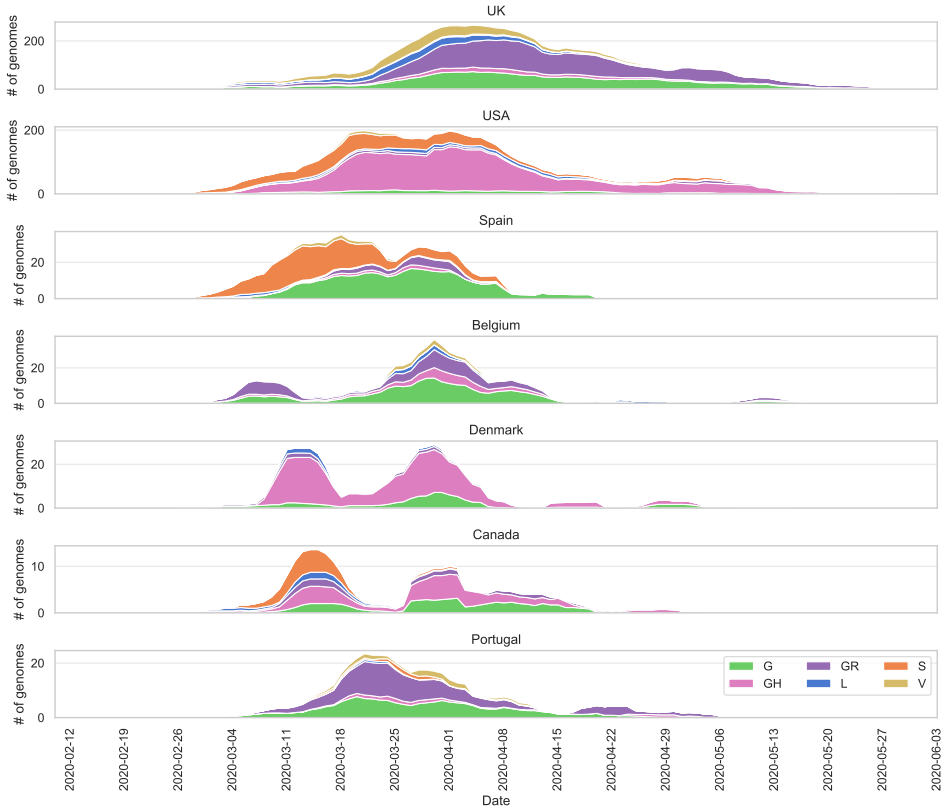


Figure S2.2: Changes in the percent breakdown of non-silent and silent SNPs observed on SARS-CoV-2 genome in the Netherlands over the course of pandemic.

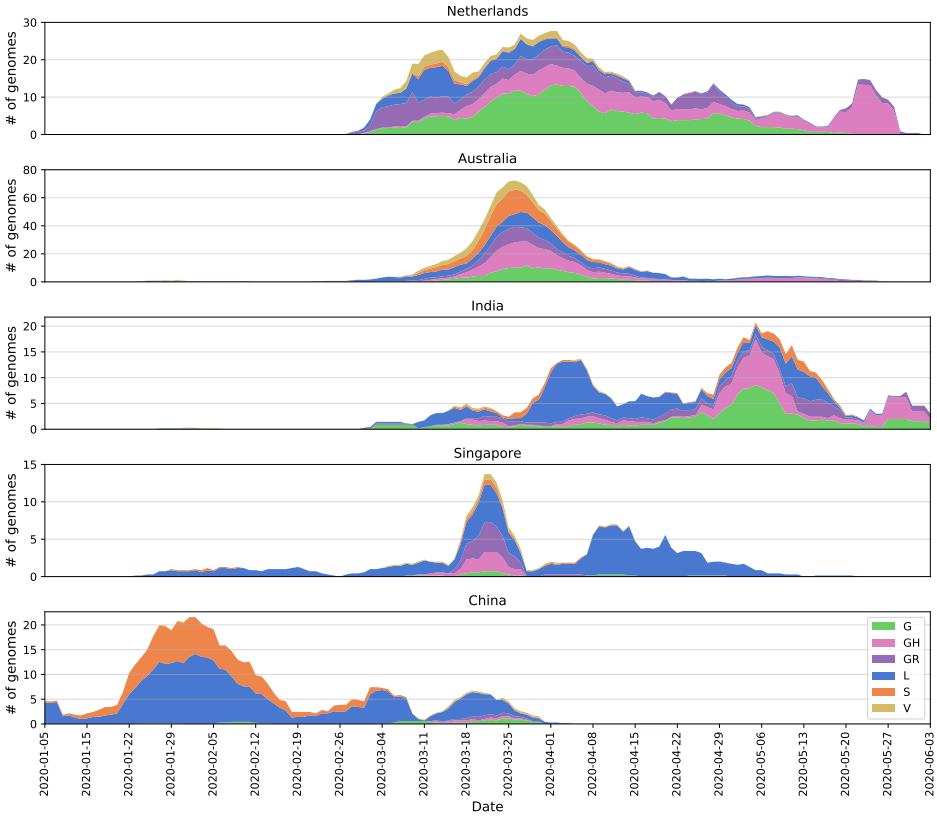


Figure S2.3: Distribution of SARS-CoV-2 clades in a selection among the 12 most sampled countries (Australia, India, Singapore, China) in comparison to the Netherlands: y-axis shows absolute number of genomes, and x-axis shows collection date. Moving average over seven days was calculated for six clades (see the legend for clade names and colors) discarding intervals of fewer than one genome per day.

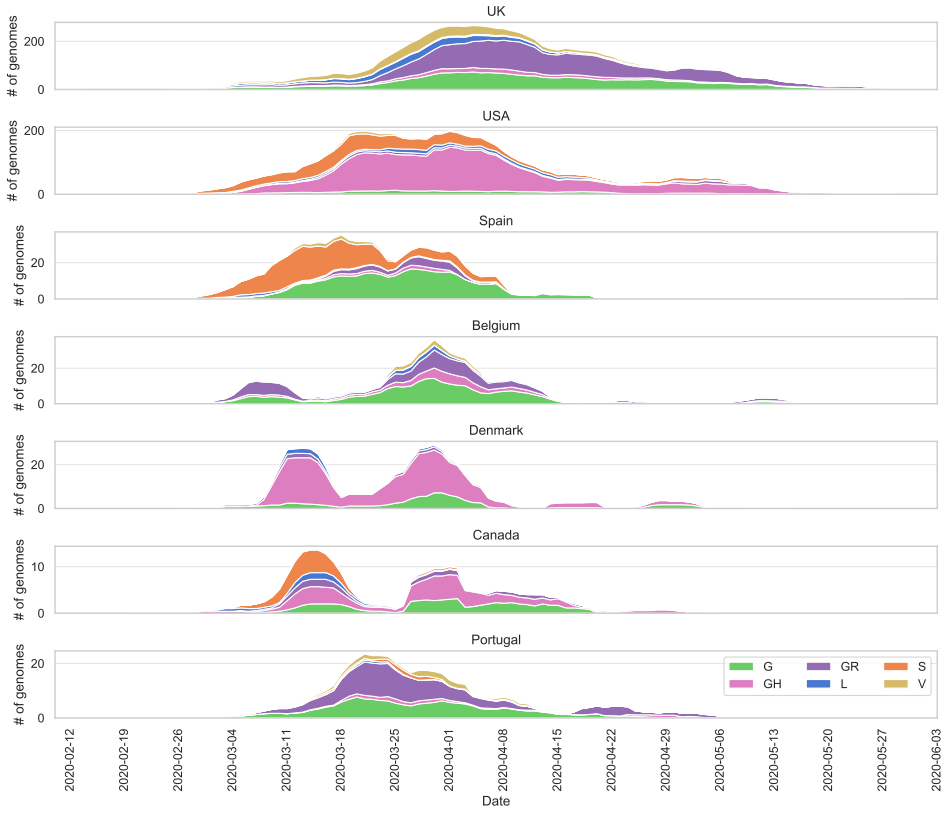


Figure S2.4: Distribution of SARS-CoV-2 clades in the UK, the USA, Spain, Belgium, Denmark, Canada and Portugal: moving average over seven days was calculated for six clades (see the legend for clade names and colors) discarding intervals of less than one genome per day.

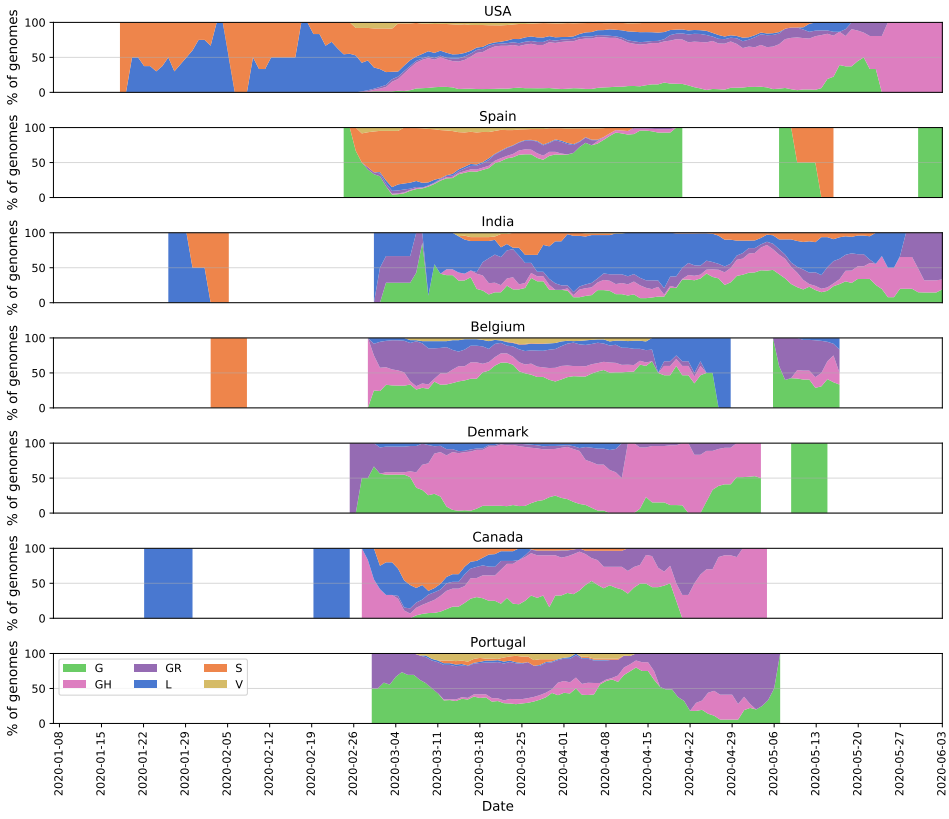


Figure S2.5: Distribution of SARS-CoV-2 clades in the UK, the USA, Spain, Belgium, Denmark, Canada and Portugal: y-axis shows % breakdown, and x-axis shows collection date. Moving average over seven days was calculated for six clades (see the legend for clade names and colors) discarding intervals of less than one genome per day.

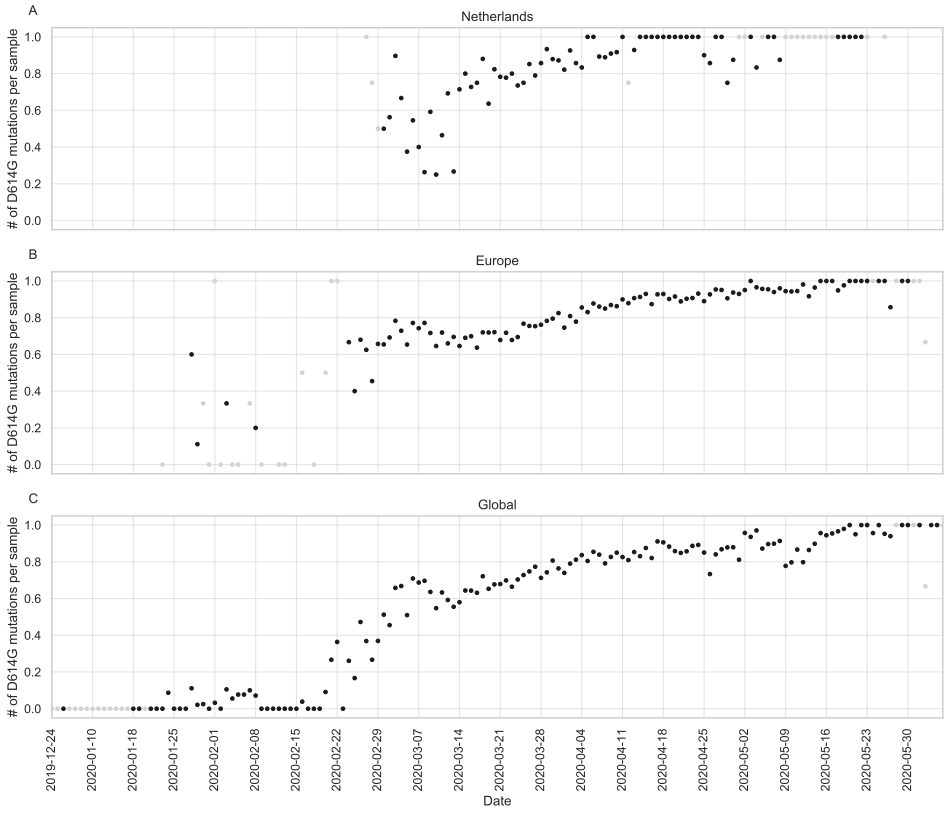


Figure S2.6: Number of S:D614G mutations observed per sample over the course of pandemic in the Netherlands (A), Europe (B) and globally (C): data points corresponding to dates with fewer than five samples are colored gray to indicate uncertainty.

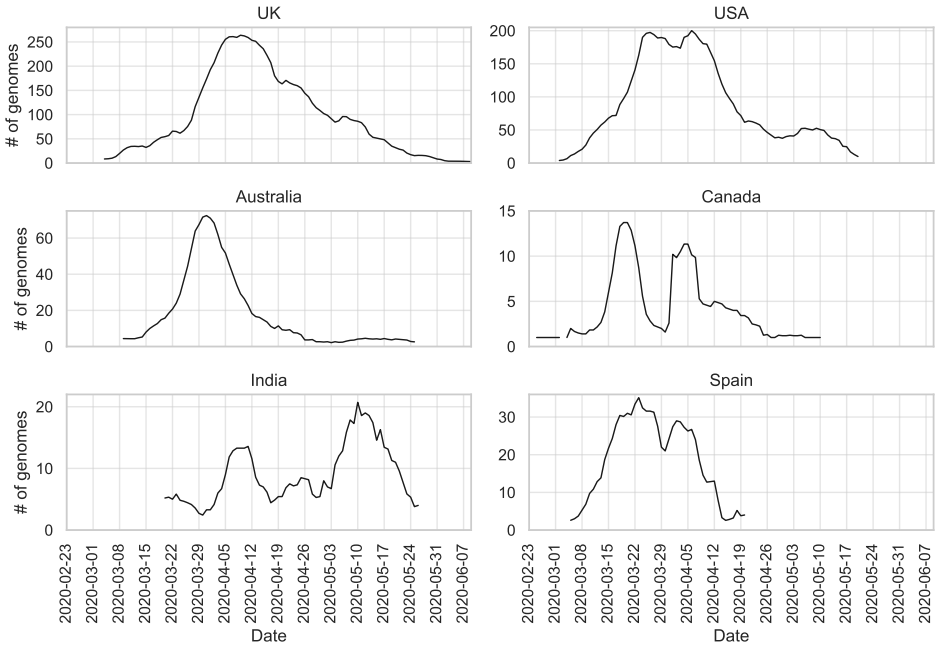


Figure S2.7: Change in number of genomes in the most-sampled countries in our dataset, numbers are averaged over a period of 7 days and periods with fewer than one sample are removed.

References

[1] Aysun Urhan and Thomas Abeel. Emergence of novel sars-cov-2 variants in the netherlands. *Scientific reports*, 11(1):6625, 2021.

[2] J. Cohen and D. Normile. New sars-like virus in china triggers alarm. *Science*, 367:234–5, 2020.

[3] Health organization. who coronavirus disease (covid-19) dashboard, 2020. Accessed 9 Sep 2020.

[4] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, and J. Song. A novel coronavirus from patients with pneumonia in china, 2019. *N Engl J Med*, 382:727–33, 2020.

[5] K.G. Andersen, A. Rambaut, W.I. Lipkin, E.C. Holmes, and R.F. Garry. The proximal origin of sars-cov-2. *Nat Med*, 89:44–8, 2020.

[6] F. Wu, S. Zhao, B. Yu, Y.M. Chen, W. Wang, and Z.G. Song. A new coronavirus associated with human respiratory disease in china. *Nature*, 579:265–9, 2020.

[7] N.D. Grubaugh, J.T. Ladner, P. Lemey, O.G. Pybus, A. Rambaut, and E.C. Holmes. Tracking virus outbreaks in the twenty-first century. *Nature Microbiology*, 4:10–9, 2019.

[8] T. Seemann, C. Lane, N. Sherry, S. Duchene, Silva A.G. da, and L. Caly. Tracking the covid-19 pandemic in australia using genomics, 2020.

[9] Z. Song, X. Zhou, Y. Cai, S. Feng, T. Zhang, and Y. Wang. Infection groups differential (igd) score reveals infection ability difference between sars-cov-2 and other coronaviruses, 2020.

[10] C. Wang, Z. Liu, Z. Chen, X. Huang, M. Xu, and T. He. The establishment of reference sequence for sars-cov-2 and variation analysis. *J Med Virol*, 2020.

[11] Oude Munnink BB, Nieuwenhuijse DF, Stein M, O.'Toole Á, Haverkate M, and Mollers M. Rapid sars-cov-2 whole-genome sequencing and analysis for informed public health decision-making in the netherlands. *Nat Med*, 2020.

[12] D.J. Baker, G.L. Kay, A. Aydin, T. Le-Viet, S. Rudder, and A.P. Tedim. Coronahit: large scale multiplexing of sars-cov-2 genomes using nanopore sequencing. *bioRxiv*, 2020.

[13] A.S. Gonzalez-Reiche, M.M. Hernandez, M.J. Sullivan, B. Ciferri, H. Alshammary, and A. Obla. Introductions and early spread of sars-cov-2 in the new york city area, 2020.

[14] Thanh Le T, Andreadakis Z, Kumar A, Gómez Román R, Tollefsen S, and Saville M. The covid-19 vaccine development landscape. *Nature reviews. Drug discovery*, 19:305–6, 2020.

[15] S. Cleemput, W. Dumon, V. Fonseca, W.A. Karim, M. Giovanetti, and L.C. Alcantara. Genome detective coronavirus typing tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics*, 2020.

[16] Y. Tang, T.D.A. Serdan, L.N. Masi, S. Tang, R. Gorjao, and S.M. Hirabara. Epidemiology of covid-19 in brazil: using a mathematical model to estimate the outbreak peak and temporal evolution. *Emerg Microbes Infect*, 9:1453–6, 2020.

[17] Y. Shu and J. McCauley. Gisaid: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*, 22(30494), 2017.

[18] A. Maitra, M.C. Sarkar, H. Raheja, N.K. Biswas, S. Chakraborti, and A.K. Singh. Mutations in sars-cov-2 viral rna identified in eastern india: Possible implications for the ongoing outbreak in india and impact on viral structure and host susceptibility. *J Biosci*, 45:1–18, 2020.

[19] I. Jungreis, R. Sealfon, and M. Kellis. Sarbecovirus comparative genomics elucidates gene content of sars-cov-2 and functional impact of covid-19 pandemic mutations, 2020.

[20] S. Laha, J. Chakraborty, S. Das, S.K. Manna, S. Biswas, and R. Chatterjee. Characterizations of sars-cov-2 mutational profile, spike protein stability and viral transmission. *Infect Genet Evol*, 85(104445), 2020.

[21] J.-S. Kim, J.-H. Jang, J.-M. Kim, Y.-S. Chung, C.-K. Yoo, and M.-G. Han. Genome-wide identification and characterization of point mutations in the sars-cov-2 genome. *Osong Public Heal Res Perspect*, 11:101–11, 2020.

[22] H. Harvala, D. Frampton, P. Grant, J. Raffle, R.B. Ferns, and Z. Kozlakidis. Emergence of a novel subclade of influenza a(h3n2) virus in london. *Eurosurveillance*, 22(30466), 2016.

[23] D. Benvenuto, M. Giovanetti, A. Ciccozzi, S. Spoto, S. Angeletti, and M. Ciccozzi. The 2019-new coronavirus epidemic: Evidence for virus evolution. *J Med Virol*, 92:455–9, 2020.

[24] National center for biotechnology information (ncbi)[internet, 1988. Accessed 1 Jan 2020.

[25] W.M. Zhao, S.H. Song, M.L. Chen, D. Zou, L.N. Ma, and Y.K. Ma. The 2019 novel coronavirus resource. *Yi Chuan*, 42:212–21, 2020.

[26] K. Katoh. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res*, 30:3059–66, 2002.

[27] D. Mercatelli, L. Triboli, E. Fornasari, F. Ray, and F.M. Giorgi. coronapp: a web application to annotate and monitor sars-cov-2 mutations, 2020.

[28] Centers for disease control and prevention. a cdc 2019-novel coronavirus (2019-ncov) real-time rt-pcr diagnostic panel. Accessed 13 Jan 2021.

[29] World Health Organization. Molecular assays to diagnose covid-19, 2020. Accessed 23 Jun 2020.

[30] Institut pasteur paris. protocol: Real-time rt-pcr assays for the detection of sars-cov-2. Accessed 13 Jan 2021.

[31] C.D.C. China. China cdc primers and probes for detection 2019-ncov. Accessed 13 Jan 2021.

[32] B.Q. Minh, H. Schmidt, O. Chernomor, D. Schrempf, M. Woodhams, and A. Haeseler. Iq-tree 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*, 37:1530–4, 2019.

[33] J. Huerta-Cepas, F. Serra, and P. Bork. Ete 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*, 33:1635–8, 2016.

[34] P. Zhou, Yang X. Lou, X.G. Wang, B. Hu, L. Zhang, and W. Zhang. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579:270–3, 2020.

[35] John hopkins center for health security. comparison of national rt-pcr primers , probes , and protocols for sars-cov-2 diagnostics. centerforhealthsecurity.org, 2020. Accessed 24 Jun 2020.

[36] L. Du, Y. He, Y. Zhou, S. Liu, B.J. Zheng, and S. Jiang. The spike protein of sars-cov - a target for vaccine and therapeutic development. *Nature Reviews Microbiology*, 7:226–36, 2009.

[37] J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, and S. Fan. Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor. *Nature*, 581:215–20, 2020.

[38] A.C. Walls, Y.J. Park, M.A. Tortorici, A. Wall, A.T. McGuire, and Veesler D. Structure. Function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell*, 181:281–292, 2020.

[39] T. Koyama, D. Weeraratne, J.L. Snowdon, and L. Parida. Emergence of drift variants that may affect covid-19 vaccine development and antibody treatment. *Pathogens*, 9(324), 2020.

[40] X. Tang, C. Wu, X. Li, Y. Song, X. Yao, and X. Wu. On the origin and continuing evolution of sars-cov-2.

[41] N.O.S. Alle scholen, cafés en restaurants tot en met 6 april dicht om coronavirus. *NOS.nl*, 2020. Accessed 26 Jun 2020.

[42] Coronavirus. Pm says everyone should avoid office, pubs and travelling - bbc news. *BBC News Services*, 2020. Accessed 18 Jan 2021.

[43] Australia closes borders to stop coronavirus | 7news.com.au. 7news, 2020. Accessed 18 Jan 2021.

[44] Travel health notices. *Government of Canada*, 2020. Accessed 18 Jan 2021.

[45] Coronavirus: Sánchez decreta el estado de alarma durante 15 días | españa | el paÍs. *El Pais*, 2020. Accessed 18 Jan 2021.

[46] Fact Sheet. Dhs notice of arrival restrictions on china, iran and certain countries of europe | homeland security. *Homeland Security*, 2020. Accessed 18 Jan 2021.

[47] Government of india ministry of home affairs. *ORDER*, (40-3/2020-DM-I(A)), 2020. Accessed 18 Jan 2021.

[48] Zo ziet de versoepeling van de coronamaatregelen er in de komende maanden uit | nos. *NOS*, 2020. Accessed 18 Jan 2021.

[49] Trump gives governors 3-phase plan to reopen economy. *APNews*, 2020. Accessed 18 Jan 2021.

[50] Nsw pubs and clubs to reopen on friday for dining after coronavirus shutdown - abc news. *ABC News*, 2020. Accessed 18 Jan 2021.

[51] Spanish government does u-turn, will allow children aged 14 and under out for walks | society | el paÍs in english. *El Pais*, 2020. Accessed 18 Jan 2021.

[52] Our plan to rebuild: The uk government's covid-19 recovery strategy - gov.uk. *Cabinet Office*, 2020. Accessed 18 Jan 2021.

[53] Epidemiologische situatie covid-19 in nederland 22 mei 2020 | rivm. *RIVM*, 2020. Accessed 18 Jan 2021.

[54] R.S. Sikkema, S.D. Pas, D.F. Nieuwenhuijse, Á. O'Toole, J. Verweij, and A. Linden. Covid-19 in health-care workers in three hospitals in the south of the netherlands: a cross-sectional study. *Lancet Infect Dis*, 2020.

[55] Dorp L, Acman M, Richard D, Shaw LP, Ford CE, and Ormond L. Emergence of genomic diversity and recurrent mutations in sars-cov-2. *Infect Genet Evol*, 83(104351), 2020.

[56] S. Dellicour, K. Durkin, S.L. Hong, B. Vanmechelen, J. Martí-Carreras, and M.S. Gill. A phylodynamic workflow to rapidly gain insights into the dispersal history and dynamics of sars-cov-2 lineages, 2020.

[57] Petrone Fauver, JR, Hodcroft ME, Shioda EB, Ehrlich K, Watts HY, and A.G. Coast-to-coast spread of sars-cov-2 during the early epidemic in the united states. *Cell*, 181:990–6, 2020.

[58] J. Tian, G. Zhang, Y. Ju, N. Tang, J. Li, and R. Jia. Five novel carbapenem-hydrolysing oxa-type β-lactamase groups are intrinsic in acinetobacter spp. *J Antimicrob Chemother*, 73:3279–84, 2018.

[59] M. Vanaerschot, S.A. Mann, J.T. Webber, J. Kamm, S.M. Bell, and J. Bell. A sars-cov-2 variant that occurs worldwide and has spread in. *bioRxiv*, 2020.

[60] E. Mahase. Covid-19: What have we learnt about the new variant in the uk? *BMJ*, 371:m4944, 2020.

[61] Genomic sequencing of sars-cov-2: a guide to implementation for maximum impact on public health. Accessed 19 Jan 2021.

[62] E. Volz, V. Hill, J.T. McCrone, A. Price, D. Jorgensen, and Á. O'Toole. Evaluating the effects of sars-cov-2 spike mutation d614g on transmissibility and pathogenicity. *Cell*, 184:64–75, 2020.

[63] B. Korber, W. Fischer, S.G. Gnanakaran, H. Yoon, J. Theiler, and W. Abfalterer. Spike mutation pipeline reveals the emergence of a more transmissible form of sars-cov-2. *bioRxiv*, 4(2020), 2020.

[64] Maddalena Dilucca, Sergio Forcelloni, Alexandros G Georgakilas, Andrea Giansanti, and Athanasia Pavlopoulou. Temporal evolution and adaptation of sars-cov 2 codon usage. *bioRxiv*, pages 2020–05, 2020.

[65] T Koyama, D Platt, and L Parida. Variant analysis of covid-19 genomes.[preprint]. *Bull World Heal Organ*, 98(7), 2020.

[66] Christian Jean Michel, Claudine Mayer, Olivier Poch, and Julie Dawn Thompson. Characterization of accessory genes in coronavirus genomes. *Virology journal*, 17(1):1–13, 2020.

3

A comparative study of pan-genome methods for microbial organisms: Acinetobacter baumannii pan-genome reveals structural variation in antimicrobial resistance-carrying plasmids

"Now, if you'll excuse me, I'm going to go home and have a heart attack."

— Quentin Tarantino

This chapter is based on Aysun Urhan and Thomas Abeel. A comparative study of pan-genome methods for microbial organisms: *Acinetobacter baumannii* pan-genome reveals structural variation in antimicrobial resistance-carrying plasmids. *Microbial Genomics*, 7(11), 2021. [1].

Abstract

Microbial organisms have diverse populations, where using a single linear reference sequence in comparative studies introduces reference-bias in downstream analyses, and leads to a failure to account for variability in the population. Recently, pan-genome graphs have emerged as an alternative to the traditional linear reference with many successful applications and a rapid increase in the number of methods available in the literature. Despite this enthusiasm, there has been no attempt at exploring these graph construction methods in depth, demonstrating their practical use. In this study, we aim to develop a general guide to help researchers who may want to incorporate pan-genomes in their analyses of microbial organisms. We evaluated the state-of-the art pan-genome construction tools to model a collection of 70 *Acinetobacter baumannii* strains. Our results suggest that all tools produced pan-genome graphs conforming to our expectations based on previous literature, and that their approach to homologue detection is likely to be the most influential in determining the final size and complexity of the pan-genome. The graphs overlapped most in the core pan-genome content while the cloud genes varied significantly among tools. We propose an alternative approach for pan-genome construction by combining two of the tools, Panaroo and Ptolemy, to further exploit them in downstream analyses, and demonstrate the effectiveness of our pipeline for structural variant calling in beta-lactam resistance genes in the same set of *A. baumannii* isolates, identifying various transposon structures for carbapenem resistance in chromosome, as well as plasmids. We identify a novel plasmid structure in two multidrug-resistant clinical isolates that had previously been studied, and which could be important for their resistance phenotypes.

Data Summary

A dataset of 70 *Acinetobacter baumannii* strains has been curated from a published dataset used in a comparative study of adaptation in niche environments by removing the oldest assemblies of low quality [2]. This particular dataset was selected as the use-case for evaluating pan-genomes because (i) it comprises only full-length genome assemblies, (ii) it includes strains isolated from different environments and thus is diverse, and (iii) the original study provides a common ground on which a baseline evaluation can be performed to compare the results of different tools. Sequence and annotation data have been obtained from the NCBI RefSeq database [3]; the accession numbers of assemblies used in this work are listed in Table S3.1 (available in the online version of this paper). The scripts and code developed for this work can be found in the github repository at https://github.com/AbeelLab/abaumannii-pangenome.

3.1 Introduction

As the amount of DNA sequence data available has increased dramatically, the conventional, reference-based approach in bioinformatics is being re-examined. Relying on a single linear reference sequence in comparative genomic studies can lead to reference-bias in downstream analyses, and to failure to account for population variance which may be valuable [4].

Pan-genome graphs have been proposed as an alternative to a linear reference to model a collection of DNA sequences [5], representing genes shared across multiple genomes

in a compact structure, and hence, they have found many applications in many tasks such as genome alignment, read mapping and variant calling [6–9]. Several methods have been developed in the literature to construct gene-based pan-genome graphs, and we can summarize these methods in roughly three steps: (i) identifying homologue genes based on all-vs-all pairwise alignments [10–12] and clustering [13, 14], (ii) paralogue splitting and (iii) linking families to preserve the genomic order. For paralogue splitting, two different approaches stand out: tree-based ones that make use of the phylogeny in gene families and synteny-based ones in which the neighbourhood of each gene family guides the paralogue splitting process. The final step may vary depending on the output of the tool and, in some algorithms, it may be absent unless a final graph is produced.

Currently, there has been no attempt at bringing these graph construction methods to a common ground, assessing both their weaknesses and strengths independent of their computational performance. In this study, we evaluate the state-of-the-art pan-genome construction tools to propose general guidelines and rules-of-thumb to help with researchers who may want to incorporate pan-genomes in their analyses, particularly in those of microbial organisms. The aim is to explore what questions each tool might be useful in answering, and in what ways we can make use of these answers to gain valuable biological insight. We performed a comparative study on a collection of 70 *A. baumannii* strains of different isolation sources that has previously been published [2]. *A. baumannii*, a multi-drug-resistant bacteria classified as an ESKAPE pathogen, is among the leading causes of nosocomial infections, and thus plays a vital role in understanding antibiotic resistance [15]. Studies have established genes associated with several traits including virulence, pathogenicity and adaptation to its niche, probably acquired through horizontal transfer in large clusters via plasmids [16]. *A. baumannii*, as a population, has a diverse gene repertoire, and exhibits large, structural rearrangements; hence it has the prominent characteristics of bacterial genomes and presents as a good example use-case for application of pan-genome graphs in bacterial species. Given its typical average genome size and plasmid content for bacteria, it should not pose any additional challenges to the algorithms which would interfere with the comparison. In this work, first, we verify that our results confirm the original analyses, and are in parallel with previous studies on *A. baumannii*. Next, we propose to combine two of the pan-genome construction tools we have evaluated, Panaroo and Ptolemy, to further exploit them in downstream analyses; the effectiveness of this approach is demonstrated by calling structural variants in *A. baumannii* species to gain more insights in the data set. We analysed different structures of transposons carrying the $bla_{\text{OXA-23}}$ carbapenemase gene in the set of *A. baumannii* strains. In addition, we explore *A. baumannii* plasmids, and locate novel structures that might be involved in transferring multiple antimicrobial resistance genes.

3.2 Methods

In this section, we first describe our approach for comparing state-of-the-art pan-genome tools. The aim in the first part of this study is to evaluate existing tools in both qualitative and quantitative terms, and to provide an overview of the current field. In the second part, two of these tools, Panaroo and Ptolemy, are used in conjunction for calling structural variants. The final pan-genome graph serves as a compact model of a set of genomes, utilizing Panaroo's error correction mechanisms while it also retains the sequence continuity in

each genome with the guidance of Ptolemy's indexing and anchoring.

3.2.1 Data preprocessing

The *A. baumannii* dataset was curated from a previous study by Yakkala et al. which analysed niche-specific adaptations of *A. baumannii* [2]. We removed the oldest, low-quality assemblies to retain 70 in total. Nucleotide sequences and annotations were downloaded from the NCBI RefSeq database (assembly accessions are listed in Table S3.1) [3]. The genomes were not re-annotated in our study since all the assemblies had been annotated by NCBI's prokaryotic annotation pipeline, and thus they had gone through the same process. NCBI annotations were corrected and the corresponding nucleotide sequences were appended to the GFF input files using the python script convert_refseq_to_prokka_gff.py provided by Panaroo (see Supplementary Text) [17].

3.2.2 Tools

In this study, we have compared the tools Roary (v3.13.0), Ptolemy (v1.0), PPanGGOLiN (v1.0.13), PIRATE (v1.0.3) and Panaroo (v1.1.2) [17–21]. The set of tools are by no means comprehensive; however, they are diverse enough in their methodology and at the same time sufficiently similar in their purpose to make comparison meaningful. In addition, the pan-genome representation is consistent across these tools; the pan-genome is a graph in which the nodes are formed by at least one gene (a node may contain multiple genes forming an orthologous cluster) and the edges indicate sequence continuity between two nodes. All the tools were run in their default settings according to instructions provided by their authors. Tools which allow for some options without the need for parameter tuning were also run with these options. A full list of commands and arguments used in this study can be found in the Supplementary Text.

3.2.3 Qualitative and quantitative assessment

In the first part, we compared different tools qualitatively in their usage first in terms of software availability input/output file formats and compatibility with existing downstream analyses. Input is usually sequences with their annotations in FASTA and GFF files, or GenBank and GFF3 files that contain both the nucleotide sequences and annotations in a single file. If a tool provides sequence annotation as well, then FASTA sequences alone can be used. Since these tools are often run within a pipeline, once a pan-genome is obtained, it might be used for aligning reads and whole genomes, calling structural variants or performing genome-wide association studies (GWAS). To establish compatibility, tools produce outputs in commonly used file formats such as DOT, GML, GEXF or GFA for graphs, NEWICK for phylogenic trees and tab-separated text files for the remaining types of outputs. In addition to these, we attempt to compare the core algorithms of the tools by breaking down pan-genome construction into multiple steps in (i) detection of homologue genes, as there are different methods (blast, DIAMOND, CD-HIT, minimap2) to determine sequence similarity; (ii) paralogue identification (and splitting) to differentiate paralogues from orthologues and find repeats in a genome, which can be achieved using the local context of genes (synteny), phylogenetic information or graph-based approaches; (iii) type of final output, a directed/undirected graph if a graph is produced, or gene clusters; and (iv) additional functions the tools provide for correction of annotations, assembly errors,

or pre-/post-processing for variant calling, converting file formats, etc [10–12, 22].

For quantitative comparison, the numbers of nodes, edges and connected components were used as metrics to assess the graph size. Pan-genome content was measured based on the average number of genomes per node, and the soft-core thresholding approach, which is implemented frequently in the literature to classify gene clusters [23]: core genome is observed in over 99%, soft core in 95–99%, shell in 15–95% and the cloud genome is observed in less than 15% of the strains in the dataset. Unique genes are defined as the singleton nodes on the graph; they are present in only a single strain. Finally, we established a pairwise comparison in core pan-genome content using the Jaccard index: $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$, where A and B are two different sets of core genes identified by different tools.

3.2.4 Replication of Yakkala *et al.*

Pan-genome graphs were also used to replicate the following findings of the previous study from which the dataset had been obtained [2]:

Identify genes related to different carbon catabolism, and iron acquisition in environmental *A. baumannii* strain isolated from soil, DS002. Find antimicrobial resistance genes associated with biofilm formation, efflux pumps and beta-lactamases in the clinical strains. We processed pan-genome graphs constructed by all five of the tools to identify the nodes which contain genes from only the environmental *A. baumannii* strain DS002; these nodes represent the unique genome content of DS002. Next, we extracted all the genes contained in these nodes to analyse gene enrichment. In our replication study, gene enrichment analyses were performed using the python package GOATOOLS (v0.9.9) [24]. Gene ontology (GO) hierarchy in OBO format was retrieved from the Gene Ontology Website[25], and all the annotated ORFs in our dataset were mapped to GO terms using the ID mapping tool on UniProt [26]. Gene enrichment was performed with correction for false discovery rate (fdr option in GOAtools), and GOATOOLS was also used for plotting GO subgraphs in python. The in-house python script used to perform the analysis (see `runGOE.py` is provided in the Supplementary Text.

3.2.5 Combining methods

Individual steps from two of the tools in our comparative study, Panaroo and Ptolemy, were combined. Briefly, Panaroo uses CD-HIT with a high threshold for sequence identity to obtain gene clusters. These clusters are then collapsed according to synteny information, which is also used to find missing genes and correct for possible errors in assembly and annotation [17]. The error correction in Panaroo can disrupt sequence continuity by breaking up genomes. For that reason, we used Ptolemy to index genomes and connect the nodes to retain the sequence continuity so that each genome can be traversed as a path on the graph.

Panaroo was first run with default parameters in relaxed option (`mode –relaxed`) in order to get an initial estimate of the pan-genome that consists of gene families as nodes in the graph output file `final_graph.gml`. Next, all sequences were indexed with Ptolemy (`extract`), and Panaroo's gene families were reformatted to match Ptolemy's indexing and conform to the format of the syntenic anchor input file in python (see the script `createSA.py` in Supplementary Text). Gene families were then used as input to the canonical quiver construction step in Ptolemy's algorithm (`canonical-quiver`).

The final output is a directed graph stored in a GFA file. The Supplementary Text provides the full list of commands, as well as the in-house scripts used to perform the analysis.

3.3 Results and Discussion

3.3.1 Qualitative and quantitative comparison

To evaluate different tools for pan-genome construction, we selected a number of tools from the literature, Table 3.1 provides a qualitative overview as described in the methods section for the set of five tools we applied to our *A. baumannii* dataset. The most prominent feature among the tools is their compatibility with other software; they accept inputs in standard formats for sequence and annotation data, and produce graph outputs in formats compatible with common graph visualization software. Since all tools construct gene-based pan-genomes, sequences should be annotated with predicted ORFs beforehand, with the exception of PPanGGoLin which can run Prodigal internally for annotation. The tools differ most in their choice of sequence similarity, while usually the synteny or phylogeny (tree-based) information is used for paralogue detection with the exception of Ptolemy, which opts for a repeat graph.

Depending on the aim of pan-genome analysis, some tools could be preferred for the outputs they generate in addition to a graph, although our quantitative comparison on our *A. baumannii* dataset is limited to the graphs and we did not investigate these additional features in our use-case. Both Panaroo and Ptolemy have modules to identify structural rearrangements, while PIRATE, Panaroo and Roary can perform core gene alignment, which can be useful for downstream phylogenetic studies. Similarly, the binary gene presence/absence outputs from PPanGGoLin, PIRATE, Panaroo and Roary can also be used to make a quick and dirty tree or run pan-genome association studies.

Another feature of these tools is that they can be packaged with auxiliary scripts for pre-/post-processing, which can save user time. For instance, Panaroo and Roary both include scripts to perform quality control on the input data before generating a pan-genome graph. Moreover, Roary, Panaroo and PIRATE provide scripts for querying the pan-genome. All tools, except for Ptolemy, have built-in modules to plot pan-genome statistics in various ways. For visualizing the pan-genome graph, we found Ptolemy and PIRATE to be the most straightforward since the GFA outputs can be used directly in Bandage [27]. However, depending on the use-case, Panaroo might be preferred for its GML output, which can be visualized more extensively using Cytoscape and combining additional metadata with the graph [28]. Finally, we note that among these projects, Roary is the only one that is no

Table 3.1: Summary overview of qualitative features of pan-genome tools implemented in this study.

Method	Software	Input	Graph output	Pan-genome	Sequence homology	Paralogue identification
Roary (v3.13.0)	Conda package	GFF3	DOT	Directed graph	BLAST	Synteny
Ptolemy (v1.0)	Java executable	FASTA+GFF	GFA	Directed graph	minimap2	Graph-based
PPanGGoLin (v1.0.13)	Conda package	GBK or FASTA	GEXF	Undirected graph	MMseq2	Synteny
PIRATE (v1.0.3)	Conda package	GFF3	GFA	Directed graph	BLAST (/DIAMOND)	Synteny
Panaroo (v1.1.2)	Conda package	GFF3	GML	Directed graph	CD-HIT	Synteny

longer maintained actively. However, the PPanGGoLin, PIRATE and Panaroo packages have all been extended since completion of our study.

The flowchart in Fig. 3.1 also includes tools not implemented in our study, to provide a guide to help users choose among the state-of-the-art pan-genome tools. For some applications, the flowchart in Fig. 3.1 can lead to multiple tools to choose from. In that case, one can differentiate the tools according to the (i) required inputs or (ii) additional outputs they produce and whether they could benefit from these in the downstream analysis. For instance, Panaconda and PanX both provide visualization of the results [28, 29]. Panaconda's GEXF graph can be viewed using JS visualizer, while panX, having an accompanying web-based interactive application, has more extensive options to visualize the outputs. PanX also provides several statistics on genes (count, length, distribution, etc.), and a phylogenetic tree, all of which can be manipulated and adjusted through its graphical interface. Note that it is not possible to perform the analysis for one's own dataset using the web interface alone. Moreover, Panaconda requires input in PATRIC's feature tab format, in comparison to GBK format as in PanX, which might be less convenient to prepare depending on the data available [29].

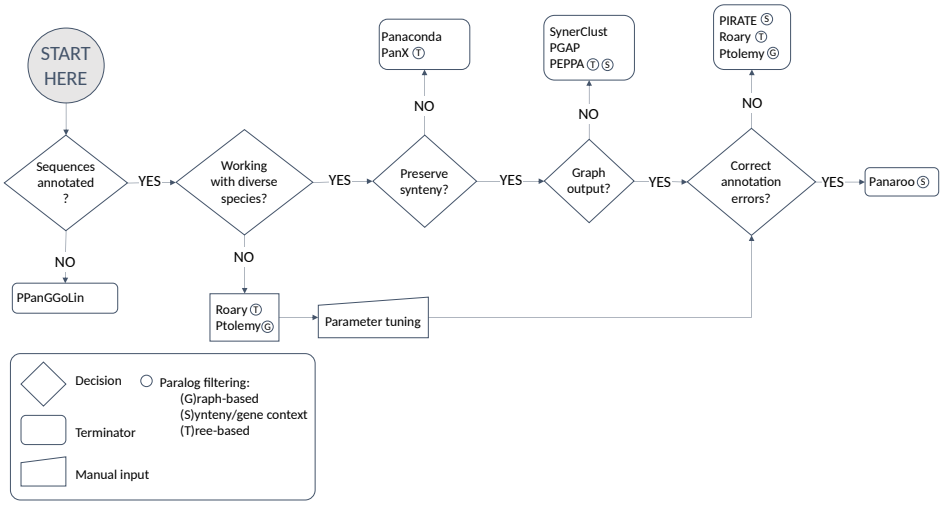


Figure 3.1: Flowchart for users to choose among the current state-of-the-art pan-genome construction tools in the literature, accompanying the analyses in this study..

If one ends up having to decide between SynerClust, PGAP and PEPPA, the input requirements could be considered. PEPPA has the fewest requirements, with only annotations (GFF) and the nucleotide sequence; for PGAP, protein sequences and more extensive gene annotations, including their functional descriptions and COG classifications, must be available [30–32]. For SynerClust, a phylogenetic tree of the genomes needs to be available along with gene annotations (GFF). In terms of the outputs, these three tools are not significantly different except for PGAP, which also performs species evolution and gene function enrichment analysis, and could eliminate the need to run additional tools downstream. Note that, in general, for tools that require annotations beyond to the most

basic ORF predictions, the performance of the upstream annotation step may become even more important, possibly more than the choice of the pan-genome construction tool itself.

While all tools used on the *A. baumannii* dataset for quantitative comparison were run in default settings, we note that all tools considered in Fig. 3.1 are flexible enough to allow for parameter tuning to tailor for different use-cases. The subset of tools implemented in our study are mostly suited for use in microbial organisms out-of-the-box. It is estimated that the *A. baumannii* pan-genome contains 20000 genes [33, 34], although some studies report fewer than or approximately 10000 genes [2, 35]. We presume this number should vary depending on the diversity among the strains in a particular dataset, in addition to the method of choice. *A. baumannii* can colonize a variety of ecological niches; its extreme adaptability is driven by a flexible genome that allows for the acquisition of new genes, and these niche-specific genes can inflate the pan-genome. Except for two strains (SDF and DS002), all genomes in our dataset were isolated from clinical sources, and thus there is little variation in their habitat, and it is likely that the smaller estimate of the pan-genome size is more applicable in our example. Regardless, with an average genome size of 3 Mbp, around 3000 coding sequences, and an estimated pan-genome of 10000 genes, we found A. baumannii to be a middle-of-the-road species among other bacteria [36]. We also note that the *A. baumannii* pan-genome has been classified as open, and that our findings may not generalize to closed pan-genomes [37].

We observe that all tools produced graphs within the range of previous studies (Table 3.2). In terms of graph size and complexity, there is little variation between PPanGGoLin, PIRATE and Panaroo, possibly due to their similar algorithms. Roary and Ptolemy, on the other hand, stand out with the largest graphs, which indicates a more stringent threshold for sequence similarity. When the synteny window size was increased in Ptolemy, the number of nodes varied by as much as 15% (~3000 nodes, see Table S3.2). PPanGGoLin and Panaroo also have the option to run in different modes (`-defrag` in PPanGGoLin, and `-relaxed` in Panaroo) that might allow for some adjustment without parameter tuning (Table 3.2); graph size decreased by 17% (~2000 nodes) in the former (PPanGGoLin with -defrag mode) and increased by less than 5% (~400 nodes) in the latter (Panaroo in `-relaxed` mode). Costa et *al.* also report significant changes in bacterial pan-genomes when the sequence identity thresholds are altered [36]. We recommend that users try different values and settings for the parameters in these tools to improve their results. External annotations, such as known orthologous clusters from the COG database, protein families from the Pfam database, or KEGG pathways can be used to check for the integrity of nodes in the pan-genome graph and possibly guide the parameter tuning, which is beyond the scope of this work.

We also report pan-genome content with the soft-core approach as outlined in the Methods section, in terms of both the number of nodes (or clusters) and the percentage with the respect to the entire pan-genome. Previous studies estimate the core content of the *A. baumannii* pan-genome to be in the range 1500–2500 [38]; in a recent analysis of 2112 *A. baumannii* strains, Mangas *et al.* identified 2221 core genes while the entire pan-genome comprised 19000 genes in total [33]. Yakkala *et al.* also found a total of 7683 genes in the pan-genome, 1344 of which are core genes and 1695 are unique (present in only one genome) [2]. In our analyses, we found the variance in the core gene size to be much smaller compared to the entire pan-genome. While the difference in core genome size

Table 3.2: Quantitative comparison of pan-genome graphs in terms of size and complexity, and the pan-genome content, all run in default configurations, except for the three that also include a different mode.

Method	**Roary**	**Ptolemy**	**PPanGGoLin**		**PIRATE**	**Panaroo**	
Settings	**Default**	**Default**	**Default**	**Defrag**	**BLAST**	**Relaxed**	**Strict**
No. of nodes	13928	20140	11329	9318	7871	10776	10336
No. of edges	21032	31946	17751	14989	11331	16270	14709
No. of connected components	10	34	7	6	416	13	13
Mean sequence length (bp)	803.7	870.5	815.5	849.8	850	824.6	832.5
Average no. of genomes per node	18.7	11.7	22.4	27.1	31.6	24.4	25
Core genes	1996 (14.3%)	1623 (8.1%)	2025 (17.9%)	2223 (23.8%)	2126 (27%)	1910 (17.7%)	2353 (22.8%)
Soft core genes	429 (3.1%)	624 (3.1%)	509 (4.5%)	389 (4.2%)	447 (5.7%)	783 (7.3%)	322 (3.1%)
Shell genes	1912 (13.7%)	1367 (6.8%)	1624 (14.3%)	1526 (16.4%)	1516 (19.3%)	1655 (15.4%)	1609 (15.6%)
Cloud	9591 (68.9%)	16526 (82.1%)	7171 (63.3%)	5180 (55.6%)	3782 (48%)	6419 (59.6%)	6052 (58.5%)
Unique	5276 (37.9%)	13963 (69.3%)	4271 (37.7%)	2522 (27.1%)	1629 (20.7%)	3299 (30.6%)	2968 (28.7%)

Table 3.3: Pairwise similarity of core genome content.

	Roary	**Ptolemy**	**PPanGGoLin**	**PIRATE**	**Panaroo**
Roary		0.68	0.87	0.87	0.84
Ptolemy	0.68		0.74	0.71	0.65
PPanGGoLin	0.87	0.74		0.91	0.85
PIRATE	0.87	0.71	0.91		0.88
Panaroo	0.84	0.65	0.85	0.88	

ranges from 130 to 503 genes, all tools predict the core content to be within the established range from previous studies. We also computed the pairwise Jaccard index for the core genome content, and observed that it varies from 0.65 to 0.91 (Table 3.3). However, when core and soft core genes are considered together, the difference in the number of genes is much smaller.

We observed the largest differences among tools in the cloud genes, as Ptolemy and Roary both stand out with the largest set of cloud genes (Table 3.2). Both tools have a relatively pared-down approach with fewer steps to identify homologous genes, which could possibly lead to a more stringent algorithm that produces clusters with fewer number of genes on average, and an inflated pan-genome. This suggests that the homologue detection step is likely to have the largest influence on the cloud gene content, although it should be further investigated by changing parameter settings, which is beyond the scope of this work. For applications in diverse species, or in cases where the strain speciation is of primary interest, cloud gene content could become more important since the cloud content reflects how an organism has evolved and diversified to adapt to different conditions and environments. In that case, we presume the cloud gene content to play the most important role in deciding which tool to use.

3.3.2 Replication of Yakkala *et al.*

Following the preliminary assessment of pan-genome graphs, we attempted to replicate the major conclusions drawn in a previous study performed on the same *A. baumannii* dataset [2]. Yakkala *et al.* had: (i) first identified genes related to the survival mechanism of *A. baumannii* in diverse environments, and observed that the non-clinical isolate DS002

carried genes which take part in detoxifying aromatic compounds to generate energy. An absence of genes with these functions in clinical isolates suggests the environmental strain had developed an adaptation mechanism in order to survive in soil that is often polluted with phenol-based insecticides. Strain DS002 also showed differences in iron acquisition mechanisms. (ii) Second, the authors reported the absence of genes involved in biofilm formation and efflux pumps, as well as modification of aminoglycoside molecules in non-clinical isolates.

In order to replicate these findings, nodes containing genes from only the soil isolate were identified in the pan-genome output, and a gene enrichment analysis was performed against the entire pan-genome. Note that for this part, only the results from the Ptolemy graph are reported here. Compared to the other tools, we found Ptolemy to have the largest set of significant terms, but when the most significant terms in common were considered, the resulting sets would overlap more than 50% (column 1 in Table S3.3, proportion of common GO terms). The remaining raw gene enrichment test results can be found in the Tables S4 and S5. Fig. 3.2 shows a bar chart of the most significant GO terms associated with these nodes; GO terms related to unique carbon catabolism and iron acquisition are highlighted in red. Conforming with Yakkala *et al.*, the unique gene content of the soil isolate dataset is preserved in the pan-genome graph as well.

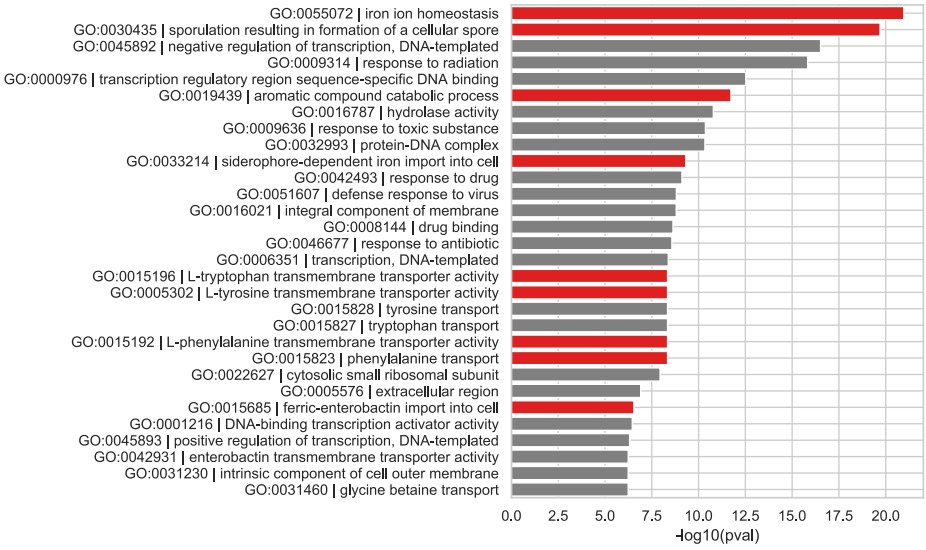


Figure 3.2: Gene enrichment analysis identifying the unique carbon catabolome and iron acquisition mechanisms in the soil isolate, DS002; bar plot of GO terms significantly enriched in DS002. GO IDs and term names are displayed on the y-axis, and the x-axis shows the $-log_{10}$(Pvalue) of each GO term. GO terms associated with specialized mechanisms for carbon catalysis and iron acquisition are in red.

3.3.3 Combining methods

To further demonstrate the power of pan-genome graphs we created a hybrid downstream application for calling structural variants in pan-genome graphs. Following the results in the first part of this study, we attempted to combine individual steps from Panaroo and Ptolemy to obtain a pan-genome that we assert is more suited for this task than using either tool on its own. We report the results from variant calling first in the context of genes involved in carbapenem and amikacin resistance and then for validating and identifying possible novel plasmid structures in our *A. baumannii* dataset.

In the preliminary exploratory part of this study, we found that Panaroo produced average-sized pan-genomes, which suggests it achieves a good balance between over- and under-clustering for our particular use-case. However, we also observed that the error correction step in Panaroo could disrupt the continuity in certain chromosomes, thereby making it difficult to place them within the context of individual genomes. It is more challenging to analyse structural differences without sufficient contextual information. To circumvent this, we attempted to re-introduce sequence continuity by making use of the indexing and path construction steps in Ptolemy. While Panaroo's error correction was shown to be highly useful for handling fragmented assemblies as well as annotation errors in the original study, our dataset consists of only chromosome-level complete assemblies and thus such errors should be negligible and we presume there will not be any major benefits from correcting the assemblies.

Unlike other tools, Ptolemy was developed modularly to the extent that each module can essentially be used independently, provided that the inputs are in the correct format. Hence, it is relatively straightforward to use Ptolemy to (i) index all the genomes, so we can keep track of the order and place of each gene within a genome, and (ii) construct a graph using indexed genes as a guide to connect genomes broken down into separate islands of gene clusters. The resulting graph contains the same set of nodes as those produced by Panaroo initially, but with an increased number of edges to establish sequence continuity. Thus, we obtain a pan-genome graph that preserves whole genomes at a coarse level, and can easily be queried for structural variant calling at small distances.

3.3.4 Structural variation in transposons carrying the *bla*$_{\text{OXA-23}}$ carbapenemase gene

Using the combined method, we explored the structural variation in β-lactamase-carrying transposons in *A. baumannii*. Beta-lactam resistance in *A. baumannii* is mainly driven by class D β-lactamase enzymes (also called oxacillinases or OXAs). The *bla*$_{\text{OXA-23}}$ gene encoding OXAs is readily carried on transposons and thus frequently observed in clinical isolates, both on the chromosome and on plasmids. It is hypothesized that *bla*$_{\text{OXA-23}}$ (red arrows in Fig. 3.3) was mobilized with the help of insertion sequence (IS) ISAb1 (green arrows in Fig. 3.3) [39]. ISAba1 acts as a promotor, and only in the presence of this sequence is the level of gene expression enough to lead to significant imipenem, meropenem and doripenem resistance [40, 41]. Hence, it is also important to investigate the context of the gene in the *A. baumannii* genome and the mechanisms through which it is mobilized among the strains [42].

So far, *bla*$_{\text{OXA-23}}$ has been observed in five contexts in the literature, Tn2006, Tn2007, Tn2008, Tn2008B and Tn2009. Among these, the *A. baumannii* dataset contains strains that

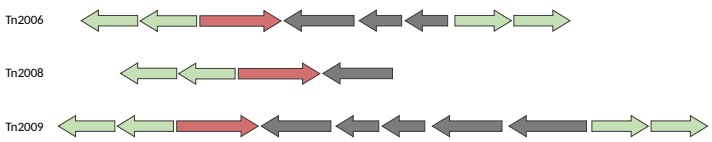


Figure 3.3: Three different structures of *A. baumannii* transposons in which the $bla_{\text{OXA-23}}$ gene is present: ISAba1 is shown with green and $bla_{\text{OXA-23}}$ with red arrows; all arrows are placed according to the direction of the ORFs.

harbour three: Tn2006, Tn2008 and Tn2009 as reported in the literature (Fig. 3.3).

It is possible to locate the $bla_{\text{OXA-23}}$ gene in our pan-genome, and extract the local neighbourhood around this gene. This allows us to visually assess different contexts. Strain 15A5, for instance, has two copies of the β-lactamase gene in a Tn2006 context in its chromosome, and Fig. 3.4 shows one of these copies (node labelled $bla_{\text{OXA-23}}$ in Fig. 3.4) and their surrounding structure. Edges are coloured according to which path they belong to among these three different structures.

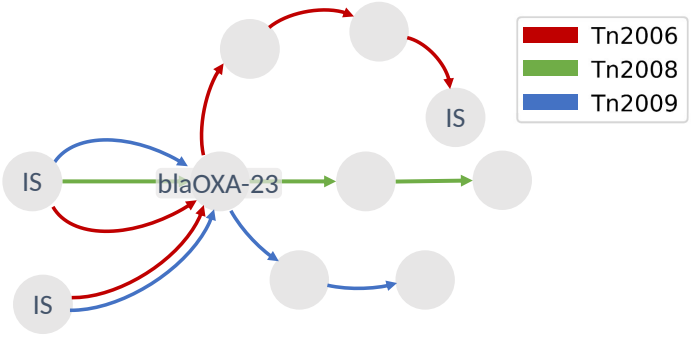


Figure 3.4: Local context of $bla_{\text{OXA-23}}$ in *A. baumannii* strains extracted from the pan-genome graph reveals three different structures of transposons on which the gene is located. Edge colours indicate which structure each path belongs to: red for Tn2006, green for Tn2008 and blue for Tn2009. Only the nodes containing IS and $bla_{\text{OXA-23}}$ are labelled; the unlabeled nodes represent ORFs.

Two ISAb1 sequences are located upstream of $bla_{\text{OXA-23}}$ in reverse direction, and the forward ISs downstream are placed after two proteins of unknown function. Both copies of the upstream IS could be identified for Tn2006 and Tn2009 (two IS nodes left of the $bla_{\text{OXA-23}}$ node in Fig. 3.4, connected with blue and red edges), whereas the downstream IS was detected only for Tn2006. In contrast, the Tn2008-carrying strains AbPK1 and CBA7 are both lacking this second instance of IS, and we did not observe it in the pan-genome graph either. According to the literature, strains BJAB0868 and BJAB07104 carry the Tn2009 transposon, but it was not possible to extract this transposon in its entirety due to the presence of four additional proteins of unknown function, since they increase the length of this syntenic region, thereby making it more difficult to capture it.

3.3.5 Exploring different plasmid structures

In addition to the resistance islands located in the chromosome, antimicrobial resistance genes related to carbapenems and amikacins are also frequently observed in plasmids. Conjugative plasmids play a crucial role in the spread of antibiotic resistance since they facilitate in transferring resistance genes by carrying transposons on which they are contained [43]. RepAci1 and RepAci2 types of plasmids are characterized by the replication proteins (RepB) and the dif modules they contain, and there is only little variance in the DNA sequence (sequence identity over 99.9%), and hence can be represented by the RepAci1 plasmid pA1-1 (accession number CP010782.1) contained in an early strain A1 (also present in our *A. baumannii* dataset). While pA1-1 does not carry resistance genes, they can be found inserted in downstream of plasmid pA1-1 on transposons [44]. Blackwell *et al.* identified several RepAci1 and RepAci2 plasmids and their variants across different strains in [43]. In this section, we analysed the context of the pA1-1 plasmid in the pan-genome graph to find four different variants of this plasmid, three of which had been studied by Blackwell *et al.* (variants 1–3 in Fig. 3.5) and a novel one carrying resistance genes related to multiple drugs in our collection of strains.

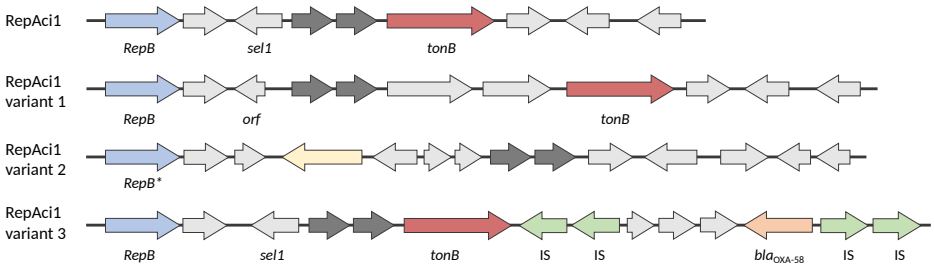


Figure 3.5: RepAci1 plasmid structure and its three variants observed in our *A. baumannii* dataset: variant 1 in p2ABAYE, variant 2 in pD36-3 and variant 3 in pABa3207a; note that variant numbers are assigned arbitrarily in this study to help follow the results.

The pA1-1 plasmid comprises the green path in Fig. 3.6; this path is shared across all plasmids in the *tonB* domain but not in *sel1*. Plasmids p2ABAYE and pD36-3 in strains AYE and D36 are both classified as RepAci1, and they share common paths with pA1-1 but diverge where the *Sel1* protein is replaced with a different *dif* module (variant 1 and variant 2 in Fig. 3.5, respectively), also reported by Blackwell *et al.* [43]. The pABa3207a plasmid from strain 3207 carries the carbapenemase gene $bla_{\text{OXA-58}}$, which had been introduced by repeating IS elements upstream (variant 3 in Fig. 3.5). It is suggested that the *RepB* protein carried by pABa3207a had been mistaken for a variant of *RepB*, and while it also appears as a separate node in the graph (labelled *RepB** protein in Figs 3.5 and 3.6), it remains within close proximity due to shared genomic context with other RepAci1 plasmids.

We also observed a novel variant of a RepAci1 plasmid, pHWBA8-1 and pAB04-1 in strains HWBA8 and Ab04-mff, respectively (yellow path in Fig. 3.6, not shown in Fig. 3.5) which were neither reported by Blackwell *et al.* nor studied yet to our knowledge. These plasmids would be interesting to study as they had been isolated from multidrug-resistant strains and they carry the tetracyline resistance genes *TetB* and *TetR*, as well as the *Sul1* gene, which has been linked to sulfonamide resistance.

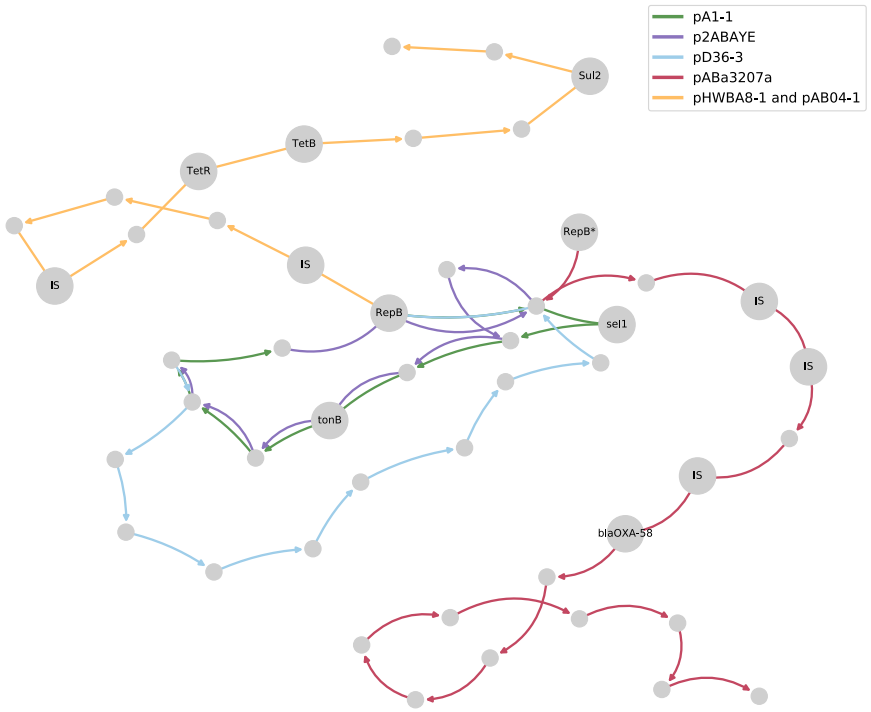


Figure 3.6: Variants of the pA1-1 plasmid sequence across *A. baumannii* strains; unique plasmid structures are differentiated by their edge colours (see key), and all the significant nodes mentioned in the text are labelled with the corresponding product name.

3.4 Conclusions

In this study, we have evaluated the state-of-the-art different pan-genome construction methods to understand the ways in which they can be the most useful to incorporate into existing pipelines and gain insight. We curated a list of tools diverse enough to describe the current literature in pan-genome construction, while still similar in their algorithms so that a meaningful comparison could be made. We provide a flowchart to guide users to select the tool most suited for their application, and we replicate a previous study analysing various survival mechanisms of *A. baumannii*.

Our results on *A. baumannii* suggest that while all the tools produced pan-genome graphs in line with previous work on the same species, they differed significantly in cloud genes. In addition, we found that graph size is likely to be influenced the most by the homologue detection step in the algorithm, and that it can be vary considerably when the parameter settings are changed. Thus, if one desires to go one step forward, and use these tools in more specialized downstream analyses, one must consider parameter tuning or moulding the algorithms available to suit one's own specific purpose. We recommend

that users utilize external databases of known annotations to validate their results for the species they are working on.

Finally, we have provided an example case of structural variant calling in the same *A. baumannii* dataset by combining two of the tools in order to explore (i) the context of the $bla_{\text{OXA-23}}$ carbapenemase gene carried on Acinetobacter transposons and (ii) different structures of RepAci1 plasmids in *A. baumannii* that play a significant role in transmission of antimicrobial resistance genes. Interestingly, we have also identified a novel variant of the RepAci1 plasmid in two clinical strains, carrying resistance genes associated with more than one resistance phenotype, and that would play an important role in understanding the mechanisms of multidrug resistance in *A. baumannii*. We assert the added benefit of combining different tools strategically instead of using any of the tools on their own. Akin to ensemble modelling in the field of machine learning, mixing and matching different methods might be a viable option to consider for constructing pan-genome graphs.

While *A. baumannii* is a good representative of bacterial organisms, our findings are limited to the particular use-case and thus may not be generalizable to species on the more extreme ends, such as *Escherichia coli*, which is reported to be have a higher genome plasticity as well as a larger average genome size (5Mbp long), or *Campylobacter jejuni*, for which the core genome forms a substantial part of the whole pan-genome although on average its genome is much smaller than that of *A. baumannii* (1.5Mbp long, with 40% core genome content). In addition, since the *A. baumannii* pan-genome has been classified as open, our findings may not generalize well to bacterial species with closed pan-genomes. We presume such extreme cases would be the most to benefit from parameter tuning.

References

[1] Aysun Urhan and Thomas Abeel. A comparative study of pan-genome methods for microbial organisms: Acinetobacter baumannii pan-genome reveals structural variation in antimicrobial resistance-carrying plasmids. *Microbial Genomics*, 7(11), 2021.

[2] H. Yakkala, D. Samantarrai, M. Gribskov, and D. Siddavattam. Comparative genome analysis reveals niche-specific genome expansion in acinetobacter baumannii strains. *PLOS ONE*, 14:e0218204, 2019.

[3] National center for biotechnology information (ncbi)[internet, 1988. Accessed 1 Jan 2020.

[4] X. Yang, W.-P. Lee, K. Ye, and C. Lee. One reference genome is not enough. *Genome Biology*, 20(104), 2019.

[5] T. Marschall, M. Marz, T. Abeel, L. Dijkstra, B.E. Dutilh, and A. Ghaffaari. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, 19:118–35, 2018.

[6] A. Dilthey, C. Cox, Z. Iqbal, M.R. Nelson, and G. McVean. Improved genome inference in the mhc using a population reference graph. *Nature Genetics*, 47:682–8, 2015.

[7] E. Garrison, J. Sirén, A.M. Novak, G. Hickey, J.M. Eizenga, and E.T. Dawson. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *nature biotechnology*, 36:875–9, 2018.

[8] E. Biederstedt, J.C. Oliver, N.F. Hansen, A. Jajoo, N. Dunn, and A. Olson. Novograph: Human genome graph construction from multiple long-read de novo assemblies. *F1000Research*, 7(1391), 2018.

[9] H. Li, X. Feng, and C. Chu. The design and construction of reference pangenome graphs, 2020.

[10] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–10, 1990.

[11] B. Buchfink, C. Xie, and D.H. Huson. Fast and sensitive protein alignment using diamond. *Nature Methods*, 12:59–60, 2014.

[12] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28:3150–2, 2012.

[13] A.J. Enright. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30:1575–84, 2002.

[14] M. Steinegger and J. Söding. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9:1–8, 2018.

[15] R. Zarrilli, S. Pournaras, M. Giannouli, and A. Tsakris. Global evolution of multidrug-resistant acinetobacter baumannii clonal lineages. *International Journal of Antimicrobial Agents*, 41:11–9, 2013.

[16] I.P. Salto, G. Torres Tejerizo, D. Wibberg, A. Pühler, A. Schlüter, and M. Pistorio. Comparative genomic analysis of acinetobacter spp. plasmids originating from clinical settings and environmental habitats. *Scientific Reports*, 8:1–12, 2018.

[17] G. Tonkin-Hill, N. MacAlasdair, C. Ruis, A. Weimann, G. Horesh, and J.A. Lees. Producing polished prokaryotic pangenomes with the panaroo pipeline. bioRxiv. 2020;:2020.01.28.922989.

[18] A.J. Page, C.A. Cummins, M. Hunt, V.K. Wong, S. Reuter, and M.T.G.G. Holden. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31:3691–3, 2015.

[19] A.N. Salazar and T. Abeel. Approximate, simultaneous comparison of microbial genome architectures via syntenic anchoring of quiver representations. *Bioinformatics*, 34:i732–42, 2018.

[20] G. Gautreau, A. Bazin, M. Gachet, R. Planel, L. Burlot, and M. Dubois. Ppanggolin: Depicting microbial diversity via a partitioned pangenome graph. bioRxiv. 2019;:836239.

[21] S.C. Bayliss, H.A. Thorpe, N.M. Coyle, S.K. Sheppard, and E.J. Feil. Pirate: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *GigaScience*, 8(598391), 2019.

[22] H. Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34:3094–100, 2018.

[23] B. Contreras-Moreira and P. Vinuesa. Get_homologues, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology*, 79:7696–701, 2013.

[24] Klopfenstein D. V., Zhang L, Pedersen BS, Ramírez F, Vesztrocy AW, and Naldi A. Goatools: A python library for gene ontology analyses. *Scientific Reports*, 8:1–17, 2018.

[25] S. Carbon and C. Mungall. Gene ontology data archive. *Dataset on Zenodo*, 2018.

[26] D480-D489. Uniprot: the universal protein knowledgebase in 2021. *Nucleic acids research*, 49:D480–9, 2021.

[27] R.R. Wick, M.B. Schultz, J. Zobel, and K.E. Holt. Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics*, 31:3350–2, 2015.

[28] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, and D. Ramage. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13, 2003.

[29] J.J. Davis, A.R. Wattam, R.K. Aziz, T. Brettin, R. Butler, and R.M. Butler. The patric bioinformatics resource center: Expanding data and analysis capabilities. *Nucleic Acids Research*, 48, 2020.

[30] C.H. Georgescu, A.L. Manson, A.D. Griggs, C.A. Desjardins, A. Pironti, and I. Wapinski. Synerclust: a highly scalable, synteny-aware orthologue clustering tool. *Microbial genomics*, 4, 2018.

[31] Y. Zhao, J. Wu, J. Yang, S. Sun, J. Xiao, and J. Yu. Pgap: Pan-genomes analysis pipeline. *Bioinformatics*, 28, 2012.

[32] Z. Zhou, J. Charlesworth, and M. Achtman. Accurate reconstruction of bacterial pan- and core genomes with peppan. *Genome Research*, 30:1667–79, 2020.

[33] E.L. Mangas, A. Rubio, R. Álvarez Marín, G. Labrador-Herrera, J. Pachón, and M.E. Pachón-Ibáñez. Pangenome of acinetobacter baumannii uncovers two groups of genomes, one of them with genes involved in crispr/cas defence systems associated with the absence of plasmids and exclusive genes for biofilm formation. *Microbial Genomics*, 5:e000309, 2019.

[34] M.R. Galac, E. Snesrud, F. Lebreton, J. Stam, M. Julius, and A.C. Ong. A diverse panel of clinical acinetobacter baumannii for research and development. *Antimicrobial Agents and Chemotherapy*, 64, 2020.

[35] A.P. Chan, G. Sutton, J. DePew, R. Krishnakumar, Y. Choi, and X.Z. Huang. A novel method of consensus pan-chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of acinetobacter baumannii. *Genome Biology*, 16(143), 2015.

[36] S.S. Costa, L.C. Guimarães, A. Silva, S.C. Soares, and R.A. Baraúna. First steps in the analysis of prokaryotic pan-genomes. *Bioinformatics and Biology Insights*, 14(117793222093806), 2020.

[37] A.P. Chan, G. Sutton, J. DePew, R. Krishnakumar, Y. Choi, and X.Z. Huang. A novel method of consensus pan-chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of acinetobacter baumannii. *Genome Biology*, 16(143), 2015.

[38] L.C.S. Antunes, P. Visca, and K.J. Towner. Acinetobacter baumannii: evolution of a global pathogen. *Pathogens and Disease*, 71:292–301, 2014.

[39] L. Poirel, S. Figueiredo, V. Cattoir, A. Carattoli, and P. Nordmann. Acinetobacter radioresistens as a silent source of carbapenem resistance for acinetobacter spp. *Antimicrobial Agents and Chemotherapy*, 52:1252–6, 2008.

[40] H. Segal, R.K. Jacobson, S. Garny, C.M. Bamford, and B.G. Elisha. Extended -10 promoter in isaba-1 upstream of blaoxa-23 from acinetobacter baumannii [3. *Antimicrobial Agents and Chemotherapy*, 51:3040–1, 2007.

[41] C. Héritier, L. Poirel, T. Lambert, and P. Nordmann. Contribution of acquired carbapenem-hydrolyzing oxacillinases to carbapenem resistance in acinetobacter baumannii. *Antimicrobial Agents and Chemotherapy*, 49:3198–202, 2005.

[42] S.J. Nigro and R.M. Hall. Structure and context of acinetobacter transposons carrying the oxa23 carbapenemase gene. *Journal of Antimicrobial Chemotherapy*, 71:1135–47, 2016.

[43] G.A. Blackwell and R.M. Hall. Mobilisation of a small acinetobacter plasmid carrying an orit transfer origin by conjugative repaci6 plasmids. *Plasmid*, 103:36–44, 2019.

[44] Y. Chen, J. Gao, H. Zhang, and C. Ying. Spread of the blaoxa-23-containing tn2008 in carbapenem-resistant acinetobacter baumannii isolates grouped in cc92 from china. *Frontiers in Microbiology*, 8 FEB:163, 2017.

3.5 Supplementary Material

3.5.1 Supplementary text

In this section we provide a list of all commands and scripts used to run pan-genome tools implemented in the study.

```
# Roary
roary gff3-input*.gff -f output/roary -p 4 -v

# Ptolemy
java -jar ptolemy.jar extract -g ptolemy-input-list.txt -o output/ptolemy
java -jar ptolemy.jar syntenic-anchors --db output -o output/ptolemy
java -jar ptolemy.jar canonical-quiver -s output/ptolemy/syntenic_anchors.txt --db
  output/ptolemy -o output/ptolemy --dump

# PPanGGoLin
ppanggolin workflow --anno ppanggolin-input-list.txt -output output/ppanggolin
```

```
ppanggolin write -p output/ppanggolin/pangenome.h5 --families_tsv -o output/ppanggolin
ppanggolin write -p output/ppanggolin/pangenome.h5 --all_gene_families -o
  output/ppanggolin -f
ppanggolin write -p output/ppanggolin/pangenome.h5 --all_genes -o output/ppanggolin -f

# PIRATE
./PIRATE -i gff3-input -s "50,70,95" -o output/pirate
perl pirate-scripts/subsample_outputs.pl -i output/pirate/PIRATE.gene_families.tsv -g
output/pirate/modified_gffs/ -o output/pirate/gene_families.prev_locus.tsv --field
  "prev_locus"
perl pirate-scripts/gene_cluster_to_binary_fasta.pl -i
  output/pirate/PIRATE.gene_families.tsv output/pirate/binary_presence_absence.fasta
FastTree -fastest -nocat -nome -noml -nosupport -nt
  output/pirate/binary_presence_absence.fasta >
  output/pirate/binary_presence_absence.nwk 2>/dev/null

# Panaroo
python convert_refseq_to_prokka_gff.py -g input.gff -f input.fasta -o output.gff #
  repeat for all inputs
panaroo -i gff3-input/*.gff -o output/panaroo -t 4 -verbose # strict mode
panaroo -i gff3-input/*.gff -o output/panaroo --mode relaxed -t 4 -verbose # relaxed
  mode

# Combining Panaroo and Ptolemy
panaroo -i gff3-input/*.gff -o output/panaroo --mode relaxed -t 4 -verbose # obtain
  gene clusters
java -jar ptolemy.jar extract -g ptolemy-input-list.txt -o output # index graph
python createSA.py output/panaroo output # create syntenic anchor file input to
  Ptolemy
java -jar ptolemy.jar canonical-quiver -s output/syntenic_anchors.txt --db output -o
  output -dump
```

Contents of `createSA.py`

```
#!/usr/bin/env python3
import os, re, sys
import pandas as pd
import networkx as nx
if __name__ == '__main__':

 panaroodir = sys.argv[1]
 ptolemydir = sys.argv[2]

 # Load and process panaroo outputs
 panaroomap = os.path.join(panaroodir, 'gene_data.csv')
 panaroograph = os.path.join(panaroodir, 'final_graph.gml')
 centroid2Loc = pd.read_csv(panaroomap, index_col=2, header=0, \
        names=['strain','location','clusterID','annot','protseq',\
          'dnaseq','gene','desc']).drop(columns=['desc'])
 centroid2Loc.update(centroid2Loc.strain.apply(lambda x:
   x.replace('_reformatted','')))
 g = nx.read_gml(panaroograph, label='id')
 g = nx.relabel_nodes(g,int)
 panaroodf = pd.DataFrame([v for k,v in g.nodes.items()], index=g.nodes())
 panaroodf.loc[:,'locustag'] = panaroodf.apply(lambda x: \
        centroid2Loc.loc[x.seqIDs].annot if isinstance(x.seqIDs,str) \
        else centroid2Loc.loc[x.seqIDs].annot.values, axis=1)

 # Load and process ptolemy outputs
 ptolemymap = os.path.join(ptolemydir, 'orf2id_mapping.txt')
 safile = os.path.join(ptolemydir, 'syntenic_anchors.txt')

 id2orf = pd.read_table(ptolemymap, delimiter='\t', header=None, index_col=2, \
                                                  names=['orf', 'strain', 'id'])
 id2orf['locustag'] = id2orf.apply(lambda x: re.sub('^'+x.strain+'_', '', x.orf),
   axis=1)
 id2orf['locustag'].update(id2orf.locustag.apply(lambda x: \
                        '_'.join(x.split('_')[1:-2]) \
```

```
                                        if len(x.split('_'))<8 else \
                                        '_'.join(x.split('_')[2:-2])))

# Match the panaroo outputs to ptolemys ORF indices
keeplocus = set(sum(panaroodf.locustag.apply(lambda x: \
               [x] if isinstance(x,str) \
                        else list(x)),[]))
keeplocus = set(id2orf.locustag.values).intersection(keeplocus)
df = id2orf[id2orf.apply(lambda x: x.locustag in keeplocus, axis=1)].dropna()
df['newid'] = df.index # df with common ORFs, renumbered (0-based)
df.to_csv(ptolemymap, sep='\t', columns=['orf','strain','newid'], \
          header=False, index=False)
id2orf = df.set_index('locustag')

# Remove ORFs discarded by panaroo from all ptolemy index files in the database
keepID = df.index
df = pd.read_table(os.path.join(ptolemydir,'id2fasta.txt'), \
                   delimiter='\t', header=None, \
                                index_col=0, names=['id','seq']).loc[keepIDs]
df.to_csv(os.path.join(ptolemydir,'id2fasta.txt'), \
          sep='\t', header=False, index=True)
df = pd.read_table(os.path.join(ptolemydir,'global_z.txt'), \
                   delimiter='\t', header=None, \
                                index_col=0).apply(lambda x:
                                  set(map(int,x[1].split(','))), axis=1)
df = df.apply(lambda x: ','.join(map(str,x.intersection(keepID))))
df = df.drop(labels=df[df.apply(lambda x: not(x))].index)
df.to_csv(os.path.join(ptolemydir,'global_z.txt'), \
          sep='\t', header=False, index=True)
df = pd.read_table(os.path.join(ptolemydir,'global_z_prime.txt'), \
          delimiter='\t', header=None, index_col=0)
df.loc[keepID].dropna()[1].to_csv(os.path.join(ptolemydir,'global_z_prime.txt'), \
                                          sep='\t', header=False, index=True)

# Writing out the syntenic anchor file
df = panaroodf.apply(lambda x:
  id2orf.loc[set(x.locustag).intersection(keeplocus),'newid'].values,axis=1)
anchors = df.apply(lambda x: ''.join(['\t'.join([str(mem), \
                                ','.join(map(str,set(x).difference([mem])))+'\n']) \
                                for mem in x if len(x)>1]))
anchors[anchors.apply(lambda x: len(x)==0)] = np.nan
anchors.dropna().to_csv(safile, sep='\t', header=False, index=False, na_rep=None)
```

Perform GO enrichment study

```
python runGOE.py go.obo orf2go.tsv studyinput.txt goestudy.out
```

Contents of `runGOE.py`

```
#!/usr/bin/env python3
from sys import argv
from goatools.obo_parser import GODag
from goatools.go_enrichment import GOEnrichmentStudy
if __name__ == '__main__':

 obofile = argv[1] # obo file obtained from GO website
 gomapfile = argv[2] # tab-separated file mapping ORFs to GO terms
 studyfile = argv[3] # study input (ORFs)
 if len(argv) > 4:
    outfile = argv[4]
 else:
    outfile = 'goestudy.out'

 godag = GODag(obofile, optional_attrs={'relationship'}) # load obo file
 with open(gomapfile, 'r') as f: # load mapping orf -> go IDs
    gomap = {}
    for line in f:
        orf,go = line.strip().split('t')
```

```
        go = go.split(' ')
        gomap[orf] = go
pop = set(gomap.keys())
with open(studyfile, 'r') as f: # load study orf
   study = set([line.strip() for line in f])
   study = study.intersection(pop)

# Create the GOE study object
g = GOEnrichmentStudy(pop, gomap, godag, propagate_counts=False, \
                      alpha=0.01, methods='fdr_bh')
results = g.run_study(study) # run study
g.wr_tsv(outfile, results) # save the study results to a tab-separated file
```


3.5.2 Supplementary tables and figures

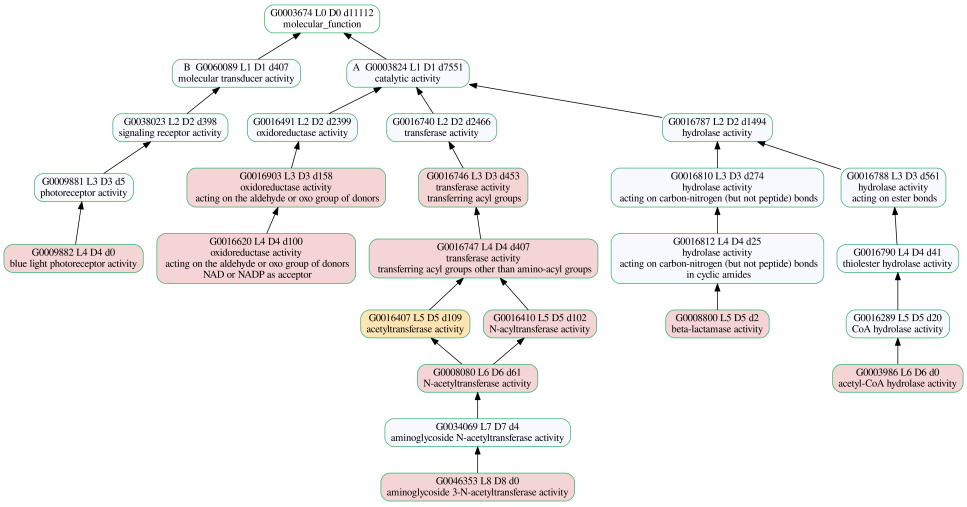


Figure S3.1: A section of GO hierarchy subgraph for the terms significantly enriched in clinical isolates, nodes are color-coded according to the p-value: red for p-value of less than 0.005, orange for p-value between 0.005 and 0.01, and gray for p-values larger than 0.01.

Table S3.1: Genbank accession numbers of Acinetobacter baumannii assemblies used in this study.

Accession	Strain	Accession	Strain
AP013357.1	NCGM 237	CP017656.1	KAB08
AP014649.1	IOMTU 433	CP018143.1	HRAB-85
CP024911.1	AB307-0294	CP018254.1	AF-401
CP003846.1	BJAB07104	CP018256.1	AF-673
CP003847.1	BJAB0715	CP018332.1	A1296
CP003849.1	BJAB0868	CP018421.1	XDR-BJ83
CP003856.1	TYTH-1	CP018664.1	ATCC 17978
CP007535.2	AC29	CP020574.1	15A5
CP007577.1	AC30	CP020578.1	SSA12
CP008706.1	AB5075-UW	CP020579.1	SAA14
CP009256.1	AB031	CP020581.1	SSMA17
CP009257.1	AB030	CP020584.1	JBA13
CP009534.1	AbH12O-A2	CP020586.1	CBA7
CP010397.1	6200	CP020590.1	15A34
CP010779.1	XH386	CP020591.1	SSA6
CP010781.1	A1	CP020592.1	USA2
CP012006.1	Ab04-mff	CP020595.1	USA15
CP012952.1	D36	CP020597.1	HWBA8
CP013924.1	KBN10P02143	CP020598.1	WKA02
CP014215.1	YU-R612	CP021342.1	B8342
CP014528.1	XH858	CP021347.1	B8300
CP014539.1	XH859	CP021782.1	A85
CP014540.1	XH857	CP024124.1	AYP-A2
CP014541.1	XH856	CP024576.1	AbPK1
CP015364.1	3207	CP024611.1	Ab4977
CP015483.1	ORAB01	CP024612.1	Ab4653
CP016298.1	CMC-MDR-Ab59	CP024613.1	Ab4568
CP017152.1	DU202	CP001172.1	AB307-0294_2
CP017642.1	KAB01	CP025266.1	SMC_Paed_Ab_BL01
CP017644.1	KAB02	CP027704.1	DS002
CP017646.1	KAB03	CU459141.1	AYE
CP017648.1	KAB04	CU468230.2	SDF
CP017650.1	KAB05	LN865143.1	CIP70.10
CP017652.1	KAB06	LN868200.1	R2090
CP017654.1	KAB07	LN997846.1	R2091

Table S3.2: Ptolemy graph size at different values of synteny neighborhood size (f).

	# of nodes	# of edges	# of CCs
f = 5	19,541	32,420	28
f = 10*	20,140	31,946	34
f = 15	21,614	33,184	45
f = 20	22,149	33,106	47

Table S3.3: Proportion of common GO terms identified as significantly enriched in isolate DS002 using outputs of different pan-genome tools.

	Ptolemy	Panaroo	PIRATE	PPanGGoLin	Roary
Ptolemy		0.15	0.05	0.67	0.20
Panaroo	0.71		0.32	0.97	0.68
PIRATE	0.50	0.63		1.00	0.94
PPanGGoLin	0.66	0.20	0.11		0.26
Roary	0.69	0.50	0.36	0.93	

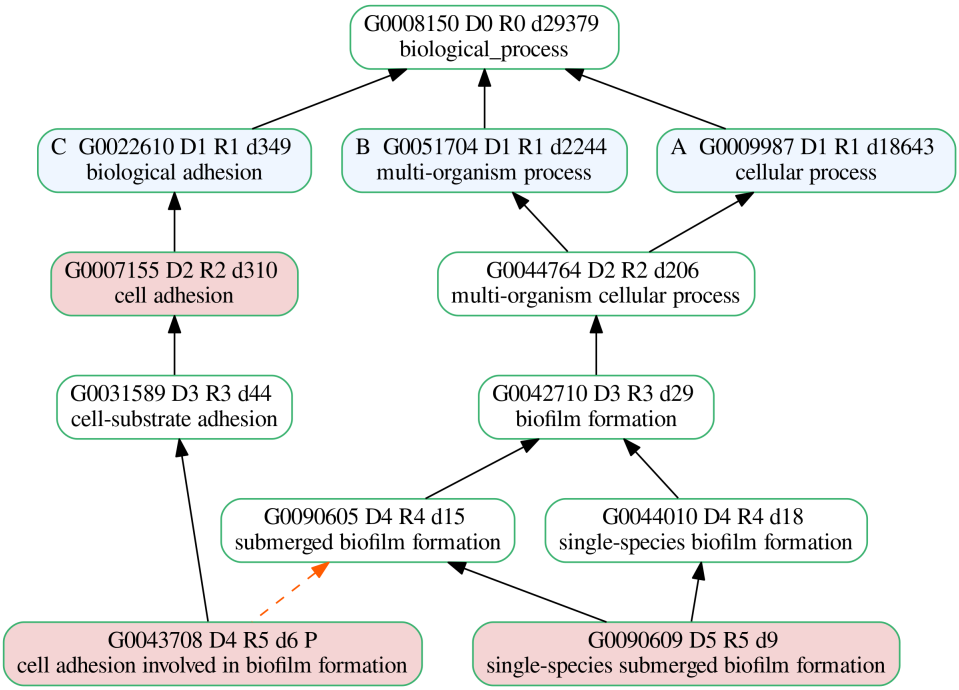


Figure S3.2: GO hierarchy of significant terms related to biofilm formation in drug resistance mechanisms of *A. baumannii*; nodes are color-coded according to the p-value: red for p-value of less than 0.005, orange for p-value between 0.005 and 0.01, and gray for p-values larger than 0.01.

4

SAFPred: Synteny-aware gene function prediction for bacteria using protein embeddings

"If you dream something, it might happen.
If you never dream it, it will never happen."

— Bruce Dickinson

This chapter is based on Urhan et al. SAFPred: Synteny-aware gene function prediction for bacteria using protein embeddings *Bioinformatics*, 2024. [1].

Abstract

Today, we know the function of only a small fraction of the protein sequences predicted from genomic data. This problem is even more salient for bacteria, which represent some of the most phylogenetically and metabolically diverse taxa on Earth. This low rate of bacterial gene annotation is compounded by the fact that most function prediction algorithms have focused on eukaryotes, and conventional annotation approaches rely on the presence of similar sequences in existing databases. However, often there are no such sequences for novel bacterial proteins. Thus, we need improved gene function prediction methods tailored for bacteria. Recently, transformer-based language models - adopted from the natural language processing field - have been used to obtain new representations of proteins, to replace amino acid sequences. These representations, referred to as protein embeddings, have shown promise for improving annotation of eukaryotes, but there have been only limited applications on bacterial genomes.

To predict gene functions in bacteria, we developed SAFPred, a novel synteny-aware gene function prediction tool based on protein embeddings from state-of-the-art protein language models. SAFpred also leverages the unique operon structure of bacteria through conserved synteny. SAFPred outperformed both conventional sequence-based annotation methods and state-of-the-art methods on multiple bacterial species, including for distant homolog detection, where the sequence similarity to the proteins in the training set was as low as 40%. Using SAFPred to identify gene functions across diverse enterococci, of which some species are major clinical threats, we identified 11 previously unrecognized putative novel toxins, with potential significance to human and animal health.

4.1 Introduction

With increasing volumes of sequencing data from high-throughput technologies, the observed diversity of protein sequences is increasing faster than our knowledge of its function. Given costs and the inability to scale experimental and other manual approaches for function prediction, computational approaches have a critical role in deciphering functional diversity. Most state-of-the-art gene function prediction methods have focused on eukaryotes, leaving a gap in our understanding of the vast landscape of diversity among bacteria, which represent some of the most phylogenetically and metabolically diverse taxa.

As with previous tools, we define gene function prediction as the process of mapping terms from the Gene Ontology (GO) knowledgebase to ORFs where the start and stop positions have been annotated [2, 3]. Conventional approaches for gene function prediction rely on sequence homology. Initial methods employed sequence search tools such as BLAST or DIAMOND to query a database of known protein sequences and their functions [4]. While useful, these methods are limited by the completeness and fidelity of their databases. Furthermore, it is often difficult to determine appropriate thresholds, resulting in low sensitivity and specificity [3]. With increasing data, machine learning techniques have been explored; in the most recent Critical Assessment of Functional Annotation (CAFA), a challenge established to evaluate the state-of-the-art in automated function prediction, GOLabeler was the top performer for predicting molecular function ontologies by integrating sequence alignments, domain and motif information, and biophysical properties of

proteins [5].

More recently, deep learning methods leveraging ideas from natural language processing (NLP) have gained attention. Deep learning-based protein language models were recently used to extract embedding vectors for protein sequences that are analogous to word embeddings [6–8]. These vectors capture core properties of proteins beyond primary structure, in a way that is context and species agnostic, but relevant to their function in the cell, which makes them particularly useful for understudied organisms [9]. Contextualized word embeddings have demonstrated success in predicting GO terms, as well as structure and localization prediction, and refining protein family clusters [10].

Compared to eukaryotes, much less has been done to apply NLP-based methods to bacterial genes. In a recent CAFA challenge, methods consistently performed less well on bacteria than eukaryotes, suggesting room for improvement. Furthermore, the prokaryotic track was heavily biased toward a single, well-studied bacterial species, *E. coli* [3], pointing to a need to test methodologies on diverse bacteria. However, more recently Mahlich *et al.* showed that with more sophisticated deep learning methods to study bacterial function, an incredible amount of knowledge can be gained about remote homologs in novel organisms [11, 12]. Given the vast diversity of functional repertoire in bacteria, remote homology detection is of utmost importance.

Many functionally related bacterial genes are encoded in operons, co-located clusters of genes on the same strand, which are often co-regulated and co-transcribed. Thus, the context of a gene is another means to infer clues to its function [13, 14], as it is a source of information complementary to both the sequence and embeddings-based representation of a gene. Leveraging gene context and interactions was shown to improve prediction performance on eukaryotes [15, 16]; however, combining information from gene context with embeddings-based gene representations has not yet been done for gene function prediction.

We developed Synteny-Aware Function Predictor (SAFPred), a novel approach to improve bacterial gene function prediction based on protein embeddings and a comprehensive bacterial synteny database. To evaluate SAFPred, we performed extensive benchmarking using ground truth data and automated function prediction standard approaches to show that SAFPred outperformed conventional sequence-based bacterial genome annotation pipelines, HMM-based approaches, and a state-of-the-art deep learning method, when using gene synteny conservation as additional input. As part of a real-world application, we also demonstrated SAFPred's utility to predict protein functions in *Enterococcus* species, including predicting potential novel pore-forming toxins related to the delta toxin family that could not be recognized using linear sequence or protein domain information. SAFPred provides a powerful new tool for gene function prediction in bacteria, combining state-of-the-art NLP methods with a novel incorporation of syntenic information for bacteria.

4.2 Materials and methods

4.2.1 Datasets

SwissProt data set for benchmarking

We retrieved all the manually reviewed entries from the SwissProt Database (release 2021-04, retrieval date 10 November 2021) [17], which was filtered to include proteins of length 40-1000 amino acids and with at least one experimental GO annotation. We selected the evidence codes EXP, IDA, IPI, IMP, IGI, IEP, HTP, HDA, HMP, HGI, HEP, IBA, IBD, IKR, IRD, IC, and TAS. To reduce redundancy, we clustered the proteins using CD-HIT [18] at 95% sequence similarity. The final dataset comprised 107,818 proteins in total.

To benchmark the performance of our method, we created five benchmarking datasets from SwissProt, one for each of the five most numerous bacterial organisms in our dataset (Table 4.1). Each organism's dataset was split into training and test sets. The test was set composed of all proteins from the specific bacteria. We also divided each training set in different ways to create five sets where the sequence similarity (calculated using BLASTp [4] of test to training set proteins was at most 40%, 50%, 60%, 70% and 80%. This resulted in a total of 30 benchmarking sets (Table 4.1 and Supplementary Text).

Table 4.1: Total number of proteins in the benchmarking sets generated from the SwissProt dataset to evaluate function prediction tools on bacterial organisms. For each organism, the test set remained constant and was composed of all entries from the specific bacterial species, whereas the training set was restricted according to the maximum sequence similarity allowed between the test and training sets.

Organism name	# proteins in the test set	# proteins in the training set (for given similarity between train and test sets)					
		40%	50%	60%	70%	80%	95% (Full)[1]
Escherichia coli (EC)	3454	87014	96471	100445	102262	103229	104377
Mycobacterium tuberculosis (MT)	1666	95367	102531	105158	105917	106114	106152
Bacillus subtilis (BS)	1636	93363	101112	104325	105609	106015	106182
Pseudomonas aeruginosa (PA)	1014	94679	101338	104644	106186	106680	106804
Salmonella typhimurium (ST)	774	100928	104164	105384	105980	106340	107044

[1] 95% similarity was chosen to represent the full dataset to avoid redundancy

Enterococcus diversity dataset

We applied SAFPred to a set of 61,746 proteins with no experimental annotations, representing the entire protein content of 19 *Enterococcus* species, spanning four *Enterococcus* clades [19] (Table S5). This collection of genomes is representative of *Enterococcus* genomic diversity, hence we refer to it as the *Enterococcus* diversity dataset. Assemblies were downloaded from the Assembly Database in NCBI.

4.2.2 Building the bacterial synteny database, SAFPredDB

SAFPredDB is a comprehensive compilation of bacterial syntenic relationships, designed as a resource for SAFPred. It is based on genomic data from the Genome Taxonomy Database (GTDB Release 202, retrieved on 31/03/2022) [20] because GTDB assigns representative genomes based on assembly quality and provides a curated list of species, with consistent labels and IDs to cross-reference to all other databases. Starting with 45,555 representative genomes, we extracted all protein sequences from the standardized GTDB annotations and clustered them using CD-HIT at 95% sequence identity with default parameters, keeping only clusters that contained at least 10 genes, resulting in 372,308 clusters. Next, we

identified synteny by grouping clusters if at least one cluster member was located on the same contig and strand, within 2000 bp (Fig. 4.1A). This yielded 1,488,249 non-singleton candidate regions. Finally, we removed regions with an intergenic distance >300 bp, or split them into multiple regions if possible (Fig. 4.1B). At the end of this procedure, SAFPredDB consisted of 406,293 unique non-singleton regions, and the largest region was 25 genes long.

We used experimentally determined operons collected in the Operon DataBase (ODB v4) [21] to help determine threshold values used when building SAFPredDB, and to validate SAFPredDB. We downloaded both the ODB known and ODB conserved operon databases on 31/03/2022. We identified operons in ODB belonging to *E. coli* and *B. subtilis*, as (i) these two organisms form the basis of a large part of the benchmarking of SAFPredDB, (ii) we could cross-reference the protein IDs in ODB to the locus tags in their respective genome assemblies, and (iii) they are two of the most well-represented organisms in ODB. The ODB conserved operon database contained 8235 unique operons, from which we extracted descriptive statistics and common patterns found across several operons conserved among bacterial organisms. The ODB known operon database was used to model synteny features and determine thresholds, such as region length, number of genes in a region, and the maximum intergenic distance between adjacent genes in a region.

To summarize each SAFPredDB entry, we extracted protein embedding vectors for the representative sequence of clusters found in that entry. We used ESM-1b, a transformer-based protein language model [8] to extract the embeddings, and we took the average of these embeddings to obtain one embedding vector per operon (Fig. 4.1C). Then, we annotated SAFPredDB entries by assigning GO terms, if possible. Since we did not have experimental annotations, we labeled entries based on sequence similarity. We used BLASTp [4] to calculate pairwise sequence similarity between proteins in SAFPredDB entries and the non-redundant SwissProt database with experimentally determined GO terms (all 107,818 entries). We transferred GO terms from significant hits (e-value < 1e-6 and bit score > 50) using the frequency of each GO term among these hits as a predicted score. We could assign at least one GO term to 295,446 of the 372,308 clusters (79%), which in turn yielded 388,377 non-singleton entries (out of 406,293; 96%) annotated with at least one GO term (Table 5.3 in Chapter 5).

In order to keep our synteny database consistent with our benchmarking datasets, where we evaluated SAFPred on training subsets with differing sequence similarity to the proteins in the test set, we generated corresponding subsets of SAFPredDB with matching sequence similarity thresholds. We followed the same procedure as we did to generate subsets of the SwissProt training sets with different sequence similarity thresholds: we used BLAST to calculate the pairwise sequence identity of each query protein to the protein clusters that form our main database. We removed clusters if they were more than 40%, 50%, 60%, 70%, 80% and 95% similar to at least one of the query proteins in the test set. Since this operation removed or altered the content of the entries, we re-calculated the intergenic distances for the remaining clusters and again split regions where the intergenic distance exceeded our 300bp threshold, as we did when we created the main synteny database (Fig. 4.1B-C).

4.2.3 Comparison to published function prediction methods

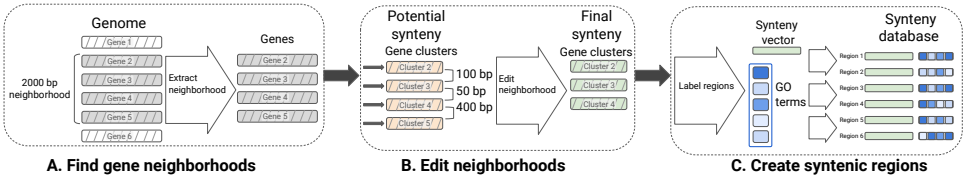


Figure 4.1: Schematic diagram of method used to construct our synteny database, SAFPredDB. Hashed boxes represent genes; solid boxes are numerical embedding vectors. A. 2000 bp-long gene neighborhoods are extracted from all genomes in GTDB; shown is an example with four genes in a single genomic neighborhood (hashed grey boxes). B. After clustering all proteins from GTDB with CD-HIT, we replace the genes with the CD-HIT clusters they belong to (hashed orange boxes) using the amino-acid sequence of the representative gene of each cluster in place of their actual amino-acid sequence. Then, we trim potential syntenic regions to remove genes separated by > 300bp, resulting in final syntenic regions (hashed green boxes). C. Once the final syntenic regions are determined, we i) annotate each region with a set of GO terms, for which we track the corresponding frequency among the gene clusters that make up the region (blue rectangles, darker shades mean GO terms are found in more genes within the region), and ii) extract numerical embedding vectors for each region (solid green boxes). We create a new representation for each region, which consists of the average embedding vector and a set of GO terms. The final synteny database is a collection of such representative embedding vectors and GO term frequency vectors; representations of six example entries are shown here.

Comparison to broadly used function prediction methods as baseline

In our SwissProt benchmarks, we compared SAFPred to two conventional methods of function prediction: i) BLAST (v. 2.12.0) [4], widely used in the literature for comparisons to function prediction tools, and ii) an HMM-based approach, as a more sophisticated baseline.

To predict function using the BLAST baseline, we transferred GO terms from significant BLAST hits taking short sequences into account (e-value < 1e-3, `-task blastp-short`) of a query protein with a predicted score of the value of the maximum sequence identity. As an alternative, we also used the GO term frequency-based approach [3], but we found the maximum sequence identity scoring method performed better in our experiments.

To predict function using the HMM-based approach, we ran the hmmscan command from the HMMER package [22] with the flags `-E 1e-3 --cpu 2 --domtblout` against the Pfam database and applied the frequency-based approach to score transferred annotations, i.e. we transferred GO terms from all significant HMM hits (e-value < 1e-3) to the query protein, using the frequency of a GO term (number of times it was observed among the significant hits) as the predicted score. To compare Pfam outputs quantitatively with those from other methods, we used Pfam2GO mapping tables (version date 2020/12/05) provided by the GO consortium to obtain GO terms corresponding to each Pfam ID in addition to the Pfam database (release 32.0) [23, 24]. Because the Pfam database is independent of the training sets we created based on the SwissProt database, we could not evaluate its dependence on the similarity threshold examined for other tools.

Comparison to a recent state-of-the-art deep learning method

We chose DeepGOPlus (v 1.0.1) [25] as a recent deep learning based comparator in our experiments. A state-of-the-art tool, it uses a supervised approach where a deep convolutional neural network model is combined with a sequence homology based method. We used the implementation provided by the authors and trained the model on the training

sets in our experiments with the optimal values reported for the hyperparameters [25]. We used the same training set for both the BLAST queries and DeepGOPlus.

4.2.4 SAFPred algorithm

SAFPred combines two nearest neighbor (nn) methods: SAFPred-nn, which is based only on amino-acid level embeddings constructed from the SwissProt database, and SAFPred-synteny, which leverages syntenic relationships drawn from our bacterial synteny database.

In SAFPred-nn, we used the ESM-1b protein language model (which we will call ESM) [8] to represent SwissProt entries. To extract amino-acid level embedding vectors, we used `bio_embeddings` (v 0.2.2) [26] with default settings. We obtained protein-level embeddings (1280 dimensional vectors for ESM) by averaging over individual amino acid embeddings. In preliminary work, we also used the ProtT5-XL-U50 model [7], but found that embeddings from ESM performed better (Supplementary Material).

For each query protein, we identified nearest neighbors in the training set based on embedding vector similarity over a threshold, which we calculate separately for each query as the 99th percentile among all pairwise similarity values. We transferred GO terms from nearest neighbors with a score equal to their cosine similarity to the query protein. As the final prediction, we keep only the maximum score for each GO term transferred from the nearest neighbors. We use cosine similarity to determine the similarity between any two embedding vectors $\vec{e_1}$ and $\vec{e_2}$ defined as: $sim(\vec{e_1}, \vec{e_2}) = (\vec{e_1} \cdot \vec{e_2})/(\|\vec{e_1}\| \cdot \|\vec{e_2}\|)$, where $\vec{e_1}$ and $\vec{e_2}$ are both real-valued vectors, $\vec{e_1} \cdot \vec{e_2}$ represents the dot product between $\vec{e_1}$ and $\vec{e_2}$, and $\|\vec{e_i}\|$ is the Euclidean norm of vector $\vec{e_i}$, for $i = 1, 2$.

The SAFPred-synteny component comprises two main steps (Fig. 4.2): (i) assigning syntenic regions to a query from the pre-computed synteny database, SAFPredDB (Fig. 4.2A) and (ii) transferring GO terms from SAFPredDB entries to the query (Fig. 4.2B). SAFPred-synteny follows the same nearest neighbor approach as SAPFred-nn to find the most suitable syntenic regions in SAFPredDB for each query point. In short, we calculate the pairwise cosine similarity between the query point and the average embedding vectors representing database entries. We assign a region to the query if the pairwise similarity between the region and query embeddings is greater than the 99th percentile among all pairwise similarity values.

In our current implementation, we do not have any restrictions on entries assigned to a query protein: given that the most suitable syntenic regions are picked among the same set of regions used to calculate the threshold, at least one region is assigned to each query point.

For all such entries assigned to the query, we also retrieve GO term frequencies. We transfer all GO terms found in the assigned entries using the frequency of the terms multiplied by the cosine similarity of the query point to the entry as the predicted score. For each GO term, the predicted score is the maximum of these values. As the final step in our algorithm, we normalize the predicted scores separately within three GO classes.

SAFPred combines the predictions from SAFPred-nn and SAFPred-synteny by taking the average of predicted scores. We also evaluated its two component predictors individually. Comparing all three methods side by side allowed us to assess the individual contributions from embeddings and our synteny database on SAFPred's performance.

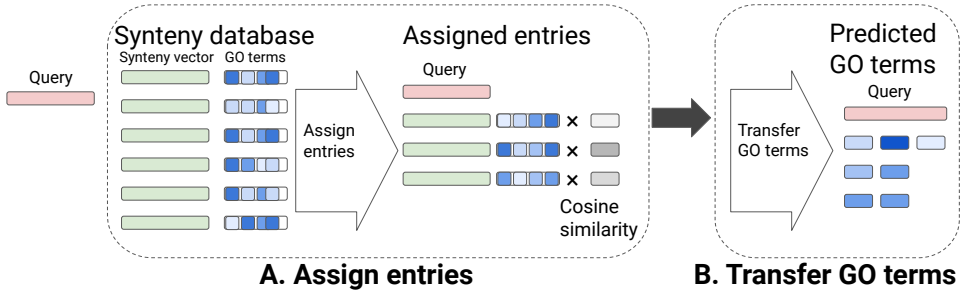


Figure 4.2: Overview of SAFPred-synteny algorithm: predicting GO terms of a query protein. A. SAFPred-synteny assigns an entry (or multiple entries) to the query protein (red filled rectangle on the left) represented using embeddings from ESM-1b LM, based on cosine similarity. Consistent with Fig. 4.1, green rectangles show synteny embeddings paired with the corresponding GO term frequencies (blue rectangles). In this example, three entries that passed the threshold are assigned to the query, and their GO term frequencies are weighted by multiplying by the cosine similarity. B. All GO terms from the assigned entries are transferred to the query, where the final predicted score of a GO term is the maximum of all the multiplied values for the term.

4.2.5 SwissProt benchmark evaluation

Using our SwissProt benchmarking datasets, we evaluated six protein prediction methods: two baselines (BLAST and Pfam), DeepGOPlus, SAFPred, and separately its two components SAFPred-nn and SAFPred-synteny, representing contributions from the embeddings representation and synteny, respectively. In order to make the outputs of all tools comparable to those of DeepGOPlus, we propagated the predicted GO term scores based on the GO hierarchy, as done previously [25]. For each GO term, we assigned the highest predicted score from among its children. This additional post-processing step was only implemented in our benchmarking comparisons across tools, and not in our function prediction across the *Enterococcus* genus.

We evaluated these function prediction methods as done for the CAFA challenges, using the maximum F1-score ($F_{\max}$) and the minimum semantic distance ($S_{\min}$) as described in [3]. We also report the coverage, defined as the percentage of test proteins annotated with at least one GO term at the threshold which maximizes the F1-score. We use leaf nodes in the GO hierarchy only, and remove all ancestor nodes between the leaves and the top of the tree.

4.2.6 Applying SAFPred to a diverse set of enterococcal genomes, including detailed analysis of pore-forming toxins

To demonstrate a practical application of SAFPred, we applied it to the *Enterococcus* diversity dataset. We ran SAFPred in default mode, comparing its output to that from three annotation approaches: (i) prokka (v. 1.14.6) [27], which runs multiple sequence homology-based function prediction tools; (ii) the Pfam database (release 32.0) [28] using the hmmscan command from HMMER (v 3.3.2) [22]; and (iii) eggNOG mapper (v 2.1.10) [29]. All tools were run using default parameters; for HMMER and eggNOG, a significant hit was defined as having e-value < 1e-3.

When examining potential novel *Enterococcus* pore-forming toxins, we performed additional analyses to assess the potential function of query proteins without experimental

annotations: (i) we performed a large-scale structure search using the query protein against AlphaFoldDB and the Protein Data Bank (PDB); (ii) we examined their similarity to known pore-forming toxins found in *Enterococcus* or closely related genera (Table S4), both in terms of structural similarity (using Foldseek), as well as in genomic context; and (iii) we assessed the presence of key structural elements, including N-terminal signal sequences, a common feature in most toxin sequences which guides toxin secretion and transportation outside the cell.

In order to compare syntenic relationships between predicted and known toxin genes, we examined five genes upstream and downstream of toxin genes predicted by SAFPred, as well as for the known delta toxin genes from Table S4, *epx1* and *epx4* [30].

To predict the structure of potential novel toxin genes identified by SAFPred, we used the Fold Sequence public server on ESMFold Atlas [31] which only allows input sequences shorter than 400 amino acids. For longer proteins, we used AlphaFold [32] in monomer mode with default settings, using the Docker implementation. We used Foldseek [33] for both protein structure search against databases and structural alignment. While the structure database search was performed with default settings, we utilized both the global (`--alignment-type 1`) and local alignment options (`--alignment-type 2`) of Foldseek. Following the guidelines available for running Foldseek, we labeled alignments depending on their structural alignment score: highly significant (> 0.7), significant (0.6 - 0.7), nonrandom (0.5 - 0.6) or random ($\leq$ 0.5). To account for large differences in the query and target sequence length, we required the alignment probability to be >0.8. We predicted the N-terminal signal sequences using the SMART server [34].

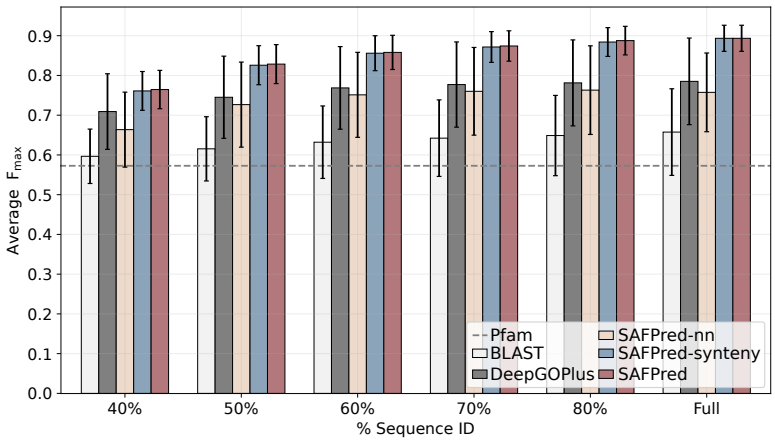


Figure 4.3: SAFPred outperformed conventional approaches to function prediction. Data is averaged across five bacterial species, for variable sequence identity to proteins in the training set (x-axis). Error bars show standard deviations. As the Pfam database is not dependent on the % Sequence identity to the training set, a single value for $F_{\max}$ for the Pfam baseline is shown (dashed line).

4.3 Results

To improve gene function annotation for bacteria, we developed SAFPred, which combines state-of-the-art protein embeddings based on NLP algorithms with bacteria-specific information about gene function inferred from bacterial synteny collected in our database, SAFPredDB, which provides meaningful insight into gene function. This combination outperformed conventional gene function prediction tools and a recent state-of-the-art method, DeepGOPlus, on bacterial genes. We also demonstrated SAFPred's performance on a real-world application where it identified potential novel variants of delta toxin in *Enterococcus*.

4.3.1 SAFPredDB: a database to leverage functional information from syntenic relationships across bacteria

To incorporate information about synteny into SAFPred, we constructed a large-scale database, SAFPredDB, of over 400,000 syntenic regions predicted from > 45,000 representative genomes from across the bacterial kingdom (Methods). We validated SAFPredDB by comparison to the experimentally determined operons found in the conserved ODB [21], a similar online database. SAFPredDB is larger and more up-to-date than ODB, which is based on a smaller, curated list of experimentally determined operons from the literature. Overall, SAFPredDB is quantitatively similar to the conserved ODB, in terms of region length, number of genes in a region and intergenic distance within regions (Fig.s). SAFPredDB provides an extensive catalog of conserved patterns of synteny within the bacterial kingdom (Table 5.1 in Chapter 5).

Table 4.2: F_{max} scores from our benchmarking for six different function prediction tools in the BPO category (MFO and CCO are shown in Table S7), for each of five bacterial species in our full SwissProt benchmarking set. The highest F_{max} score in each column is shown in bold.

	Bacterial species[1]				
	EC	***MT***	***BS***	***PA***	***ST***
Method	**F_{max} scores for BPO**				
BLAST	0.570	0.543	0.639	0.683	0.852
Pfam	0.610	0.513	0.582	0.579	0.579
DeepGOPlus	0.648	0.669	0.857	0.824	**0.928**
SAFPred-nn	0.646	0.636	0.828	0.797	0.880
SAFPred-synteny	0.872	0.837	**0.915**	0.928	0.903
SAFPred	**0.876**	**0.838**	**0.915**	**0.929**	0.902

[1] *EC*: *Escherichia coli*, *MT*: *Mycobacterium tuberculosis*, *BS*: *Bacillus subtilis*, *PA*: *Pseudomonas aeruginosa* and *ST*: *Salmonella typhimurium*.

4.3.2 SAFPred outperforms other tools in function prediction for multiple bacterial species

To assess the performance of SAFPred in assigning GO terms to proteins, we first performed benchmarking on the SwissProt database, where only the proteins with at least one

experimentally determined GO annotation were retained. We then created benchmarking datasets for five different bacterial species, dividing SwissProt entries into training and test sets, thus simulating the real-world scenario of annotating predicted proteins that lack exact matches to database entries.

We benchmarked SAFPred against three previously published tools, including i) a baseline BLAST method; ii) a basic HMM-based approach (HMMER); and iii) a state-of-the-art deep learning method (DeepGOPlus) (Methods). We also compared SAFPred against its two component algorithms run separately, SAFPred-nn (which relies solely on protein embeddings) and SAFPred-synteny (which relies solely on a database of syntenic relationships from operons), allowing us to assess contributions of the two components. We performed benchmarking separately for three categories of GO terms, including Biological Process (BPO), Molecular Function (MFO), and Cellular Component (CCO), as these are known to present different challenges for annotation [35]. Overall, SAFPred achieved the highest $F_{\max}$ scores across all five species, for all three GO categories, and on the full SwissProt benchmarking set, with *S. typhimurium* being the only exception. On this species, DeepGOPlus performed the best for BPO and MFO (Tables 4.2 and S7). We observed similar trends in prediction performance using $S_{\min}$ and the area under the precision/recall curve (Tables S8 and S9). The SAFPRed-nn predictor used alone surpassed conventional tools, showing that protein embeddings, even in a simple unsupervised model, provided a better representation of protein sequence for GO term transfer than both the amino-acid sequence itself (BLAST baseline) and the HMM profiles (Pfam baseline) (Tables 4.2 and S4.1). This agreed with recent studies on eukaryotes [36]. SAFPred-synteny used alone performed substantially better than SAFPred-nn, highlighting the usefulness of incorporating syntenic information. SAFPred-synteny performed almost as well as the full SAFPred tool.

4.3.3 SAFPred surpasses existing tools for annotating distant homologs

We were particularly motivated to develop SAFPred to increase the number of annotations for the growing number of unannotated bacterial proteins, with few or no homologs in existing databases. To emulate gene function prediction of distant homologs, we designed additional benchmarking sets where the pairs of training and test sets were generated by stratifying the full SwissProt dataset based on the maximum sequence similarity allowed between protein sequences in the training and the test set.

As we did not observe any significant differences between the species examined, we report the average $F_{\max}$ values and standard deviation for all five bacteria combined (BPO in Fig. 4.3). SAFPred was consistently the top-performing method. The difference in prediction performance (as measured by $F_{\max}$) between SAFPred and all other methods was greater as the sequence similarity between the test and the training sequences (as well as the clusters in the synteny database) increased (Fig 4.3).

Similar to the full datasets, we observed that protein embeddings (SAFPred-nn) far outperformed both conventional predictors, BLAST and Pfam, across the range of sequence similarities. Furthermore, as with the full datasets, we observed that SAFPred-synteny performed substantially better than SAFPred-nn, and almost as well as SAFPred, demonstrating the large contribution gained by adding information from synteny.

In addition, this benchmarking revealed that BLAST performance was surprisingly

consistent across levels of shared homology, while the embeddings-based methods showed improvement in performance as similarity between the training and test sets increased. This trend held for not only the average F_{max} in the remaining two ontologies (MFO and CCO), but also for each bacterial species individually (Tables S4.7-S4.10).

4.3.4 SAFPred provides more reliable predictions compared to other methods

Among the tools we benchmarked, the BLAST and HMM-based Pfam baselines had the lowest annotation coverages (i.e. the number of test genes that have at least one predicted GO term) on both the full SwissProt dataset and the sets with lower sequence similarity (Tables S15 and S16). SAFPred emerged as the all-around top-performing method in terms of balancing precision and recall. Furthermore, we found that its prediction coverage was in line with other embeddings-based nn models on the full SwissProt benchmarking, although it occasionally lagged behind the state-of-the-art in terms of coverage on our other benchmark sets. Given that SAFPred achieved the best F_{max} values across the board, the drop in coverage means SAFPred's predictions are more reliable compared to other methods.

We did observe that SAFPred's coverage decreased slightly for test sets with lower similarity to the training set (Table S16). In these benchmark tests, SAFPredDB is sparsely labeled due to a conservative annotation methodology (Supplementary Text), limiting the annotations that can be transferred based on synteny.

4.3.5 SAFPred identifies five potential novel pore-forming toxins among a diverse set of enterococcal genomes

A key goal in the development of SAFPred was predicting functions of unannotated genes in bacteria, including those associated with key bacterial features of clinical interest such as antimicrobial resistance and virulence. *Enterococcus* is a diverse genus of bacteria thought to inhabit the gastrointestinal tracts of all land animals. These organisms have an incredibly diverse functional repertoire, yet many of their predicted proteins are of unknown function [19, 37]. Uncovering this rich functional diversity is of primary interest given the ubiquity and importance of this genus. Recent targeted searches have reported the discovery of several classes of novel toxins within diverse enterococcal species, including the discovery of a new family of pore-forming delta toxins in *E. faecalis*, *E. faecium* and *E. hirae* [30] and new botulinum toxins in *E. faecium* [38]. All of these newly discovered toxins exhibit low sequence similarity to known toxin sequences in other bacterial species.

Although the previous studies focused only on three clinically relevant species of *Enterococcus*, we hypothesized that similar toxins could also be found in other diverse, less well-studied species of *Enterococcus*, providing insights into other ecologies in which these toxins may be advantageous. Thus, to search for additional novel toxin genes across the *Enterococcus* genus, we applied SAFPred to a collection of 19 *Enterococcus* genomes, each representing a different species [19], including 16 species not examined by Xiong *et al.* or Zhang *et al* [30, 38]. We looked specifically for genes that were labeled with a GO term describing toxin activity and associated with the conserved genomic context of delta toxins [30]. SAFPred associated 59 genes with the single delta toxin operon from SAPdb, consisting of an enterotoxin and a putative lipoprotein cluster, found in the unrelated

Clostridium and *Roseburia* species (Table S5). Of these 59 genes, 6 were predicted by Pfam to be pore-forming toxins (e-value < 1e-3 to PF01117 or PF03318), and 3 were annotated by Prokka as "lipoproteins" (Methods). The remaining 50 had no functional prediction prior to running SAFPred.

To explore their candidacy as delta toxin encoding, we evaluated each gene's predicted protein structure and genomic context. Eleven (of 59) had structural similarity to known toxin structural folds (Foldseek alignment probability > 0.8 and alignment score > 0.5 to proteins in the AlphaFold and the Protein Data Bank (PDB) structure databases), including several with highly significant alignments (Table S4.1; Fig. S4.34). Of these eleven, five were not previously identified as having a toxin annotation by either Prokka or Pfam - these were detected only by SAFPred. All 11 contained signal peptides at similar positions as those in known bacterial toxins. The remaining 48 proteins without structural similarity had lower SAFPredDB rankings than the 11 with structural similarity (Supplementary Text).

In the absence of experimental annotations, we continued the analysis with the 11 candidate toxins identified by SAFPred to the known pore-forming delta toxin genes previously reported in *Enterococcus*, *epx1* and *epx4*; we compared their genomic context and their neighborhoods [30]. Seven of the 11 candidate toxin genes were most similar to *epx1* structures from *E. faecalis* and *S. aureus*, including five from *E. haemoperoxidus* BAA-382, and two from *E. pernyi* ATCC882. All had surrounding genes with some degree of structural similarity to genes within the known *epx1* genomic neighborhood, including two with highly significant matches (Fig. S37A). Among the 5 putative toxin genes, the highest amino acid sequence identity to *epx1* was less than 40% (Table S4.13). Furthermore, the gene neighborhood was conserved between the five candidates from *E. haemoperoxidus* BAA-382 (Fig. S4.34).

Four of the 11 candidate toxin genes were most similar to the *E. hirae epx4* structure, including one gene from *E. haemoperoxidus* and three genes from *E. moraviensis* BAA-383 (Fig. S4.34B). Among the four putative *epx4* genes, the maximum amino acid sequence similarity we observed to *epx4* was 60% (Table S4.13). Similar to the *epx1* context, we observed that the neighboring genes of the new *epx4*-like toxins predicted by SAFPred were structurally similar to one another. Although some of the neighboring genes had lower Foldseek similarity scores, the neighborhoods had nonrandom similarity among themselves (scores ranging from 0.4 to 0.9).

4.4 Discussion

In this work, we introduce SAFPred, a novel synteny-aware, NLP-based function prediction tool for bacteria. SAFPred is distinguished from existing tools for annotating bacteria in two ways: (i) it represents proteins using embedding vectors extracted from state-of-the-art protein language models, and (ii) it incorporates additional functional information inferred from a protein's genomic neighborhood, by leveraging conserved synteny across the entire bacterial kingdom, tabulated in our synteny database SAFPredDB. This allows SAFPred to identify co-regulated genes that may be part of same functional pathways, but which have completely different sequence or protein structure. To our knowledge, SAFPred is the only bacterial gene function prediction tool with these two features.

While there have been successful uses of protein language models for gene function

prediction in eukaryotes, these methods have not yet been extensively applied to bacteria. We confirmed that protein embeddings in SAFPred surpass conventional sequence homology-based tools, providing a better representation of genes to infer gene function (Table 4.2).

To assess SAFPred's performance on bacteria, we designed a systematic, rigorous benchmarking framework based on the SwissProt database, where we further limited our training set according to its sequence similarity to the test set, in order to evaluate function predictors in the situation where there only distant homologs are known. We examined thresholds down to 40% sequence similarity, as previous work showed that proteins with identity >40% are likely to share functional similarity [39]. However, we know that BLASTp is an imperfect method for identifying homologous relationships for distantly related proteins, and we thus expect that distant homologues, including similarities in protein folds, will be present in our training set. A strength of our tool is its ability to identify functional relationships in distant homologs that sequence comparisons are unable to identify. As we have observed in our function prediction of enterococcal toxins, SAFPred can identify functional linkages to proteins with structural folds that share less than 30% sequence similarity. Thus, we expect that SAFPred's performance would surpass conventional methods when sequence similarity is even lower than the thresholds we implemented in our SwissProt benchmarks. Future work will include benchmarks where we can evaluate SAFPred's performance on unseen genes and assess its generalizability.

Although bacterial gene neighborhoods have been used previously for function prediction, this practice has mostly been manual and is absent from current automated annotation tools. We consistently achieved the best performance when synteny was used in conjunction with the embeddings representation within the SAFPred framework. Either component alone resulted in lower performance, while the biggest gain in prediction performance came from the use of synteny relationships. We demonstrate that conserved synteny and protein embeddings provide complementary information for predicting gene function, in particular when there are fewer homologs available (Fig. 4.3). We presume the overall improvement in prediction accuracy stems from both more accurate function prediction and homolog detection since SAFPred consistently outperforms other methods, even when the sequence similarity between training and test set is low. In future work, a different experiment should be designed to study homolog detection specifically in addition to expanding the set of comparator tools to provide more insight.

We demonstrated that SAFPred improves homolog detection for 19 diverse enterococcal species. Following the recent discovery of several types of novel toxin genes in enterococci, we focused on toxin discovery. SAFPred predicted 11 candidate delta toxin genes, which showed low sequence similarity to known toxins (< 30%) but significant structural homology to known toxin protein structural folds. Several of these candidates also shared similar genomic neighborhood patterns with those of known toxin genes. Although six of these candidate toxins could also be identified based on their Pfam domains, five of these could not be annotated using any of the existing gene prediction tools. These five genes are strong candidates for further experimental validation of their toxin activities. SAFPred also identified 48 additional genes with functional linkages to toxin operons, but without structural homology to known toxins. The function of these genes should be investigated in future studies as well.

One limitation of SAFPred is its reliance on a predicted synteny database, which may contain syntenic linkages that do not share a function, in addition to actual operons. Also, in the absence of ground truth, both the operon predictions and the functions we assigned to these operons are limited by the existing databases (Supplementary Text). To minimize false positives, we adopted a conservative approach which in turn resulted in a sparsely annotated training set, lowering the prediction coverage of SAFPred (Tables S4.11 and S4.12). One way to alleviate this problem is to routinely pick unlabeled entries from our database, prioritizing the most common ones, to perform experiments and identify their functions. With each new experimental annotation available, additional entries can be labeled. We expect this iterative approach to rapidly increase the number of labeled entries in the database.

Another limitation of the current version of SAFPredDB is its focus on broadly conserved patterns; it represents synteny across the entire bacterial kingdom. Since our goal was to develop an all-purpose bacterial gene annotation tool, we deliberately designed our database to be inclusive and to cover as many syntenic regions as possible. Thus, syntenic patterns or operons associated with rare traits, or functional pathways unique to novel species are not present in the default SAFPredDB, but are straightforward to add for specific analyses, as SAFPredDB can be tailored and reconstructed using the latest releases of its source databases. We provide scripts to customize and keep it up to date. For instance, a version of SAFPredDB incorporating metagenomic data could be used to study new functions in uncultured bacteria. Or, to design an all-purpose annotation pipeline for prokaryotes, SAFPredDB could be expanded to cover the diversity of prokaryotes. Although we used only GO terms to describe gene function, the new database could incorporate additional features, such as enzymatic activity and pathways to better capture functional traits. Finally, different representations of synteny vectors in the database, other than taking the average of embeddings, could be explored.

Currently, SAFPred assigns every query gene the same number of entries, equal to 1% of all entries available in the dataset, in order to be as inclusive as possible in learning about unannotated genes. To help disambiguate real matches from false positives, SAFPred reports a rank for each of the matching entries based on their similarity to the query. Although we have not determined whether a universal ranking threshold exists, our detailed examination of toxin operons in *Enterococcus* suggested this ranking can be a reliable proxy for confidence. While SAFPred reported 48 additional genes associated with the delta toxin operon, the delta toxin operon ranked among the top two entries for only the 11 candidate genes that showed structural similarity to the toxin fold. Thus, the order of assigned entries could be used as a proxy to infer confidence.

We demonstrated that conserved synteny and protein embeddings both provide useful information for predicting the protein function. SAFPred outperforms conventional sequence-based bacterial genome annotation pipelines, as well as more sophisticated HMM-based approaches and more recently developed deep learning methods. SAFPred can not only infer beyond the linear sequence, at the level of protein fold, but it can also successfully utilize conserved synteny among bacterial species to predict gene function.

Data Availability

All data used for the analyses in this article are publicly available in the Uniprot database and the Assembly Database at NCBI (accession IDs of *Enterococcus* assemblies listed in Table S5). The code and the scripts developed in this work are public at https://github.com/AbeelLab/safpred.

4.5 Supplementary Material

4.5.1 Supplementary Text: SwissProt Dataset for Benchmarking

SAFPred was developed to improve gene function prediction in bacteria, and to evaluate our method on this specialized task we built a bacterial benchmarking dataset based on the SwissProt database [17].

SwissProt Database (release 2021-04, retrieval date 10 November 2021) forms the basis of our benchmarking dataset. First, we limited the entries to proteins of length [40,1000] to minimize the effect of protein length on our study. Then, we restricted the dataset to include only experimental GO annotations, we retained SwissProt entries with at least one GO term with the evidence codes EXP, IDA, IPI, IMP, IGI, IEP, HTP, HDA, HMP, HGI, HEP, IBA, IBD, IKR, IRD, IC, and TAS. Finally, we clustered the proteins using CD-HIT at 95% sequence similarity threshold to reduce redundancy [18]. At the end of this procedure, we had 107,818 proteins in total. This dataset reflects the full scope of organisms present in SwissProt, with 2005 unique tax IDs, 737 of which belong to bacteria including proteobacteria, terrabacteria, archaea and spirochaetes.

To create our bacterial benchmark datasets, we selected the five bacterial organisms that had the largest number of entries in SwissProt: *Escherichia coli* (EC), *Mycobacterium tuberculosis* (MT), *Bacillus subtilis* (BS), *Pseudomonas aeruginosa* (PA), and *Salmonella typhimurium* (ST). For each bacteria, we obtained the "Full" dataset by reserving all proteins that belong to the organism to the test set. I.e., the "Full" *EC* dataset comprises a test set made up of only *E. coli* genes (3454 entries), and a training set of the remainder of entries from SwissProt (104377 entries).

Next, to evaluate SAFPred on more challenging prediction scenarios where the goal is to predict gene function with few or no homologs in the existing databases, we created additional train/test set pairs for each bacteria. We removed proteins from the training set if they were more than 40%, 50%, 60%, 70% and 80% similar to any protein in the corresponding test set. This led to 6 train/test pairs for bacteria; the test set size remained constant, while the training set decreases in size as the sequence similarity threshold increases because more proteins are removed from the training set. In total, we had 30 train/test pairs (Table 4.1 in the manuscript).

In addition to the training set derived from SwissProt, SAFPred relies on the synteny database, SAFPredDB. Thus, we followed the same procedure to create additional databases corresponding to the preset sequence similarity threshold. We ran BLASTp on all genes in SAFPredDB, and removed members from the database if they were more than 40%, 50%, 60%, 70%, 80% and 95% similar to any protein in the corresponding test set. This operation altered the contents of the entries, thus we recalculated the intergenic distance for each entry in SAFPredDB and split the regions if the intergenic distance exceeded our 300bp threshold.

Table S4.1: Known pore-forming toxin genes found in *Enterococcus* or closely related genera. Protein structures for all but two genes have been solved experimentally and were downloaded from PDB. For Leucotoxin and Alveolysin, AlphaFold was used to predict their structure since they were not found in the databases.

Gene	Sequence ID[1]	Structure ID[2]	Species
Alpha-hemolysin	P09616	3m3r	*S. aureus*
Alveolysin	P23564	-	*P. alvei*
Cytolysin	P19247	4owl	*V. vulnificus*
Gamma-hemolysin component A	P0A071	3b07	*S. aureus*
Gamma-hemolysin component B	A0A0H3JX61	4p1x	*S. aureus*
Heat-labile enterotoxin B chain	P01558	3ziw	*C. perfringens*
Hemolysin	P09545	1xez	*V. cholerae*
Leucotoxin LukE	O54081	-	*S. aureus*
Pore-forming toxin epx1	WP_104660001.1	7t4e	*E. faecalis*
Pore-forming toxin epx4	WP_053766529.1	7t4d	*E. hirae*

[1]Uniprot IDs for the first 8 rows, and NCBI protein IDs for the last two rows.
[2]PDB identifiers.

Table S4.2: SAFPred associated 59 enterococcal proteins with the delta toxin entry from SAFPredDB (entry ID 395786), found in *Clostridium* and *Roseburia*. The contents of this entry are shown here.

Entry	Cluster	Gene	Uniprot ID
395786	75445122	Heat-labile enterotoxin B chain	P01558
	84731674	Uncharacterized lipoprotein YsaB	Q83J37

4.5.2 Analysis of false positive enterococcal toxin predictions

To detect novel toxins, we first collected known pore-forming toxin genes observed in *Enterococcus* or closely related genera. Table S4.1 lists all such toxins, along with their Uniprot IDs and PDB identifiers, if their structure is available in the PDB database.

Among the 10 toxin genes, we located the heat-liable enterotoxin B chain in an entry in SAFPredDB. Forty-eight genes were associated with the toxin operon in SAFPredDB, but did not have structural similarity to a known toxin protein fold. The SAFPredDB entry codes for two membrane-associated proteins: 1) the toxin gene itself, and 2) another membrane protein of unknown function. Based on FoldSeek search results, 33 of these 48 genes appeared to perform other functions related to the cell membrane, involving signaling, secretion, and pore formation. To gain further insight into why these 48 additional genes without toxin structural folds were matched to the delta toxin entry in SAFPredDB, we compared the Euclidean similarity of embeddings vectors extracted from all 59 genes to the embedding vector of the known delta toxin operon (Table S4.2). While we did not find any significant differences in terms of the absolute similarity values between those 11 that had structural similarity to the delta toxin operon and the rest, we noted that the SAFPredDB toxin entry was consistently ranked in the top two among all SAFPredDB entries assigned to the 11 likely toxins, whereas for the remaining 48 genes this was not the case. This

suggests that the ranking could be used in future versions of SAFPred as a proxy to infer confidence in region assignments and to reduce false positives.

4.5.3 Supplementary Tables

Table S4.3: Genome metadata and assembly statistics of the 19 *Enterococcus* genomes used in the *Enterococcus* diversity dataset.

Species	Clade[1]	NCBI Accession	Genome size (bp)	N50 (bp)	No. of components[2]	Completeness[3]	Contamination[4]	No. of CDS[5]
E. caccae	I	GCA_000407145.1	3551922	1036503	7	98.99	1.01	3244
E. haemoperoxidus	I	GCA_000407165.1	3581919	2345301	2	100	1.52	3207
E. moraviensis	I	GCA_000407445.1	3586110	878836	6	98.99	0	3337
E. durans	II	GCA_000407265.1	3170574	972116	11	98.99	0	3031
E. hirae	II		2882665	676112	6	98.99	0	2667
E. pernyi	II	GCA_000407465.1	3078858	2163032	14	98.99	0	2891
E. phoeniculicola	II	GCA_000407505.1	3910972	889011	11	100	0	3507
E. villorum	II	GCA_000407205.1	3060724	1302785	4	98.99	0	2821
E. avium	III	GCA_000407245.1	4614282	652010	13	98.99	0	4504
E. gilvus	III	GCA_000407545.1	4179913	3038269	5	98.99	0	4111
E. malodoratus	III	GCA_000407185.1	4627134	1188036	7	98.99	0	4500
E. pallens	III	GCA_000407485.1	5437454	1015307	13	100	0	5196
E. raffinosus	III	GCA_000407525.1	4310366	1007288	9	98.99	0	4202
E. asini	IV	GCA_000407365.1	2572706	2038454	2	98.99	0	2429
E. cecorum	IV	GCA_000492155.1	2408477	1060610	6	98.99	0	2375
E. columbae	IV	GCA_000407225.1	2545099	421089	9	98.99	0	2327
E. dispar	IV	GCA_000407585.1	2812918	2801753	4	98.99	0	2635
E. saccharolyticus	IV	GCA_000407285.1	2604038	2432614	2	98.99	0	2586
E. sulfureus	IV	GCA_000407605.1	2301651	733717	8	98.99	0	2176

[1] *Enterococcus* clades, defined by [19].
[2] Number of contigs/scaffolds or chromosomes.
[3,4] Calculated using CheckM [40].
[5] Total CDS predicted by Prokka [27].

Table S4.4: $F_{\max}$ values for the full SwissProt dataset for the molecular function and cellular component ontologies.

$F_{\max}$ values for each of five species[1,2]					
Method	***EC***	***MT***	***BS***	***PA***	***ST***
Molecular function					
BLAST	0.613	0.593	0.625	0.699	0.814
Pfam	0.650	0.549	0.571	0.534	0.559
SAFPred-nn	0.675	0.723	0.814	0.854	0.837
SAFPred-synteny	0.880	0.869	0.893	0.938	0.878
SAFPred	0.885	0.869	0.893	0.938	0.877
DeepGOPlus	0.686	0.755	0.841	0.883	0.911
Cellular component					
BLAST	0.569	0.397	0.638	0.700	0.871
Pfam	0.625	0.541	0.608	0.560	0.616
SAFPred-nn	0.731	0.500	0.898	0.900	0.917
SAFPred-synteny	0.920	0.847	0.943	0.945	0.918
SAFPred	0.922	0.847	0.943	0.945	0.917
DeepGOPlus	0.745	0.567	0.885	0.887	0.936

[1] *EC: Escherichia coli*, *MT: Mycobacterium tuberculosis*, *BS: Bacillus subtilis*, *PA: Pseudomonas aeruginosa* and *ST: Salmonella typhimurium.*
[2] Lowest value for each species and for each GO category are shown in bold.

Table S4.5: S_{min} values for the full SwissProt dataset for each of five species[1,2].

Method	*EC*	*MT*	*BS*	*PA*	*ST*
Biological process					
BLAST	20.97	22.93	15.99	19.02	8.53
Pfam	108.50	161.14	126.22	122.86	129.13
SAFPred-nn (T5)	18.61	18.44	9.84	12.76	8.83
SAFPred-nn	17.44	18.24	9.02	12.18	8.37
SAFPred-synteny	**7.10**	**7.60**	**3.27**	**4.41**	4.90
SAFPred	**7.10**	**7.60**	**3.27**	**4.41**	4.90
DeepGOPlus	16.45	14.94	5.68	10.39	**2.85**
Molecular function					
BLAST	10.97	11.07	7.99	8.25	5.51
Pfam	30.87	52.74	35.04	35.48	37.17
knn (T5)	10.01	8.99	4.74	4.52	5.72
knn	9.12	8.04	4.77	4.49	5.67
SAP-operon	**3.63**	**3.51**	**1.88**	**1.66**	2.72
SAP	**3.63**	**3.51**	**1.88**	**1.66**	2.72
DeepGOPlus	8.68	6.36	2.57	2.67	**1.68**
Cellular component					
BLAST	6.46	5.74	3.87	4.17	2.53
Pfam	30.29	35.91	34.37	31.40	32.15
knn (T5)	4.87	4.99	1.76	1.81	2.37
knn	5.02	5.00	1.84	2.05	2.29
SAP-operon	**1.64**	**1.85**	**0.49**	**0.82**	1.35
SAP	**1.64**	**1.85**	**0.49**	**0.82**	1.35
DeepGOPlus	4.92	4.63	1.55	1.77	**1.18**

[1] *EC: Escherichia coli*, *MT: Mycobacterium tuberculosis*, *BS: Bacillus subtilis*, *PA: Pseudomonas aeruginosa* and *ST: Salmonella typhimurium*.
[2] Lowest value for each species and for each GO category are shown in bold.

Table S4.6: Area under the precision/recall curve[1] for the full SwissProt dataset, for all 5 bacterial organisms[2] and three GO categories.

Method	*EC*	*MT*	*BS*	*PA*	*ST*
Biological process					
BLAST	0.519	0.497	0.686	0.642	0.642
Pfam	0.648	0.556	0.582	0.623	0.623
SAFPred-nn (T5)	0.529	0.514	0.76	0.714	0.714
SAFPred-nn	0.544	0.533	0.80	0.714	0.714
DeepGOPlus	0.564	0.600	0.862	0.797	0.797
SAFPred-synteny	**0.855**	**0.805**	**0.907**	**0.908**	**0.908**
SAFPred	**0.855**	**0.805**	**0.907**	**0.908**	**0.908**
Molecular function					
BLAST	0.569	0.573	0.658	0.644	0.644
Pfam	0.686	0.599	0.571	0.595	0.595
SAFPred-nn (T5)	0.523	0.602	0.764	0.774	0.774
SAFPred-nn	0.560	0.640	0.790	0.780	0.780
DeepGOPlus	0.613	0.719	0.857	0.899	0.899
SAFPred-synteny	**0.872**	**0.849**	**0.887**	**0.921**	**0.921**
SAFPred	**0.872**	**0.849**	**0.887**	**0.921**	**0.921**
Cellular component					
BLAST	0.541	0.324	0.701	0.724	0.724
Pfam	0.675	0.598	0.608	0.634	0.634
SAFPred-nn (T5)	0.641	0.344	0.858	0.849	0.849
SAFPred-nn	0.633	0.335	0.879	0.853	0.853
DeepGOPlus	0.747	0.512	0.931	0.917	0.917
SAFPred-synteny	**0.919**	**0.831**	**0.954**	**0.945**	**0.945**
SAFPred	**0.919**	**0.831**	**0.954**	**0.945**	**0.945**

[1] The highest value for each species and for each GO category are shown in bold.

[2] *EC: Escherichia coli*, *MT: Mycobacterium tuberculosis*, *BS: Bacillus subtilis*, *PA: Pseudomonas aeruginosa* and *ST: Salmonella typhimurium*

Table S4.7: F_{max} values for the entire *E. coli* SwissProt benchmark sets.

Method	40	50	60	70	80	Full
Biological process						
BLAST	0.505	0.526	0.540	0.556	0.560	0.570
Pfam	0.610	0.610	0.610	0.610	0.610	0.610
SAFPred-nn (T5)	0.464	0.510	0.574	0.596	0.607	0.625
SAFPred-nn	0.536	0.584	0.623	0.635	0.639	0.646
DeepGOPlus	0.568	0.601	0.624	0.636	0.644	0.648
SAFPred-synteny	0.720	0.792	0.824	0.840	0.858	0.877
SAFPred	0.724	0.800	0.830	0.846	0.865	0.877
Molecular function						
BLAST	0.554	0.573	0.585	0.599	0.603	0.613
Pfam	0.650	0.650	0.650	0.650	0.650	0.650
SAFPred-nn (T5)	0.475	0.528	0.595	0.611	0.618	0.632
SAFPred-nn	0.558	0.598	0.656	0.666	0.669	0.675
DeepGOPlus	0.605	0.641	0.649	0.674	0.680	0.686
SAFPred-synteny	0.725	0.803	0.833	0.850	0.865	0.886
SAFPred	0.739	0.810	0.840	0.855	0.872	0.886
Cellular component						
BLAST	0.502	0.519	0.532	0.546	0.554	0.569
Pfam	0.625	0.625	0.625	0.625	0.625	0.625
SAFPred-nn (T5)	0.560	0.610	0.691	0.712	0.732	0.738
SAFPred-nn	0.610	0.650	0.707	0.715	0.723	0.731
DeepGOPlus	0.645	0.705	0.721	0.706	0.726	0.745
SAFPred-synteny	0.769	0.835	0.866	0.887	0.905	0.925
SAFPred	0.774	0.842	0.873	0.893	0.911	0.925

Table S4.8: $F_{\max}$ values for the entire *M. tuberculosis* SwissProt benchmark sets.

Method	40	50	60	70	80	Full
Biological process						
BLAST	0.525	0.521	0.531	0.532	0.539	0.543
Pfam	0.513	0.513	0.513	0.513	0.513	0.513
SAFPred-nn (T5)	0.520	0.597	0.613	0.615	0.618	0.618
SAFPred-nn	0.575	0.617	0.629	0.630	0.633	0.636
DeepGOPlus	0.627	0.645	0.670	0.667	0.667	0.669
SAFPred-synteny	0.698	0.750	0.786	0.812	0.825	0.838
SAFPred	0.703	0.750	0.787	0.813	0.828	0.838
Molecular function						
BLAST	0.589	0.582	0.580	0.584	0.591	0.593
Pfam	0.549	0.549	0.549	0.549	0.549	0.549
SAFPred-nn (T5)	0.586	0.654	0.672	0.677	0.683	0.681
SAFPred-nn	0.654	0.706	0.714	0.718	0.722	0.723
DeepGOPlus	0.709	0.736	0.745	0.752	0.750	0.755
SAFPred-synteny	0.748	0.800	0.830	0.851	0.864	0.869
SAFPred	0.748	0.801	0.832	0.856	0.866	0.869
Cellular component						
BLAST	0.401	0.397	0.394	0.396	0.396	0.397
Pfam	0.541	0.541	0.541	0.541	0.541	0.541
SAFPred-nn (T5)	0.434	0.513	0.520	0.517	0.510	0.507
SAFPred-nn	0.431	0.504	0.505	0.505	0.502	0.500
DeepGOPlus	0.570	0.577	0.575	0.573	0.572	0.567
SAFPred-synteny	0.634	0.700	0.753	0.804	0.835	0.846
SAFPred	0.638	0.704	**0.756**	0.807	0.835	0.846

Table S4.9: F_{max} values for the entire *P. aeruginosa* SwissProt benchmark sets.

Method	40	50	60	70	80	Full
Biological process						
BLAST	0.650	0.666	0.680	0.686	0.684	0.683
Pfam	0.579	0.579	0.579	0.579	0.579	0.579
SAFPred-nn (T5)	0.629	0.731	0.786	0.798	0.797	0.796
SAFPred-nn	0.681	0.769	0.794	0.799	0.797	0.797
DeepGOPlus	0.749	0.785	0.812	0.820	0.816	0.824
SAFPred-synteny	0.754	0.835	0.879	0.892	0.916	0.927
SAFPred	0.755	0.837	0.878	0.896	0.922	0.927
Molecular function						
BLAST	0.712	0.721	0.716	0.708	0.702	0.699
Pfam	0.534	0.534	0.534	0.534	0.534	0.534
SAFPred-nn (T5)	0.679	0.801	0.848	0.858	0.856	0.853
SAFPred-nn	0.765	0.849	0.863	0.863	0.857	0.854
DeepGOPlus	0.810	0.856	0.883	0.894	0.882	0.883
SAFPred-synteny	0.789	0.879	0.913	0.921	0.932	0.938
SAFPred	0.784	0.882	0.915	0.925	0.939	0.938
Cellular component						
BLAST	0.658	0.687	0.692	0.701	0.701	0.700
Pfam	0.560	0.560	0.560	0.560	0.560	0.560
SAFPred-nn (T5)	0.743	0.838	0.876	0.894	0.900	0.898
SAFPred-nn	0.780	0.868	0.894	0.900	0.900	0.900
DeepGOPlus	0.819	0.847	0.874	0.877	0.900	0.887
SAFPred-synteny	0.823	0.890	0.914	0.927	0.934	0.945
SAFPred	0.830	0.893	0.914	0.931	0.943	0.945

Table S4.10: $F_{\max}$ values for the entire *S. typhimurium* SwissProt benchmark sets.

Method	40	50	60	70	80	Full
Biological process						
BLAST	0.675	0.728	0.775	0.800	0.822	0.852
Pfam	0.579	0.579	0.579	0.579	0.579	0.579
SAFPred-nn (T5)	0.634	0.739	0.795	0.848	0.875	0.869
SAFPred-nn	0.774	0.855	0.889	0.910	0.919	0.880
DeepGOPlus	0.811	0.860	0.896	0.911	0.922	0.928
SAFPred-synteny	0.811	0.871	0.894	0.906	0.908	0.905
SAFPred	0.811	0.871	0.895	0.907	0.910	0.905
Molecular function						
BLAST	0.674	0.723	0.751	0.782	0.811	0.814
Pfam	0.559	0.559	0.559	0.559	0.559	0.559
SAFPred-nn (T5)	0.589	0.694	0.753	0.818	0.856	0.829
SAFPred-nn	0.761	0.844	0.868	0.894	0.904	0.837
DeepGOPlus	0.800	0.845	0.875	0.892	0.904	0.911
SAFPred-synteny	0.796	0.863	0.884	0.893	0.889	0.883
SAFPred	0.798	0.862	0.885	0.895	0.890	0.883
Cellular component						
BLAST	0.644	0.705	0.737	0.774	0.815	0.871
Pfam	0.616	0.616	0.616	0.616	0.616	0.616
SAFPred-nn (T5)	0.794	0.848	0.871	0.898	0.918	0.916
SAFPred-nn	0.824	0.881	0.909	0.923	0.941	0.917
DeepGOPlus	0.819	0.868	0.898	0.901	0.917	0.936
SAFPred-synteny	0.856	0.896	0.912	0.920	0.922	0.918
SAFPred	0.858	0.899	0.916	0.922	0.923	0.918

Table S4.11: Prediction coverage[1] (in %) for the full SwissProt dataset, for all 5 bacterial organisms[2] and three GO categories.

Method	*EC*	*MT*	*BS*	*PA*	*ST*
Biological process					
BLAST	65.66	66.14	75.58	73.11	87.61
Pfam	81.41	71.19	80.69	75.03	84.42
SAFPred-nn (T5)	84.78	91.40	91.58	91.00	97.15
SAFPred-nn	88.29	90.16	94.88	94.12	97.82
DeepGOPlus	88.79	81.92	94.80	95.80	97.49
SAFPred-synteny	92.48	82.36	90.68	92.32	90.45
SAFPred	92.66	82.36	90.68	92.32	90.45
Molecular function					
BLAST	74.04	69.51	75.25	81.36	89.23
Pfam	88.46	88.80	87.28	88.82	87.44
SAFPred-nn (T5)	89.64	96.53	91.47	98.59	97.06
SAFPred-nn	88.35	96.00	93.56	98.33	97.23
DeepGOPlus	98.05	92.80	89.69	94.86	96.25
SAFPred-synteny	92.85	92.98	88.29	93.06	88.42
SAFPred	93.33	92.98	88.29	93.06	88.42
Cellular component					
BLAST	60.29	58.41	68.377	77.51	87.05
Pfam	81.65	76.35	81.981	84.43	89.68
SAFPred-nn (T5)	82.39	67.22	93.437	93.60	96.15
SAFPred-nn	81.36	64.92	94.153	93.43	96.76
DeepGOPlus	89.86	95.08	97.017	94.29	97.98
SAFPred-synteny	93.32	92.86	95.943	91.35	96.96
SAFPred	93.53	92.86	95.943	91.35	96.96

[1] The number of test proteins annotated with at least one GO term at the threshold which maximizes the F1-score.

[2] *EC: Escherichia coli*, *MT: Mycobacterium tuberculosis*, *BS: Bacillus subtilis*, *PA: Pseudomonas aeruginosa* and *ST: Salmonella typhimurium*

Table S4.12: Prediction coverage[1] (in %) for the entire SwissProt benchmark evaluation, averaged over 5 bacterial organisms for three GO categories. The standard deviation across the 5 bacterial organisms is indicated in parentheses.

Method	Maximum sequence % ID between the training and the test set					
	40%	50%	60%	70%	80%	Full
Biological process						
BLAST	71.16% (3.30%)	70.20% (5.82%)	70.21% (5.00%)	71.76% (5.62%)	72.02% (4.49%)	73.62% (7.99%)
Pfam	78.55% (4.77%)	78.55% (4.77%)	78.55% (4.77%)	78.55% (4.77%)	78.55% (4.77%)	78.55% (4.77%)
SAFPred-nn (T5)	85.71% (4.99%)	88.23% (6.43%)	90.09% (3.74%)	89.03% (4.80%)	89.93% (4.49%)	91.18% (3.92%)
SAFPred-nn	85.33% (4.07%)	91.58% (6.72%)	93.23% (3.52%)	92.40% (4.24%)	92.76% (3.78%)	93.06% (3.41%)
DeepGOPlus	95.62% (1.56%)	94.48% (3.45%)	93.84% (4.99%)	94.44% (3.16%)	90.25% (6.16%)	91.76% (5.73%)
SAFPred-synteny	85.48% (3.18%)	87.36% (1.05%)	87.93% (3.90%)	89.95% (4.43%)	91.42% (0.83%)	89.66% (3.74%)
SAFPred	84.04% (1.82%)	87.93% (0.92%)	88.11% (3.90%)	89.17% (3.60%)	91.73% (0.91%)	89.69% (3.77%)
Molecular function						
BLAST	75.42% (3.39%)	75.52% (2.93%)	74.33% (3.73%)	74.51% (4.34%)	75.17% (4.20%)	77.88% (6.82%)
Pfam	88.16% (0.67%)	88.16% (0.67%)	88.16% (0.67%)	88.16% (0.67%)	88.16% (0.67%)	88.16% (0.67%)
SAFPred-nn (T5)	89.55% (4.20%)	92.16% (6.19%)	94.44% (3.49%)	94.46% (3.41%)	94.62% (3.10%)	94.66% (3.47%)
SAFPred-nn	89.67% (5.32%)	92.59% (6.82%)	94.83% (3.31%)	94.68% (3.32%)	94.69% (3.49%)	94.69% (3.55%)
DeepGOPlus	93.79% (2.32%)	93.98% (3.29%)	94.46% (2.71%)	93.98% (3.00%)	95.25% (2.48%)	94.33% (2.89%)
SAFPred-synteny	87.07% (3.14%)	88.57% (1.60%)	90.65% (1.65%)	91.35% (1.97%)	91.36% (2.33%)	91.12% (2.26%)
SAFPred	85.37% (2.80%)	89.20% (1.75%)	90.96% (1.49%)	91.69% (2.02%)	91.79% (2.50%)	91.21% (2.34%)
Cellular component						
BLAST	71.81% (8.69%)	67.46% (5.39%)	68.98% (7.05%)	70.76% (7.77%)	67.96% (7.00%)	70.33% (10.75%)
Pfam	82.82% (4.33%)	82.82% (4.33%)	82.82% (4.33%)	82.82% (4.33%)	82.82% (4.33%)	82.82% (4.33%)
SAFPred-nn (T5)	71.44% (12.05%)	82.58% (11.93%)	86.49% (8.77%)	86.49% (9.561%)	86.47% (9.99%)	86.56% (10.77%)
SAFPred-nn	74.37% (12.89%)	83.43% (12.24%)	86.28% (11.03%)	86.31% (11.50%)	86.56% (11.99%)	86.12% (11.86%)
DeepGOPlus	94.66% (1.63%)	95.86% (1.52%)	93.86% (1.86%)	95.23% (1.89%)	94.88% (2.90%)	94.84% (2.82%)
SAFPred-synteny	85.93% (4.14%)	91.17% (3.14%)	90.54% (5.40%)	91.42% (3.94%)	93.50% (2.48%)	94.09% (2.07%)
SAFPred	85.87% (5.42%)	91.01% (3.82%)	92.08% (3.44%)	92.91% (4.08%)	93.89% (2.17%)	94.13% (2.05%)

[1] The number of test proteins annotated with at least one GO term at the threshold which maximizes the F1-score.
[2] Because the Pfam method uses a different training set for predictions, it is not possible to compare the change in the coverage across different train/test pairs.

Table S4.13: Maximum % sequence identity of *Enterococcus* toxin genes to the 11 putative novel toxins predicted by SAFPred.

epx gene	epx gene locustag	predicted gene locustag	maximum ID
epx1	GCA_015300525.1_02780	Ente_haem_BAA-382_V2_01251	38.89
epx1	GCA_002945555.1_00127	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_003319525.1_02660	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_004120285.1_01193	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_004125665.1_01702	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_004125915.1_01060	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_004125935.1_01411	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_004125945.1_01477	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_004125955.1_00929	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_004126025.1_01270	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_005236595.1_02731	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_005237765.1_02590	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_005238535.1_01041	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_015302065.1_02752	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_015333445.1_02601	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_015335045.1_02694	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_015336125.1_02821	Ente_haem_BAA-382_V2_01826	25.76
epx1	GCA_015338255.1_02588	Ente_haem_BAA-382_V2_01826	25.76
epx4	GCA_017356565.1_01452	Ente_mund_ATCC882_V5_00469	60.00
epx4	GCA_015506475.1_02489	Ente_haem_BAA-382_V2_01251	28.85
epx4	GCA_015507715.1_02409	Ente_haem_BAA-382_V2_01251	28.85
epx4	GCA_018089265.1_02226	Ente_haem_BAA-382_V2_01251	28.85

4.5.4 Supplementary Figures

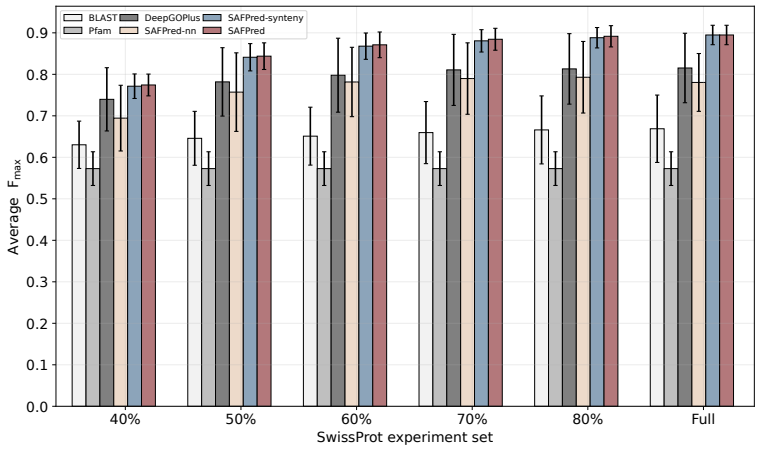


Figure S4.1: average $F_{\max}$ values in Molecular Function Ontology (MFO), across benchmarking datasets for five bacteria in our experimental setup, and the error bars show the corresponding standard deviation of each method, across five species. Note that bar plots for the Pfam baseline are identical for all 6 experiment sets because Pfam uses a different training set, independent of our experimental design.

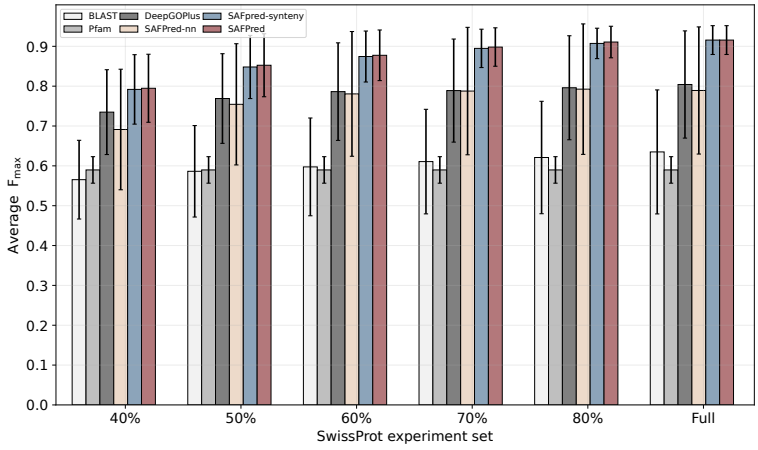


Figure S4.2: average $F_{\max}$ values in Cellular Component Ontology (CCO), across benchmarking datasets for five bacteria in our experimental setup, and the error bars show the corresponding standard deviation of each method, across five species. Note that bar plots for the Pfam baseline are identical for all 6 experiment sets because Pfam uses a different training set, independent of our experimental design.

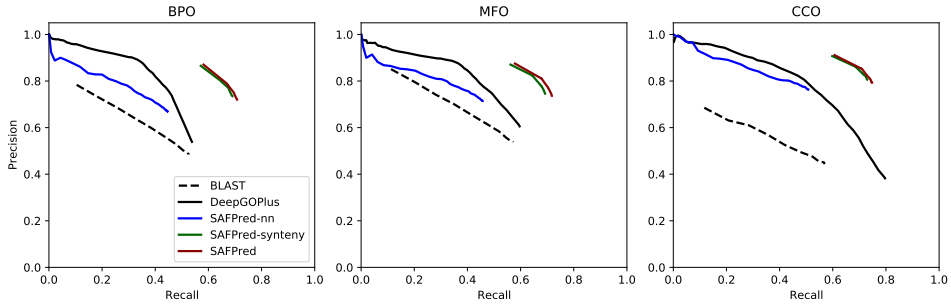


Figure S4.3: Precision recall curves for the *E. coli* dataset where the % sequence ID between the training and the test is limited to 40%.

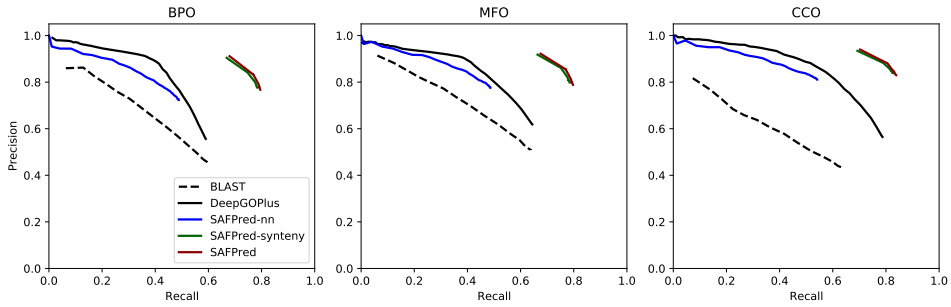


Figure S4.4: Precision recall curves for the *E. coli* dataset where the % sequence ID between the training and the test is limited to 50%.

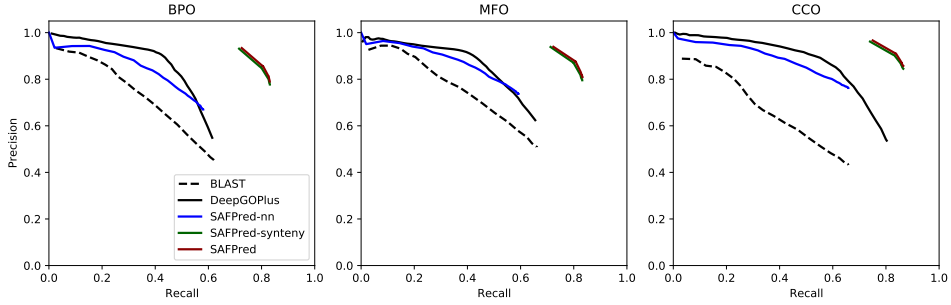


Figure S4.5: Precision recall curves for the *E. coli* dataset where the % sequence ID between the training and the test is limited to 60%.

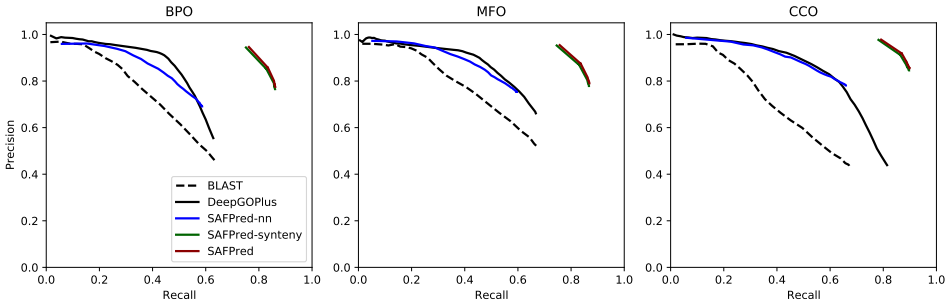


Figure S4.6: Precision recall curves for the *E. coli* dataset where the % sequence ID between the training and the test is limited to 70%.

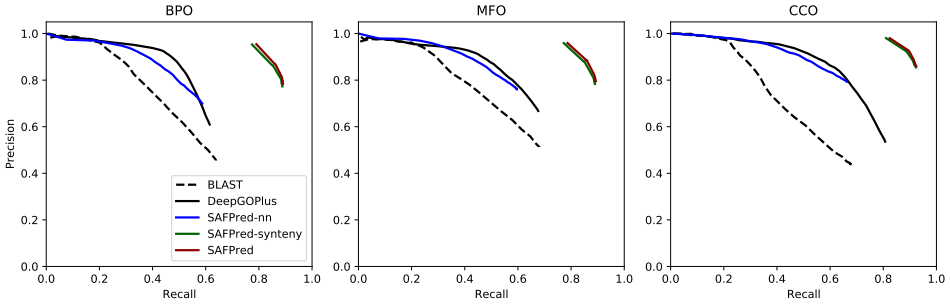


Figure S4.7: Precision recall curves for the *E. coli* dataset where the % sequence ID between the training and the test is limited to 80%.

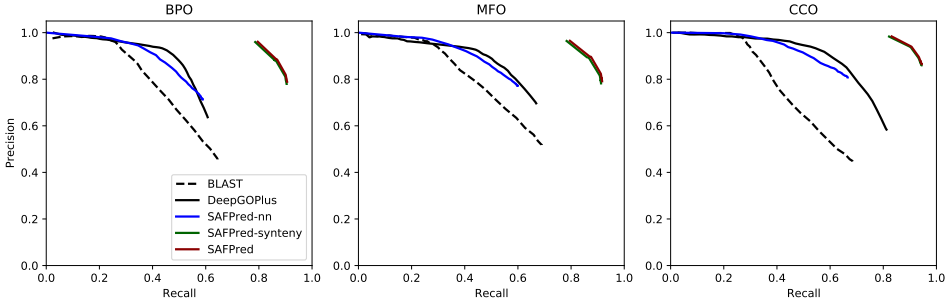


Figure S4.8: Precision recall curves for the *E. coli* dataset where the % sequence ID between the training and the test is limited to 95% (full dataset).

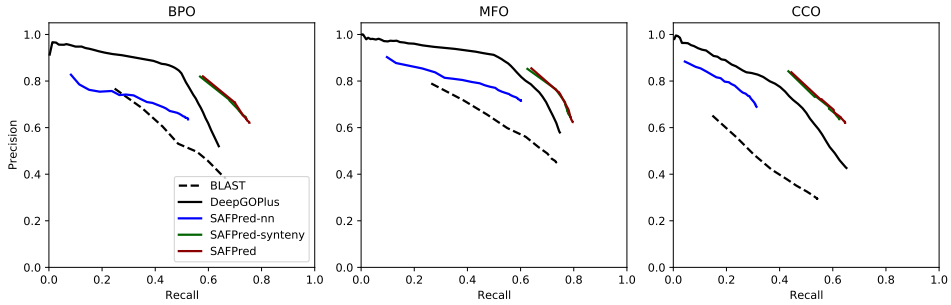


Figure S4.9: Precision recall curves for the *M. tuberculosis* dataset where the % sequence ID between the training and the test is limited to 40%.

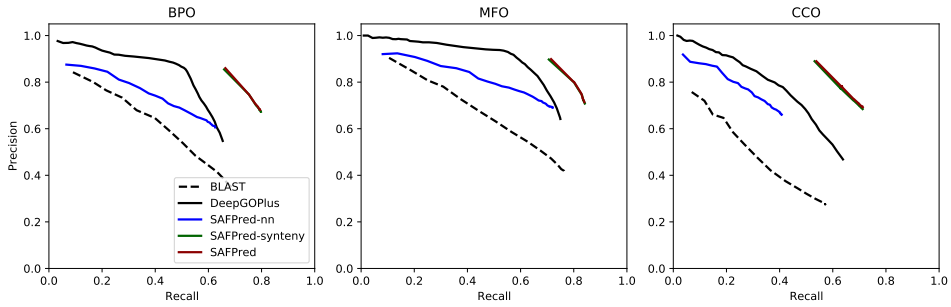


Figure S4.10: Precision recall curves for the *M. tuberculosis* dataset where the % sequence ID between the training and the test is limited to 50%.

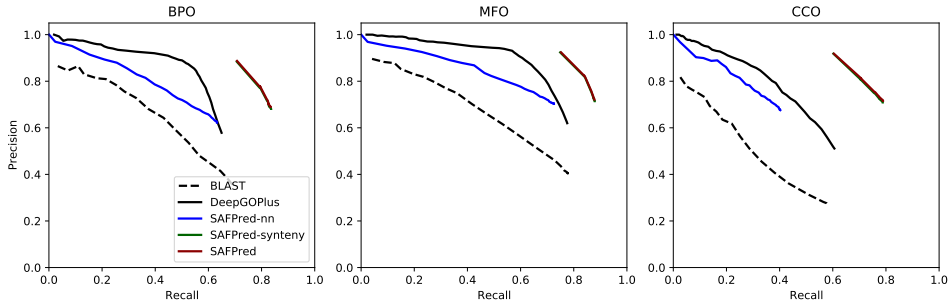


Figure S4.11: Precision recall curves for the *M. tuberculosis* dataset where the % sequence ID between the training and the test is limited to 60%.

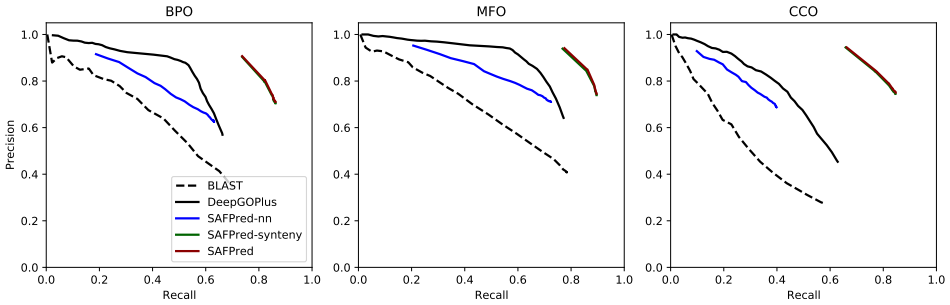


Figure S4.12: Precision recall curves for the *M. tuberculosis* dataset where the % sequence ID between the training and the test is limited to 70%.

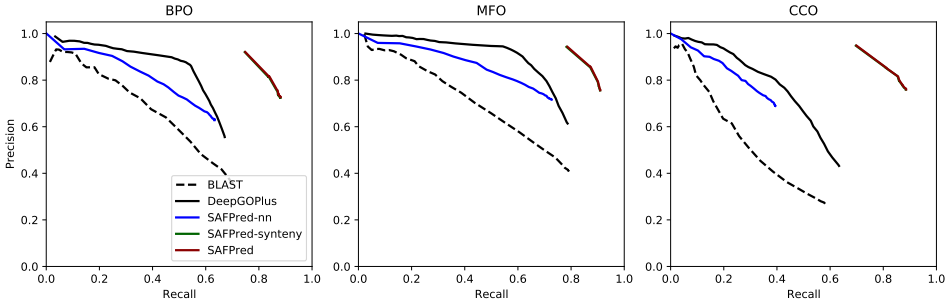


Figure S4.13: Precision recall curves for the *M. tuberculosis* dataset where the % sequence ID between the training and the test is limited to 80%.

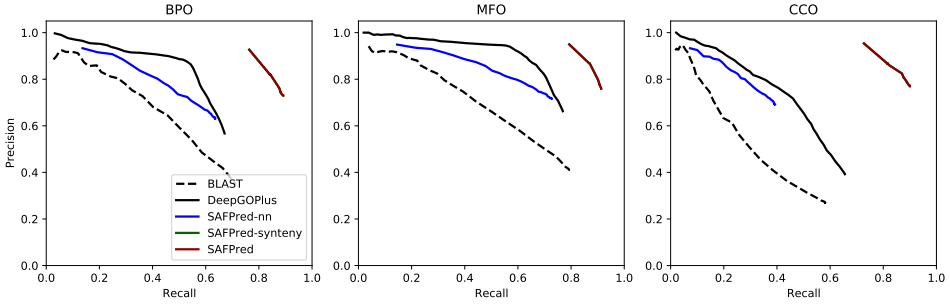


Figure S4.14: Precision recall curves for the *M. tuberculosis* dataset where the % sequence ID between the training and the test is limited to 95% (full dataset).

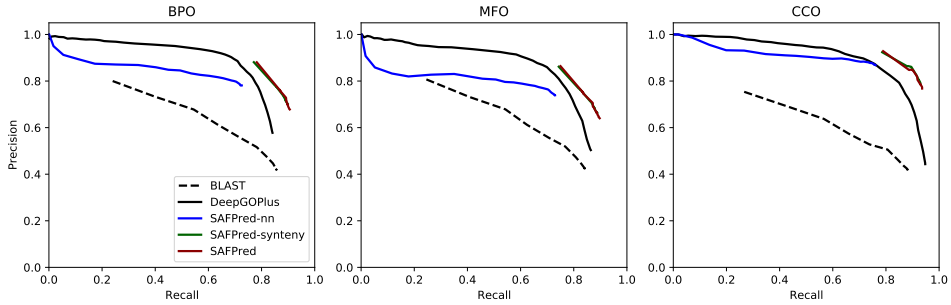


Figure S4.15: Precision recall curves for the *B. subtilis* dataset where the % sequence ID between the training and the test is limited to 40%.

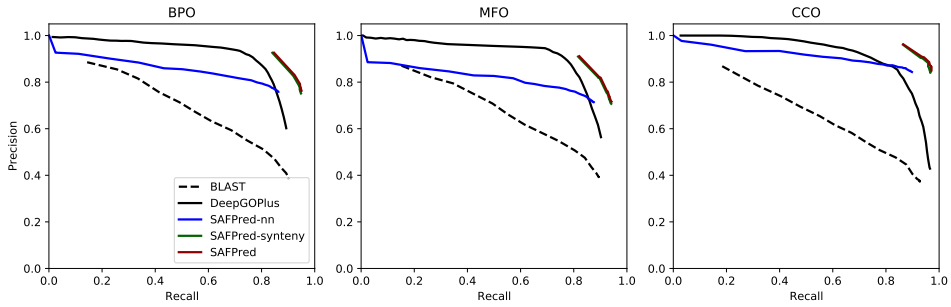


Figure S4.16: Precision recall curves for the *B. subtilis* dataset where the % sequence ID between the training and the test is limited to 50%.

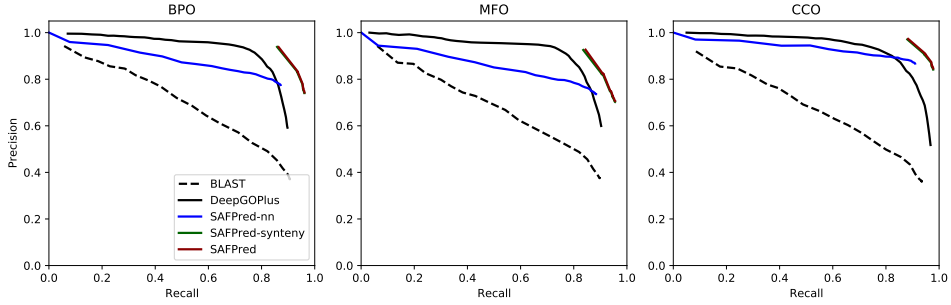


Figure S4.17: Precision recall curves for the *B. subtilis* dataset where the % sequence ID between the training and the test is limited to 60%.

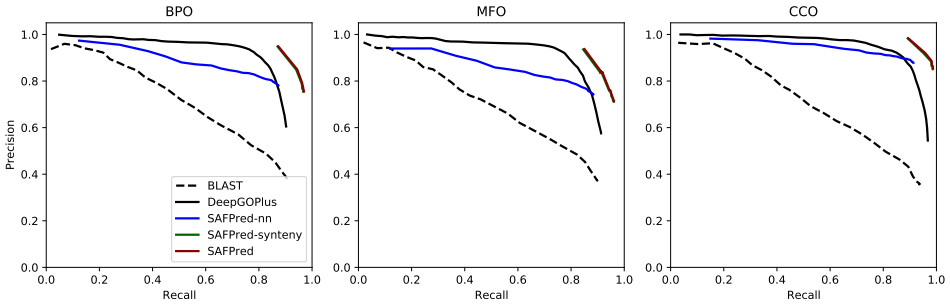


Figure S4.18: Precision recall curves for the *B. subtilis* dataset where the % sequence ID between the training and the test is limited to 70%.

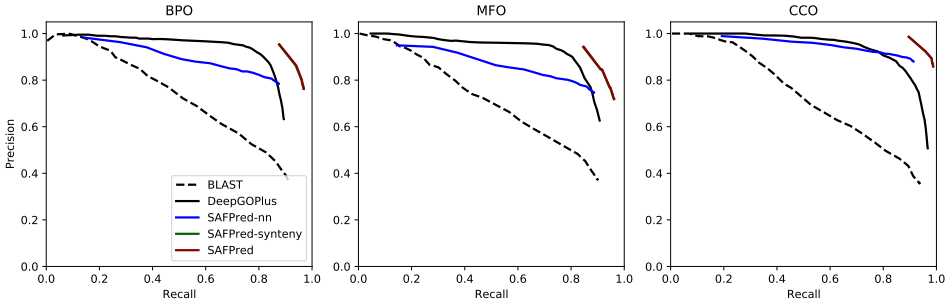


Figure S4.19: Precision recall curves for the *B. subtilis* dataset where the % sequence ID between the training and the test is limited to 80%.

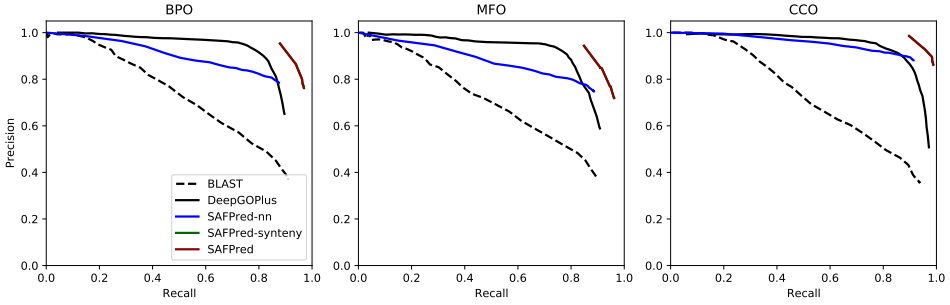


Figure S4.20: Precision recall curves for the *B. subtilis* dataset where the % sequence ID between the training and the test is limited to 95% (full dataset).

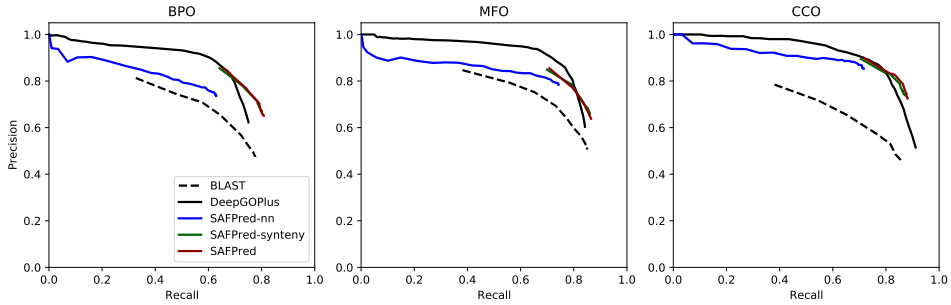


Figure S4.21: Precision recall curves for the *P. aeruginosa* dataset where the % sequence ID between the training and the test is limited to 40%.

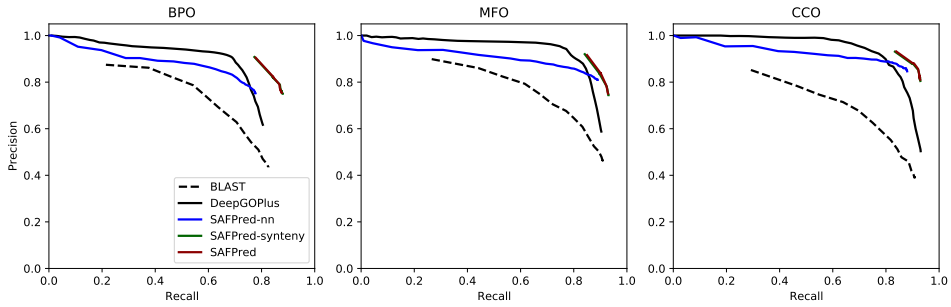


Figure S4.22: Precision recall curves for the *P. aeruginosa* dataset where the % sequence ID between the training and the test is limited to 50%.

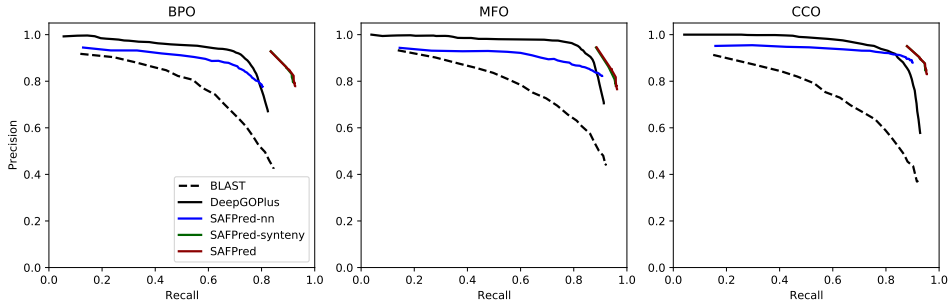


Figure S4.23: Precision recall curves for the *P. aeruginosa* dataset where the % sequence ID between the training and the test is limited to 60%.

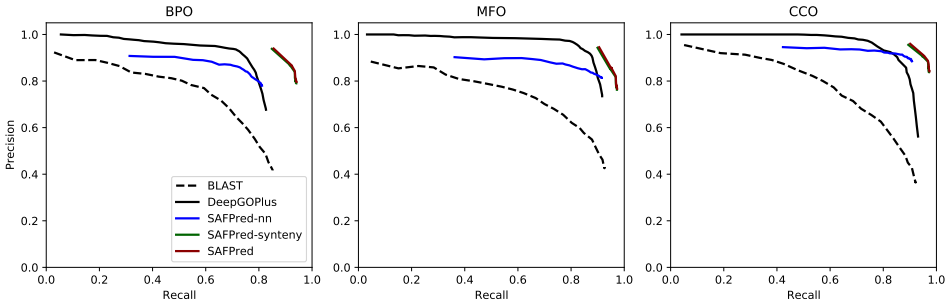


Figure S4.24: Precision recall curves for the *P. aeruginosa* dataset where the % sequence ID between the training and the test is limited to 70%.

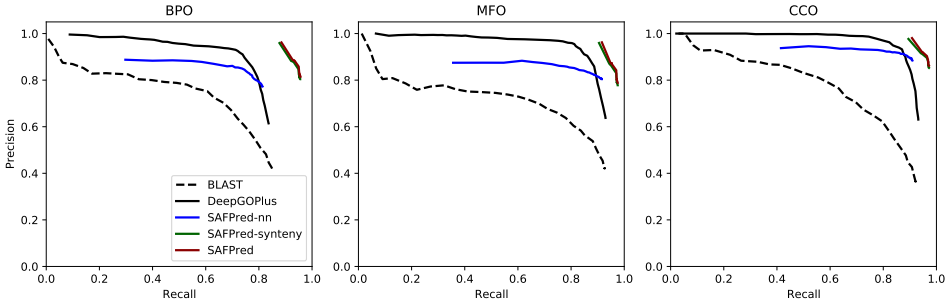


Figure S4.25: Precision recall curves for the *P. aeruginosa* dataset where the % sequence ID between the training and the test is limited to 80%.

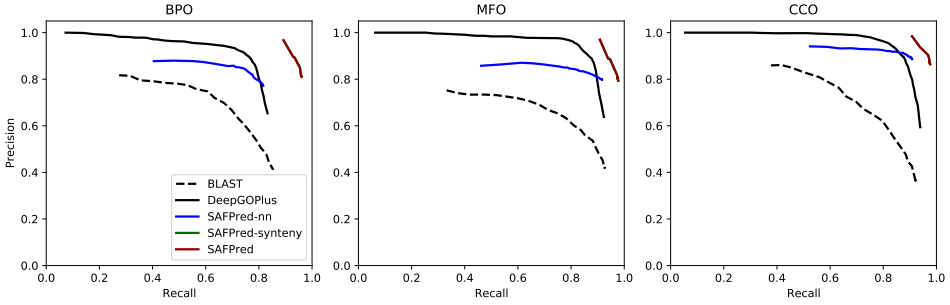


Figure S4.26: Precision recall curves for the *P. aeruginosa* dataset where the % sequence ID between the training and the test is limited to 95% (full dataset).

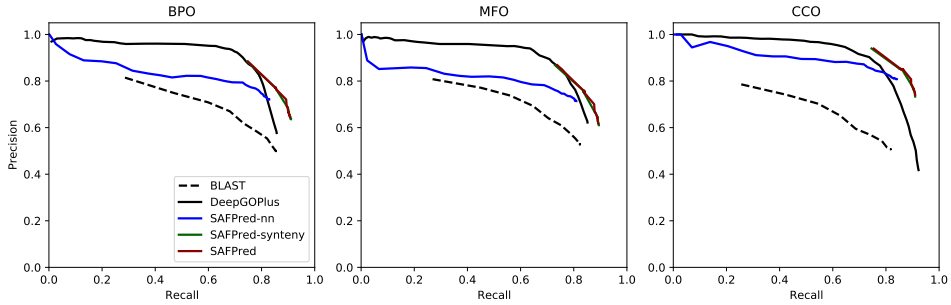


Figure S4.27: Precision recall curves for the *S. typhimurium* dataset where the % sequence ID between the training and the test is limited to 40%.

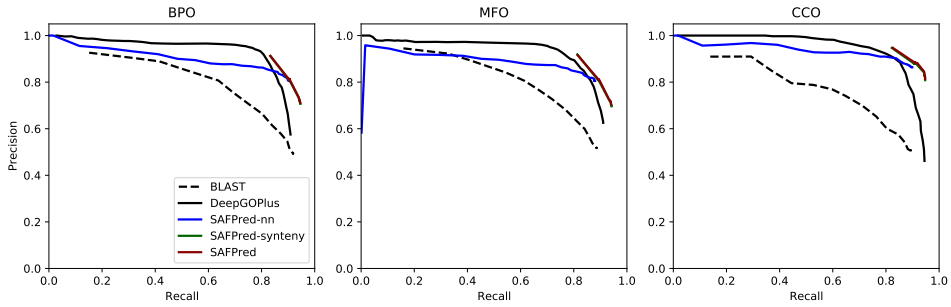


Figure S4.28: Precision recall curves for the *S. typhimurium* dataset where the % sequence ID between the training and the test is limited to 50%.

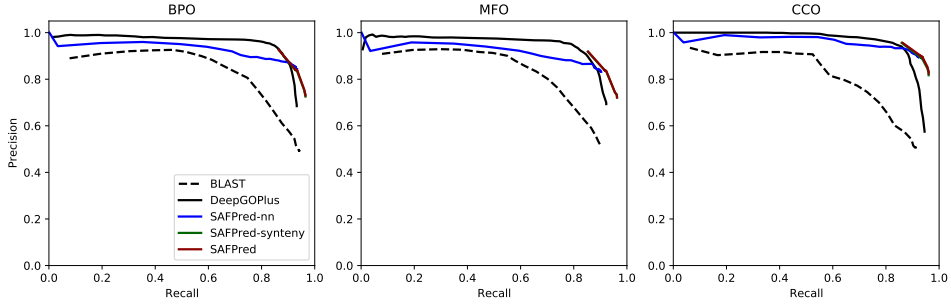


Figure S4.29: Precision recall curves for the *S. typhimurium* dataset where the % sequence ID between the training and the test is limited to 60%.

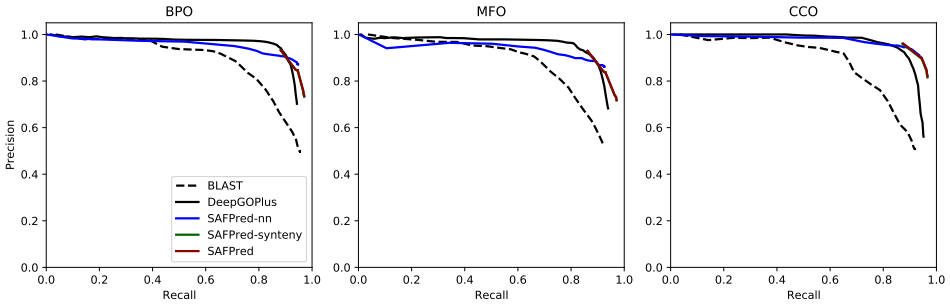


Figure S4.30: Precision recall curves for the *S. typhimurium* dataset where the % sequence ID between the training and the test is limited to 70%.

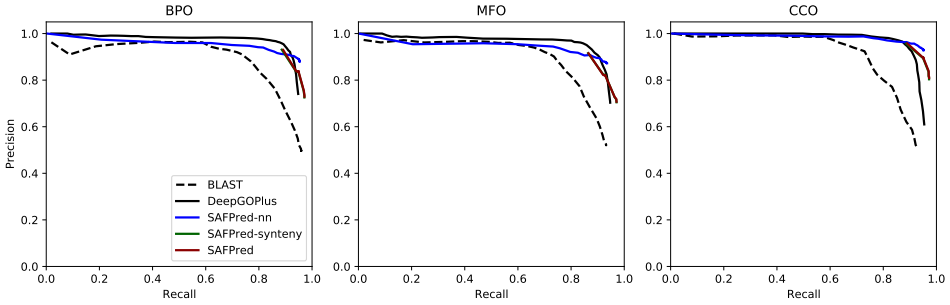


Figure S4.31: Precision recall curves for the *S. typhimurium* dataset where the % sequence ID between the training and the test is limited to 80%.

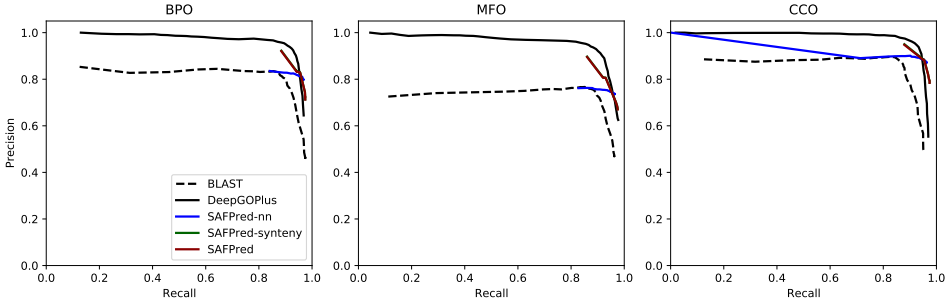


Figure S4.32: Precision recall curves for the *S. typhimurium* dataset where the % sequence ID between the training and the test is limited to 95% (full dataset).

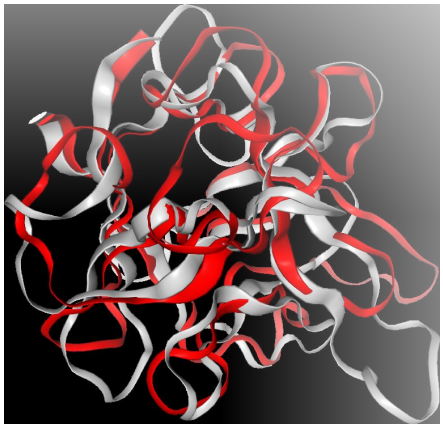


Figure S4.33: SAFPred predicts a possible new variant of pore-forming toxin gene in *E. haemoperoxidus*, distantly related to the known toxin *epx1*, as evidenced by the highly significant structural alignment to the known toxin *epx1*. Gene 0174 (red ribbon, query) had a TM- alignment score of 0.76 when we aligned its structure to *epx1* with Foldseek (gray ribbon, target), despite low sequence identity (5.8%). The protein's sequence alignment is shown below the 3D structural alignment.

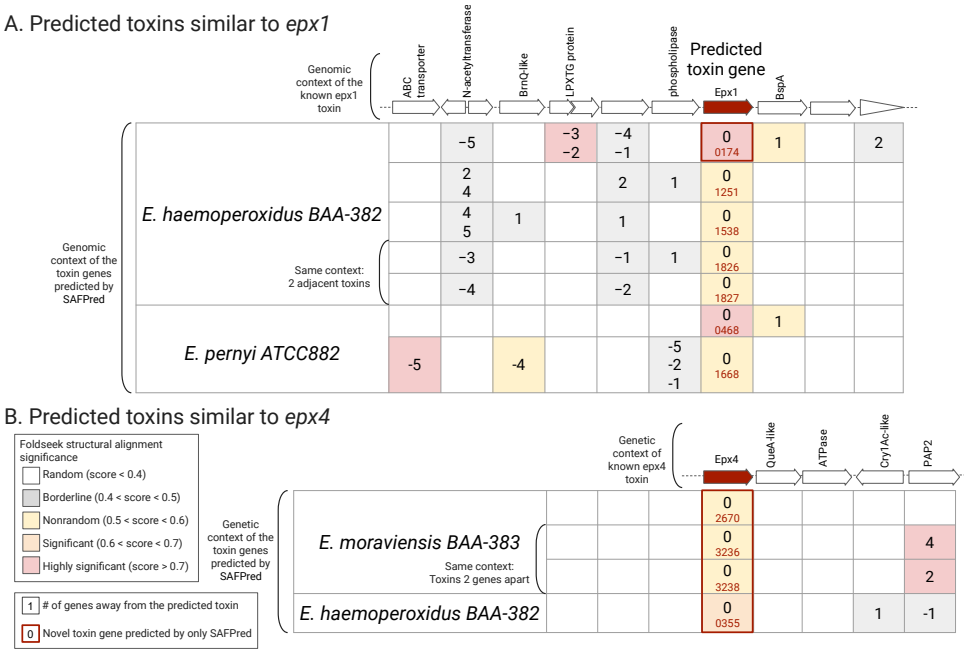


Figure S4.34: SAFPred predicts 11 likely novel toxin genes in our *Enterococcus* dataset (aligned below the column titled "Predicted toxin gene"), including five that could not be predicted by previous function prediction methods (red frames). A) seven genes with the highest similarity to *E. faecalis epx1*. These genes also shared some similarity in genomic context with *epx1*. B) four genes with the highest similarity to *E. hirae epx4*. These genes also shared some similarity in genomic context with *epx4*. The operon diagrams show the known genomic contexts of *epx1* and *exp4* found in *E. faecalis* and *E. hirae* species that were studied, respectively [30]. Beneath this, cells in the table represent the occurrence of genes with structural similarity to those in the known *epx1* or *epx4* operons. Their relative position within the operon in reference to the predicted toxin gene (# of genes away from the predicted toxin) and structural alignment score (coloring) obtained using Foldseek to the analogous gene in the operon diagram is shown. Gene locus tags, a 4-digit number given based on their location within the genome, for the predicted toxins are also placed within the cells. Additional details are shown in Figure S4.35.



Figure S4.35: Detailed view of the genomic neighborhood of 11 likely novel toxin genes (red boxes under epx columns). Each row corresponds to a gene within the genomic neighborhood of the 11 predicted toxin genes. Genes are numbered according to their relative position within the operon in reference to the predicted toxin gene (column titled "# of genes away from the predicted toxin") and the corresponding gene locus tags, a 4-digit number given based on their location within the genome, are also listed in the column titled "Gene locus tag". A) Comparison to the known *epx1* neighborhood. B) Comparison to the known *epx4* neighborhood. The two operon diagrams at the top (adapted from [30]) show the known genomic contexts of *epx1* and *exp4*, while the heatmaps below show protein structural alignment significance for proteins within this genomic context.

References

[1] Aysun Urhan, Bianca-Maria Cosma, Ashlee M Earl, Abigail L Manson, and Thomas Abeel. Safpred: Synteny-aware gene function prediction for bacteria using protein embeddings. *Bioinformatics*, 40(6):btae328, 2024.

[2] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[3] Naihui Zhou, Yuxiang Jiang, Timothy R. Bergquist, Alexandra J. Lee, Balint Z. Kacsoh, Alex W. Crocker, Kimberley A. Lewis, George Georghiou, Huy N. Nguyen, Md Nafiz Hamid, Larry Davis, Tunca Dogan, Volkan Atalay, Ahmet S. Rifaioglu, Alperen Dalkıran, Rengul Cetin Atalay, Chengxin Zhang, Rebecca L. Hurto, Peter L. Freddolino, Yang Zhang, Prajwal Bhat, Fran Supek, José M. Fernández, Branislava Gemovic, Vladimir R. Perovic, Radoslav S. Davidović, Neven Sumonja, Nevena Veljkovic, Ehsaneddin Asgari, Mohammad R. K. Mofrad, Giuseppe Profiti, Castrense Savojardo, Pier Luigi Martelli, Rita Casadio, Florian Boecker, Heiko Schoof, Indika Kahanda, Natalie Thurlby, Alice C. McHardy, Alexandre Renaux, Rabie Saidi, Julian Gough, Alex A. Freitas, Magdalena Antczak, Fabio Fabris, Mark N. Wass, Jie Hou, Jianlin Cheng, Zheng Wang, Alfonso E. Romero, Alberto Paccanaro, Haixuan Yang, Tatyana Goldberg, Chenguang Zhao, Liisa Holm, Petri Törönen, Alan J. Medlar, Elaine Zosa, Itamar Borukhov, Ilya Novikov, Angela Wilkins, Olivier Lichtarge, Po-Han Chi, Wei-Cheng Tseng, Michal Linial, Peter W. Rose, Christophe Dessimoz, Vedrana Vidulin, Saso Dzeroski, Ian Sillitoe, Sayoni Das, Jonathan Gill Lees, David T. Jones, Cen Wan, Domenico Cozzetto, Rui Fa, Mateo Torres, Alex Warwick Vesztrocy, Jose Manuel Rodriguez, Michael L. Tress, Marco Frasca, Marco Notaro, Giuliano Grossi, Alessandro Petrini, Matteo Re, Giorgio Valentini, Marco Mesiti, Daniel B. Roche, Jonas Reeb, David W. Ritchie, Sabeur Aridhi, Seyed Ziaeddin Alborzi, Marie-Dominique Devignes, Da Chen Emily Koo, Richard Bonneau, Vladimir Gligorijević, Meet Barot, Hai Fang, Stefano Toppo, Enrico Lavezzo, Marco Falda, Michele Berselli, Silvio C. E. Tosatto, Marco Carraro, Damiano Piovesan, Hafeez Ur Rehman, Qizhong Mao, Shanshan Zhang, Slobodan Vucetic, Gage S. Black, Dane Jo, Erica Suh, Jonathan B. Dayton, Dallas J. Larsen, Ashton R. Omdahl, Liam J. McGuffin, Danielle A. Brackenridge, Patricia C. Babbitt, Jeffrey M. Yunes, Paolo Fontana, Feng Zhang, Shanfeng Zhu, Ronghui You, Zihan Zhang, Suyang Dai, Shuwei Yao, Weidong Tian, Renzhi Cao, Caleb Chandler, Miguel Amezola, Devon Johnson, Jia-Ming Chang, Wen-Hung Liao, Yi-Wei Liu, Stefano Pascarelli, Yotam Frank, Robert Hoehndorf, Maxat Kulmanov, Imane Boudellioua, Gianfranco Politano, Stefano Di Carlo, Alfredo Benso, Kai Hakala, Filip Ginter, Farrokh Mehryary, Suwisa Kaewphan, Jari Björne, Hans Moen, Martti E. E. Tolvanen, Tapio Salakoski, Daisuke Kihara, Aashish Jain, Tomislav Šmuc, Adrian Altenhoff, Asa Ben-Hur, Burkhard Rost, Steven E. Brenner, Christine A. Orengo, Constance J. Jeffery, Giovanni Bosco, Deborah A. Hogan, Maria J. Martin, Claire O'Donovan, Sean D. Mooney, Casey S. Greene, Predrag Radivojac, and Iddo Friedberg. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20(1):244, 2019.

[4] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[5] Ronghui You, Zihan Zhang, Yi Xiong, Fengzhu Sun, Hiroshi Mamitsuka, and Shanfeng Zhu. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics*, 34(14):2465–2473, 03 2018.

[6] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, 20(1):723, 2019.

[7] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing, 2020.

[8] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[9] A Hoarfrost, A Aptekmann, G Farfañuk, and Y Bromberg. Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nature communications*, 13(1):2606, 2022.

[10] Maria Littmann, Michael Heinzinger, Christian Dallago, Tobias Olenyi, and Burkhard Rost. Embeddings from deep learning transfer go annotations beyond homology. *Scientific reports*, 11(1):1–14, 2021.

[11] Yannick Mahlich, Martin Steinegger, Burkhard Rost, and Yana Bromberg. Hfsp: high speed homology-driven function annotation of proteins. *Bioinformatics*, 34(13):i304–i312, 2018.

[12] Yannick Mahlich, Chengsheng Zhu, Henri Chung, Pavan K Velaga, M Clara De Paolis Kaluza, Predrag Radivojac, Iddo Friedberg, and Yana Bromberg. Learning from the unknown: exploring the range of bacterial functionality. *Nucleic Acids Research*, 51(19):10162–10175, 2023.

[13] Antoine de Daruvar, Julio Collado-Vides, and Alfonso Valencia. Analysis of the cellular functions of escherichia coli operons and their conservation in bacillus subtilis. *Journal of molecular evolution*, 55:211–221, 2002.

[14] Xin Li, Hsinchun Chen, Jiexun Li, and Zhu Zhang. Gene function prediction with gene interaction networks: a context graph kernel approach. *IEEE Transactions on Information Technology in Biomedicine*, 14(1):119–128, 2009.

[15] Stavros Makrodimitris, Marcel Reinders, and Roeland van Ham. A thorough analysis of the contribution of experimental, derived and sequence-based predicted protein-protein interactions for functional annotation of proteins. *Plos one*, 15(11):e0242723, 2020.

[16] Shuwei Yao, Ronghui You, Shaojun Wang, Yi Xiong, Xiaodi Huang, and Shanfeng Zhu. Netgo 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information. *Nucleic acids research*, 49(W1):W469–W475, 2021.

[17] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 11 2018.

[18] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[19] François Lebreton, Abigail L Manson, Jose T Saavedra, Timothy J Straub, Ashlee M Earl, and Michael S Gilmore. Tracing the enterococci from paleozoic origins to the hospital. *Cell*, 169(5):849–861, 2017.

[20] Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. Gtdb: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794, 09 2021.

[21] Shujiro Okuda and Akiyasu C Yoshizawa. Odb: a database for operon organizations, 2011 update. *Nucleic acids research*, 39(suppl_1):D552–D555, 2010.

[22] Sean R Eddy. Accelerated profile hmm searches. *PLoS computational biology*, 7(10):e1002195, 2011.

[23] Alex Mitchell, Hsin-Yu Chang, Louise Daugherty, Matthew Fraser, Sarah Hunter, Rodrigo Lopez, Craig McAnulla, Conor McMenamin, Gift Nuka, Sebastien Pesseat, et al. The interpro protein families database: the classification resource after 15 years. *Nucleic acids research*, 43(D1):D213–D221, 2015.

[24] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. Pfam: The protein families database in 2021. *Nucleic acids research*, 49(D1):D412–D419, 2021.

[25] Maxat Kulmanov and Robert Hoehndorf. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422–429, 07 2019.

[26] Christian Dallago, Konstantin Schütze, Michael Heinzinger, Tobias Olenyi, Maria Littmann, Amy X. Lu, Kevin K. Yang, Seonwoo Min, Sungroh Yoon, James T. Morton, and Burkhard Rost. Learned embeddings from deep learning to visualize and predict protein sets. *Current Protocols*, 1(5):e113, 2021.

[27] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014.

[28] Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, et al. Interpro in 2022. *Nucleic Acids Research*, 51(D1):D418–D427, 2023.

[29] Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, Christian von Mering, and Peer Bork. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309–D314, 11 2018.

[30] Xiaozhe Xiong, Songhai Tian, Pan Yang, Francois Lebreton, Huan Bao, Kuanwei Sheng, Linxiang Yin, Pengsheng Chen, Jie Zhang, Wanshu Qi, Jianbin Ruan, Hao Wu, Hong Chen, David T. Breault, Hao Wu, Ashlee M. Earl, Michael S. Gilmore, Jonathan Abraham, and Min Dong. Emerging enterococcus pore-forming toxins with mhc/hla-i as receptors. *Cell*, 185(7):1157–1171.e22, 2022.

[31] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[32] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[33] Michel van Kempen, Stephanie Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L.M. Gilchrist, Johannes Soeding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *bioRxiv*, 2023.

[34] Jörg Schultz, Frank Milpetz, Peer Bork, and Chris P. Ponting. Smart, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences*, 95(11):5857–5864, 1998.

[35] Predrag Radivojac, Wyatt T. Clark, Tal Ronnen Oron, Alexandra M. Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, Gaurav Pandey, Jeffrey M. Yunes, Ameet S. Talwalkar, Susanna Repo, Michael L. Souza, Damiano Piovesan, Rita Casadio, Zheng Wang, Jianlin Cheng, Hai Fang, Julian Gough, Patrik Koskinen, Petri Törönen, Jussi Nokso-Koivisto, Liisa Holm, Domenico Cozzetto, Daniel W. A. Buchan, Kevin Bryson, David T. Jones, Bhakti Limaye, Harshal Inamdar, Avik Datta, Sunitha K. Manjari, Rajendra Joshi, Meghana Chitale, Daisuke Kihara, Andreas M. Lisewski, Serkan Erdin, Eric Venner, Olivier Lichtarge, Robert Rentzsch, Haixuan Yang, Alfonso E. Romero, Prajwal Bhat, Alberto Paccanaro, Tobias Hamp, Rebecca Kaßner, Stefan Seemayer, Esmeralda Vicedo, Christian Schaefer, Dominik Achten, Florian Auer, Ariane Boehm, Tatjana Braun, Maximilian Hecht, Mark Heron, Peter Hönigschmid, Thomas A. Hopf, Stefanie Kaufmann, Michael

Kiening, Denis Krompass, Cedric Landerer, Yannick Mahlich, Manfred Roos, Jari Björne, Tapio Salakoski, Andrew Wong, Hagit Shatkay, Fanny Gatzmann, Ingolf Sommer, Mark N. Wass, Michael J. E. Sternberg, Nives Škunca, Fran Supek, Matko Bošnjak, Panče Panov, Sašo Džeroski, Tomislav Šmuc, Yiannis A. I. Kourmpetis, Aalt D. J. van Dijk, Cajo J. F. ter Braak, Yuanpeng Zhou, Qingtian Gong, Xinran Dong, Weidong Tian, Marco Falda, Paolo Fontana, Enrico Lavezzo, Barbara Di Camillo, Stefano Toppo, Liang Lan, Nemanja Djuric, Yuhong Guo, Slobodan Vucetic, Amos Bairoch, Michal Linial, Patricia C. Babbitt, Steven E. Brenner, Christine Orengo, Burkhard Rost, Sean D. Mooney, and Iddo Friedberg. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221–227, 2013.

[36] Michael Heinzinger, Maria Littmann, Ian Sillitoe, Nicola Bordin, Christine Orengo, and Burkhard Rost. Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genomics and Bioinformatics*, 4(2), 06 2022. lqac043.

[37] Julia A. Schwartzman, Francois Lebreton, Rauf Salamzade, Melissa J. Martin, Katharina Schaufler, Aysun Urhan, Thomas Abeel, Ilana L.B.C Camargo, Bruna F. Sgardioli, Janira Prichula, Ana Paula Guedes Frazzon, Daria Van Tyne, Gregg Treinish, Charles J. Innis, Jaap A. Wagenaar, Ryan M. Whipple, Abigail L. Manson, Ashlee M. Earl, and Michael S. Gilmore. Global diversity of enterococci and description of 18 novel species. *bioRxiv*, 2023.

[38] Sicai Zhang, Francois Lebreton, Michael J. Mansfield, Shin-Ichiro Miyashita, Jie Zhang, Julia A. Schwartzman, Liang Tao, Geoffrey Masuyer, Markel Martínez-Carranza, Pål Stenmark, Michael S. Gilmore, Andrew C. Doxey, and Min Dong. Identification of a botulinum neurotoxin-like toxin in a commensal strain of enterococcus faecium. *Cell Host & Microbe*, 23(2):169–176.e6, 2018.

[39] William R Pearson. An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, 42(1):3–1, 2013.

[40] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, 25(7):1043–1055, 2015.

5

SAFPedDB: A comprehensive collection of bacterial operons and syntenic regions

"I thought how unpleasant it is to be locked out;
and I thought how it is worse, perhaps, to be locked in."

— Virginia Woolf

Parts of this chapter have been published in biorxiv?

Abstract

Contents of a genome are organized into smaller units of genes and genomic structures. These units are often arranged based on evolutionary and functional constraints. Similar organizational units within genomes are conserved and we observe them across different species and organisms. Usually, conservation, referred to as synteny, is a powerful source of information in studying evolution and inferring functional relationships. In this work, we develop a computational, bottom-up approach to model synteny in bacterial species and we build SAFPedDB. In SAFPedDB, we present an inclusive and extensive catalog of conserved synteny, and we assert its validity as a viable proxy for experimental data as well as its seamless adoption within existing pipelines.

5.1 Introduction

It has been established that genomes of living organisms are subject to several evolutionary and functional constraints that lead to certain genomic structures, and clusters of genes to be commonly conserved across different species. The arrangement and conservation of genomic structures is called synteny and the regions associated with these features are syntenic regions [1]. Especially in bacteria, these clusters are more abundant because they have a compact genome structure, with fewer non-coding regions and intergenic spaces and thus a higher density of genes. In addition, we observe less variation in genome size between closely related bacterial species compared to eukaryotes [2]. Thus, gene-dense bacterial genomes of similar size that are also under strong selective pressures to maintain adaptive fitness are more likely to exhibit synteny [3].

Operons, a specific kind of such syntenic regions, are one of the most prominent features of bacterial genomes [4]. While operons are present in eukaryotes as well, they are not as central to most eukaryotic genomes [5]. Genes within an operon are often co-regulated and co-transcribed and thus they often function in the same metabolic pathway, and carry complementing functions [6]. For that reason, operons can provide valuable information to infer gene function through guilt by association since the genes located on the same operon have functional similarities. Even when a syntenic region is not associated with an operon, they offer us clues on the origins and evolution of genomic structures in bacteria as they are conserved within several species, genera or the entire bacterial kingdom. These features render synteny invaluable to investigate in detail [2].

Recognizing the importance of synteny in evolutionary analyses, several statistical models and algorithms have been proposed to model and describe syntenic regions [7, 8]. In addition, there are many databases established to catalog known bacterial operons, making use of experimental studies [9]. However, such databases are often restricted to specific organisms or strains, and biased toward model organisms that are more widely studied since the databases are limited to experimental data available. While computational approaches can solve a lot of these issues with experimental databases, the best-performing models and algorithms prioritize theory over practicality and thus they have not been adopted as widely as they should be in the field [2].

Although theoretical models are invaluable to understand synteny at a fundamental level, with the exponential increase in the amount of genomic data available, researchers have been more drawn to explore data-driven, machine learning techniques for several

applications in bioinformatics [10, 11]. More recently, deep learning has gained rapid traction when researchers began to employ ideas developed in the field of natural language processing (NLP) [12, 13]. Based on the analogy between human language and genomic data, i.e. the language of life, several methods have been adopted successfully to solve bioinformatics problems. In particular, we observed protein language models (pLM) to be useful as an alternative representation of genomic data [14].

In this work, we developed a purely computational, bottom-up approach to build an extensive database of bacterial operons and syntenic regions, titled SAFPredDB. SAFPredDB has two notable novelties: (i) it is an inclusive and extensive catalog of conserved synteny since it is not restricted to experimental data and (ii) it incorporates state-of-the-art deep learning models to represent the syntenic regions, allowing for quick look-up and seamless adoption within existing bioinformatics pipelines. SAFPredDB serves as a repository for our gene function prediction tool SAFPred, which exploits the bacterial synteny cataloged in SAFPredDB [14].

We demonstrate the validity of our algorithm for building SAFPredDB by first comparing it to existing databases of experimentally determined bacterial operons and then showcasing individual entries to asses its operon predictions. Based on our analyses, we present SAFPredDB to be one of the most up-to-date and inclusive collections of bacterial operons and syntenic regions. SAFPredDB is equipped with metadata that would prove useful in downstream applications to investigate the origins and evolution of a region. SAFPredDB, can be expanded to include more annotations and features, allowing it to be an invaluable tool that can be incorporated within the existing bioinformatics pipelines.

5.2 Building SAFPredDB

We built SAFPredDB, a comprehensive bacterial synteny database through a bottom-up, data-based approach. We start at the level of the gene and move up within the predefined constraints of our operon model to obtain a collection of bacterial synteny regions of varying lengths, and sizes.

5.2.1 Data source

To establish a comprehensive, broad compilation of conserved syntenic regions, the choice of data source is vital. The original genomic data that SAFPedDB is based on was retrieved from the Genome Taxonomy Database (GTDB), which is a collection of all the bacterial representative genomes [15]. We downloaded the entire database (Release 202, retrieved on 31/03/2022), comprising 258,406 genomes in total, 45,555 of which were representative assemblies. For all of these representative genomes, we extracted their protein sequences along with the standardized annotations GTDB provides. To reduce redundancy and the computational cost of downstream processing, we clustered the protein sequences using CD-HIT [16] at 95% sequence identity with default parameters. CD-HIT output was filtered to keep only the clusters that contained at least 10 genes. This step resulted in 372,308 clusters of bacterial genes in total.

5.2.2 Synteny model

To form our database of syntenic regions, we first defined a *synteny model*. Our definition of a synteny region and its constraints in the model were determined based on our survey of known syntenic regions and bacterial operons determined through experiments as well as the operon theory that describes gene regulation in bacteria [9]. According to our model, a region would be classified as syntenic if

1. All genes are located on the same contig.
2. All genes are on the same DNA strand.
3. The maximum distance between any two genes within the region is 2000 bp.
4. The maximum intergenic distance - the distance between two consecutive genes - is 300 bp.

We formed the initial synteny database by grouping gene clusters that satisfy the first three criteria above: initial groups consisted of gene clusters where at least one member of the cluster is located on the same contig and same strand within 2000 bp distance (subdiagram A. in Fig. 5.1). This yielded 1,488,249 such non-singleton *candidate regions*.

Next, to meet the final condition, we iterated over the candidate regions to either remove those with an intergenic distance larger than 300 bp, or split into multiple regions if possible (subdiagram B. in Fig. 5.1).

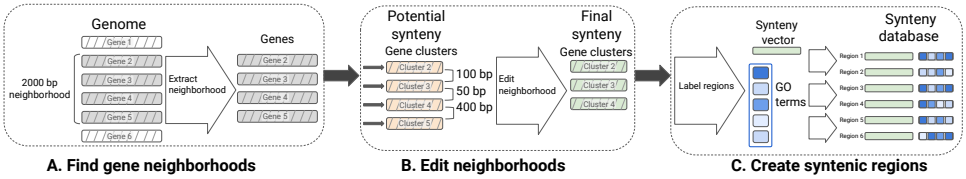


Figure 5.1: Schematic overview of the SAFPredDB construction, from [14]. Hashed boxes represent genes; solid boxes are numerical embedding vectors. A. 2000 bp-long gene neighborhoods are extracted from all genomes in GTDB; shown is an example with four genes in a single genomic neighborhood (hashed grey boxes). B. After clustering all proteins from GTDB with CD-HIT, we replace the genes with the CD-HIT clusters they belong to (hashed orange boxes) using the amino-acid sequence of the representative gene of each cluster in place of their actual amino-acid sequence. Then, we trim potential syntenic regions to remove genes separated by > 300bp, resulting in the final syntenic region (hashed green boxes). C. Once the final operon structures are determined, we i) annotate each region with a set of GO terms, for which we track the corresponding frequency among the gene clusters that make up the region (blue rectangles, darker shades mean GO terms are found in more genes within the syntenic region), and ii) extract numerical embedding vectors for each syntenic region (solid green boxes). We create a new representation for each region, which consists of the average embedding vector and a set of GO terms. The final synteny database is a collection of such representative embedding vectors and GO term frequency vectors; representations of six example entries from the database are shown here.

5.2.3 Vector representation of synteny

In SAFPredDB, we opted for a numerical vector representation of the syntenic regions. The representation is a pair of vectors that contain a numerical representation of the amino-acid sequences as well as a numerical description of the probable function of the region.

The first component, the amino-acid representation is based on the ESM-1b model. ESM-1b is a transformer-based protein language model developed by Rives et al., and it is the same model we used when designing our gene function prediction algorithm, SAFPred [14]. Here in SAFPredDB, we use the same model to stay consistent within our function prediction framework and ensure that our database can be used out of the box without the need to convert vectors or change dimensions. Hence, we followed the same procedure as in SAFPred. To extract amino-acid level embedding vectors, we used `bio_embeddings` (v 0.2.2) (v 0.2.2) [18] with default settings. Then, we obtained protein-level embeddings by taking the average over individual amino-acid embeddings to get 1280x1 dimensional vectors.

The second component, the probable GO-function of the region, is based on the Gene Ontology (GO) database [19]. GTDB includes Prokka annotations for the protein sequences on the database, which are derived from a collection of sequence and HMM-based annotation tools. We re-annotated the sequences by assigning GO terms. In the absence of experimental annotations, we used only the GO terms with the experimental evidence codes: EXP, IDA, IPI, IMP, IGI, IEP, HTP, HDA, HMP, HGI, HEP, IBA, IBD, IKR, IRD, IC, TAS.

We assigned the GO terms based on sequence similarity. To look up similar proteins, we used the non-redundant SwissProt database (release 2021-04, retrieval date 10 November 2021) [20]. We filtered down to include proteins of sequence length [40,1000] and with at least one experimental GO annotation. Consistent with our approach, we selected the experiment codes: EXP, IDA, IPI, IMP, IGI, IEP, HTP, HDA, HMP, HGI, HEP, IBA, IBD, IKR, IRD, IC, TAS. To reduce the redundancy, we clustered the proteins using CD-HIT at 95% sequence similarity. The final SWissProt look-up dataset comprised 107,818 proteins in total.

To calculate pairwise sequence similarity between proteins in SAFPredDB and our non-redundant SwissProt database, we used BLASTp [21]. We transferred GO terms found in significant hits (e-value < 1e-6 and bit score > 50) using the frequency of each GO term among these hits as a predicted score. With this approach, we could assign at least one GO term to 295,446 of the 372,308 clusters of bacterial proteins (79%), which in turn yielded 388,377 non-singleton syntenic regions (out of 406,293; 96%) annotated with at least one GO term.

Thus, the default version of SAFPredDB comprises 406,293 syntenic regions, and each region is an entry with 2 numerical vectors: an embedding vector to represent the aminoacid sequence, and a GO term frequency vector to summarize the probable function of the region.

5.2.4 Further refinement of SAFPredDB

SAFPredDB, at its default state, is an all-around, comprehensive database of bacterial syntenic regions and operons. Depending on the downstream analysis, it is possible to edit the database and further refine it. As an example, here we describe the procedure for removing specific proteins and the synteny entries associated with them.

As part of our benchmarking study to asses SAFPred's prediction performance, we adjusted the SAFPredDB database to be consistent with the corresponding experiment. The goal of the experiment was to assess SAFPred's predictive performance in more challenging

scenarios where there are no homologs of the query point, i.e. the protein sequence whose function we want to predict, in the databases. For that reason, we removed proteins and the regions they were located in if the proteins were homologous to any of the points in the test set.

We used BLASTp to calculate the pairwise sequence identity of each query point to the protein clusters that form SAFPredDB. We removed clusters if they were more similar than a preset threshold of similarity to at least one of the query points in the test set. Since this operation altered the content of syntenic regions (unless they were removed completely), we re-calculated the intergenic distance for the remaining clusters and we split the regions where the intergenic distance exceeded our threshold, 300 bp. This final step is the same as we did when we created the main synteny database (subdiagrams B and C. in Fig. 5.1).

The current version of SAFPredDB is a catalog of broadly conserved syntenic patterns in bacteria; operons or syntenic regions observed rarely in novel species, or uncultured bacteria from metagenomic samples are not included in our database. In addition, considering the exponential growth in genomic data, contents of SAFPredDB will become outdated quickly. SAFPredDB can be reconstructed from newer database releases as well as genomic data other than whole genome assemblies, such as metagenomic sequences. Annotations of SAFPredDB entries can be expanded to include additional functional descriptors or ontologies. Moreover, if the goal is to study a specific genotype, SAFPredDB can be edited to remove unrelated entries to streamline the database. We provide python scripts to rebuild, and edit SAFPredDB to tailor it to any bioinformatics pipeline desired[1].

5.2.5 Assigning query points to syntenic regions

In this section, we briefly summarize how SAFPredDB was used within our gene function method SAFPred, and a detailed explanation of our method SAFPred can be found in the original manuscript [14]. Our approach demonstrates an example procedure to incorporate SAFPredDB into existing bioinformatics pipelines.

SAFPredDB serves as a repository for SAFPred to exploit bacterial synteny for gene function prediction, SAFPred achieves this through a nearest neighbor approach. Within the framework of SAFPred, gene function prediction is a computational task where protein sequences are assigned GO terms. Hence the goal is to figure out which GO terms should be transferred to a query point. SAFPred identifies the nearest neighbors of the query point in the SAFPredDB, and retrieves all syntenic regions the nearest neighbors are located in; these syntenic regions are *assigned* to the query point. The GO term frequency vectors of all assigned regions are extracted and the GO terms are transferred proportional to the similarity of the region to the query point.

5.3 SAFPredDB is a comprehensive collection of bacterial operons and syntenic regions

In this section, we will investigate SAFPredDB as a universal collection of syntenic regions and the validity of our synteny model as well as the bottom-up, purely computational approach we developed. We start by comparing it to known, experimentally determined

[1]You can find the relevant python scripts and documentation at https://github.com/AbeelLab/SAFPredDB

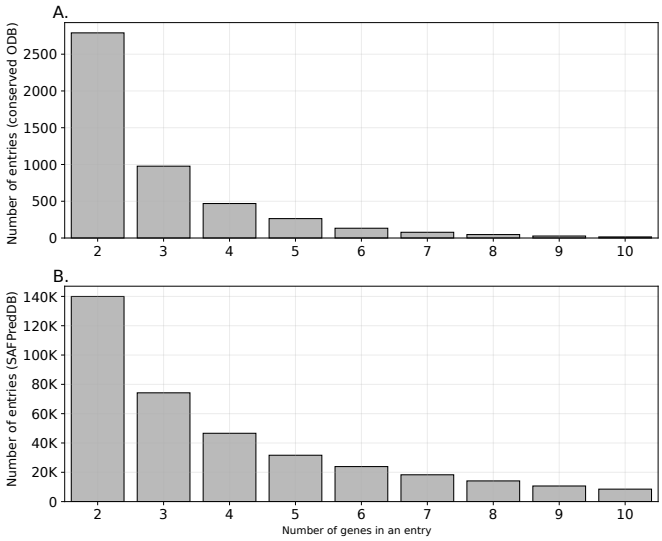


Figure 5.2: The entries in SAFPredDB have a similar distribution of the number of genes in a region to the operons in ODB: A. operons in ODB. B. syntenic regions in SAFPredDB. Only non-singleton regions are shown.

operons found in Operon DataBase (ODB v4) [9]. Then we explore the contents of SAFPredDB in detail and the final step of our algorithm when query points are assigned to syntenic regions.

5.3.1 Synteny model in SAFPredDB approximates experimentally determined operons in ODB

In SAFPredDB, we built a large-scale database that not only approximates the known experimental operons in bacteria, but it provides adequate predictions for potential operons as well as syntenic regions conserved across the bacterial kingdom.

ODB was among the sources we utilized when designing our synteny prediction pipeline, in particular, their conserved operon database [9]. The conserved operon database from ODB, referred to as *ODB conserved* in this text, is essentially an expansion on their known operons where the additional operons were determined from orthologous genes found in multiple genomes that are located consecutively on the same strand of the contig. Our goal in designing our synteny prediction pipeline was to achieve an end product similar to the *ODB conserved* database, but more extensive - representing a broader range of diversity within the bacterial kingdom - and more up-to-date.

We used information from ODB to define our synteny model and the threshold values of the parameters such as operon length, number of genes in an operon, and intergenic distance between adjacent genes in an operon (Synteny model). When we compare SAFPredDB to ODB conserved, we find that our database, is on an aggregate level, quantitatively similar to it. In terms of region length, the number of genes in a region, and the intergenic distance within regions (Figure 5.2-5.4)

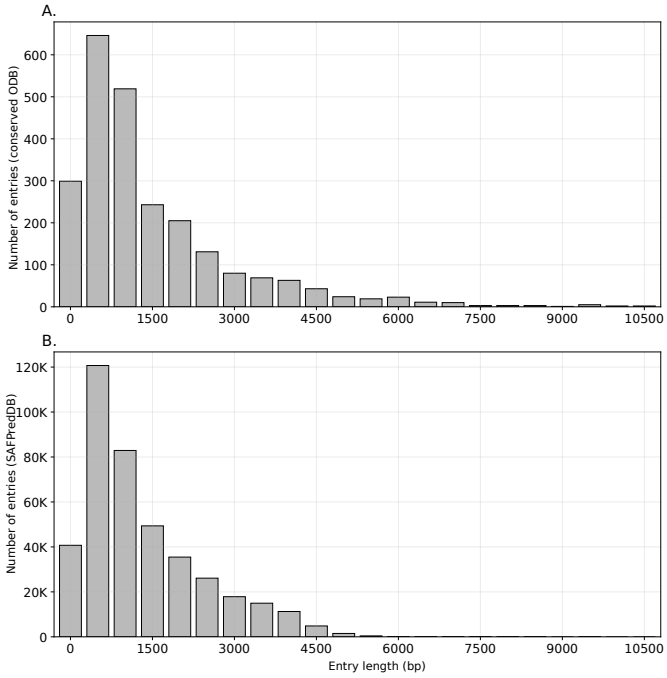


Figure 5.3: The entries in SAFPredDB have similar region length (bp) distribution to the operons in ODB: A. operons in ODB. B. sytenic regions in SAFPredDB. Only non-singleton regions are shown.

We emphasize that our synteny database was not derived from ODB but was built from scratch; as explained previously, we used the CDS of representative bacterial genome assemblies in the GTDB database as the starting point, and we predicted putative operons and identified syntenic regions based on our own synteny model, whereas ODB is a curated list of experimentally determined operons obtained from the literature.

5.3.2 SAFPredDB predicts experimentally determined operons accurately

Our goal in designing our synteny prediction algorithm was to approximate the bacterial operon and synteny landscape using a synteny model derived from known bacterial operons, exploiting large scale genomic data. In this section, we evaluate SAFPredDB as a viable proxy for experimental databases by showcasing its predictions.

We extracted ODB operons containing at least one *E. coli* gene (2845 total operons, including 1071 non-singleton operons). We chose *E. coli* since it is not only one of the most well-studied bacterial organisms in the SwissProt database, but we can identify and cross-reference their genes with the operon entries on ODB because ODB maintains the gene locus tags from the SwissProt database for this organism. However, for the remainder of the organisms in our database, it was not feasible to cross-reference the genes on ODB in a reliable manner.

We compared this to the syntenic regions in SAFPredDB that similarly contain at least

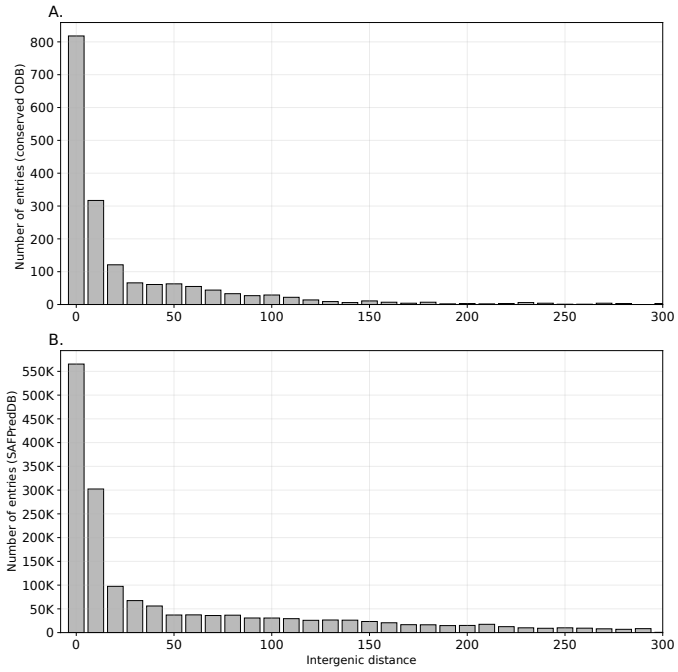


Figure 5.4: The entries in SAFPredDB have similar intergenic distance (bp) distribution to the operons in ODB: A. operons in ODB B. syntenic regions in SAFPredDB. Only non-singleton regions are shown.

one *E. coli* gene (15968 entries in total, including 15,618 non-singleton regions), of which 87% (13,610 entries in SAFPredDB) partially or fully overlap with the known ODB operons. Full concordance is not expected, as our database is substantially larger and encompasses far more genes. In addition, our database is inherently redundant; a single experimentally known operon from ODB can be split into multiple predicted operons and auxiliary regions leading to multiple entries associated with the same operon in our database. Thus, it is highly likely that the remaining 2008 entries (23% of SAFPredDB) we predicted in addition to the ODB operons can be collapsed into a smaller number of actual operons.

While it is not feasible to cross-check each individual operon in our database, here we present a predicted operon which is one of the most commonly found entries among the species in GTDB, demonstrating the accuracy of our prediction at different values in Table 5.1 below. On the right-hand side of Table 5.1 we listed the known operon IDs, names and definitions from ODB, consistent with their nomenclature. And the left-hand of contains the region ID and the relevant metadata extracted from our database. Our database is accompanied by extensive taxonomic metadata from the GTDB database, hence it is a valuable tool to explore the bacterial operon landscape. We can observe which species and lineages an operon is found in (species column in Table 5.1). In addition, our database can capture the wider context of an operon. The *relBEF* toxin-antitoxin system (bottom row in Table 5.1) is one example of such case where we observed that the individual functional unit of the operon, consisting of genes b1562, b1563 and b1564, is often found in multiple entries in our database. Some of these entries were larger than the operon itself, and they contained two IS3 family transposase genes located upstream of the actual operon. Thus, with our database it is possible to extract and analyze the surrounding genomic structures of an operon. Although this can give us additional information about conserved mechanisms related to the operon's mobilization, this can also lead to false positive findings if the goal is to detect only an operon.

5.3.3 SAFPredDB annotations and region assignments

An important feature of SAFPredDB, especially for applications in gene function prediction, is the GO term annotations. It is critical to evaluate the consistency and quality of the annotations to assess SAFPredDB's suitability to study the functional landscape of bacterial operons and syntenic regions. In this section we investigate the GO term annotations, their shortcomings and the potential impact of on the SAFPredDB's use for function prediction.

In the interest of creating a comprehensive collection of bacterial operons and conserved gene context, we used the GTDB database since it is the largest catalog of representative bacterial genomes. However, it does not catalog experimental annotations for the gene sequences. Thus, we assigned function to the entries in SAFPredDB by transferring GO terms from similar protein sequences in the SwissProt database, where only the GO terms with the experimental codes were retained using thresholds on both the pairwise identity and significance (section 5.2.3). When transferring GO terms to SAFPredDB entries, we aimed to balance minimizing false positives while retaining as many annotations as possible, in order to avoid an entry set that was too sparsely annotated to be useful for predicting gene function.

To analyze the annotation procedure, we used the same datasets we generated to benchmark our function prediction tool SAFPred in [14]. In our full SwissProt dataset

Table 5.1: Our operon prediction algorithm can reproduce experimentally known operons in *E. coli*. Three example operons and syntenic regions from SAFPredDB are shown, together with their corresponding operons in ODB, which were manually curated based on experimental studies.

Region ID	Species	*E.coli* gene	Annotation	Operon ID	Name
62671	Escherichia Shigella Citrobacter	b4460	L-arabinose transport system permease protein AraH	KO03244	araFGH
		b1900	Arabinose import ATP-binding protein AraG (EC 7.5.2.12)		
		b1901	L-arabinose-binding periplasmic protein (ABP)		
62674	Escherichia Shigella Bacillus	b1879	Flagellar biosynthetic protein FlhB	KO03228	*flhAB*
		b1880	Flagellar biosynthesis protein *FlhA*		
62755	Citrobacter	-	IS3 family transposase	-	-
	Enterobacter	-	IS3 family transposase	-	-
	Klebsiella	b1562	Toxic protein HokD	KO03197	*relBEF*
	Leclercia Escherichia	b1563	mRNA interferase toxin *RelE*		
		b1564	Antitoxin *RelB*		

experiments, 95.6% of the non-singleton entries (388,377 out of 406,293) were annotated with at least one GO term; however, in remote homology experiments, SAFPredDB was a lot more sparse in annotations (Table 5.2). Since we removed training proteins that had more than a certain predefined level of sequence homology to any of the test proteins in order to create the remote homology datasets, there were only rare proteins with no homologs left in the database. Given that such rare proteins are less likely to be studied, annotated or even carry out any function within the cell, this loss of functional information was expected. Since SAFPredDB is a major source of annotation that SAFPred relies on, SAFPred's prediction coverage also suffers from this loss of information, and we observed that the difference in coverage between SAFPred and DeepGOPlus widens as more annotations are lost (Tables S15 and S16 in [14]).

Furthermore, we note that annotations were transferred at uneven rates for different categories of GO terms, consistent with previous findings in the literature. When sequence similarity is used as the basis for transferring annotation, GO terms in the MFO category are more likely to be transferred. In previous studies, it has been shown that, unlike BPO and CCO, MFO can be modeled using the primary sequence or, features derived from it [22]. We presume this aspect of the GO database affects our findings as well because we also report that the number of syntenic regions annotated with at least one GO term is the largest for the MFO category even though the total number of MFO terms available in the SwissProt dataset is significantly smaller than that of BPO (Table 5.3).

Table 5.2: GO term annotations in SAFPredDB were more sparse in the remote homology detection experiments. Percentage of entries in SAFPredDB that were annotated with at least one GO term for each GO term category and every organism in our SwissProt benchmarks at specific sequence similarity levels (columns).

	40	50	60	70	80	Full
Organism	**Biological process (BPO)**					
E. coli	74.66%	80.23%	82.33%	83.84%	84.16%	84.83%
M. tuberculosis	80.32%	83.55%	84.65%	85.07%	85.21%	85.18%
B. subtilis	79.30%	83.82%	84.63%	85.17%	85.09%	85.15%
P. aeruginosa	79.21%	82.10%	82.79%	83.89%	84.34%	85.16%
S. typhimurium	83.56%	84.44%	84.80%	84.96%	85.03%	85.16%
	Molecular function (MFO)					
E. coli	77.37%	83.08%	85.77%	87.24%	87.84%	88.38%
M. tuberculosis	83.08%	86.40%	87.57%	88.13%	88.38%	88.59%
B. subtilis	81.55%	85.73%	87.51%	88.06%	88.48%	88.58%
P. aeruginosa	82.03%	84.94%	86.35%	87.47%	87.97%	88.59%
S. typhimurium	87.27%	87.90%	88.24%	88.45%	88.49%	88.59%
	Cellular component (CCO)					
E. coli	69.11%	75.94%	79.57%	81.25%	81.94%	82.79%
M. tuberculosis	76.02%	80.00%	81.83%	82.49%	82.82%	83.06%
B. subtilis	75.50%	79.87%	81.98%	82.76%	82.97%	83.06%
P. aeruginosa	75.71%	78.92%	80.86%	81.92%	82.44%	83.05%
S. typhimurium	81.37%	82.26%	82.61%	82.84%	82.91%	83.06%

Table 5.3: Annotation statistics and information content (IC) of syntenic regions in our database. GO terms in MFO category are more likely to be transferred when using sequence homology for annotation transfer.

Annotation statistic	BPO	MFO	CCO
# of annotated regions	268,773	311,424	268,359
Range of # of GO terms	[1, 41]	[1, 17]	[1, 20]
Average # of GO terms per gene in an annotated region	0.556	0.543	0.456
Average IC of GO terms in an annotated region	10.73	8.786	6.022
Average IC of GO terms per gene in an annotated region	3.345	2.711	1.762
Total # of GO terms in the SwissProt database	16281	6308	2565

5.4 Conclusion

In this work, we developed a purely computational, bottom-up approach to model conserved synteny and operon structures in bacterial species. The underlying hypothesis in constructing our own database was that given enough data, it is possible to model the landscape of bacterial operons and syntenic regions accurately enough. This study demonstrates the validity of our approach as we provide in SAFPredDB, a catalog of bacterial operons and syntenic regions that can mimic experimentally determined operons. SAFPredDB is one of the most extensive, and up-to-date collections of bacterial synteny. It serves as a valuable resource to uncover functional patterns, identify potential functional modules or pathways, and infer the putative biological roles of syntenic regions in bacteria.

In addition to being a viable proxy to supplement known, experimentally determined operons, SAFPredDB can be used within various bioinformatics pipelines and assist future studies in bacterial genomics. For instance, it is a valuable tool for analyzing the genomic structure surrounding an operon; equipped with relevant metadata from GTDB, one can gain insight into the region's distribution or its mobilizability across the bacterial kingdom. Analysis of bacterial mobilome is particularly significant in studying antimicrobial resistance in bacteria; several antibiotic resistance genes are found in operons flanked by mobile genetic elements that allow their spread within bacterial populations and cause severe hospital outbreaks.

Furthermore, it can be used to improve gene function prediction in bacteria, especially where sequence similarity is inadequate. Our function prediction tool SAFPred uses SAFPredDB as one of its source databases for predicting gene function [14]. In our benchmark studies of SAFPred, we demonstrate that incorporating synteny with SAFPredDB significantly improves the prediction performance, surpassing not only the conventional annotation tools but also the state of the art in the field of automated function prediction. Thus, SAFPredDB models the landscape of functions encapsulated within bacterial operons and syntenic regions in sufficient detail to be able to leverage this resource to improve bacterial function prediction. It can be an invaluable addition to existing bioinformatics pipelines to annotate new bacterial genomes, identify candidate genes involved in specific biological processes or pathways, and elucidate the functional basis of microbial phenotypes and adaptations.

Having the regions encoded as numerical vectors, it is also faster to look up entries in our database compared to conventional sequence-based databases. Since we use embedding vectors to represent syntenic regions in SAFPredDB, we bypass the need to align the query sequence against the entire database. All operations on our database, such as looking up any entry or calculating similarity are reduced down the vector calculations. This allows for seamless adoption into existing pipelines as well as keeping up to date with recent developments in deep learning.

One limitation of SAFPredDB is its sparse function annotations. Since our genomic data source, GTDB, did not contain experimental gene function annotations, we assigned functions to the entries in SAFPredDB based on sequence similiarity. To minimize false positives in operon annotations, we adopted a conservative approach which in turn resulted in a sparsely annotated training set. We observed the fallout of this procedure when we used SAFPredDB in our gene function prediction tool SAFPred and the sparse annotations led to lower prediction cover [14]. One way to alleviate this problem would be to routinely

pick unlabeled entries from SAFPredDB, prioritizing the most common ones, to perform experiments and identify their functions. With each new experimental annotation available, additional entries can be labeled from guilt by association. We expect this iterative approach to rapidly increase the number of labeled regions available in the database.

Another limitation of the current version of SAFPredDB is its focus on broadly conserved patterns; it represents conserved synteny across the entire bacterial kingdom. Since our goal was to present the most comprehensive database possible that is universally valid, we deliberately designed our algorithm for building SAFPredDB to be inclusive and to cover as many conserved syntenic regions as possible. Thus, patterns or operons associated with rare traits in bacteria, or functional pathways unique to novel species are not present in the default SAFPredDB, but are straightforward to add for specific analyses. We provide python scripts and relevant documentation to expand our database, rebuild it using the newer release of GTDB, or build additional databases from different genomic data sources, such as metagenomic samples. It is also possible to expand the GO term annotations by incorporating additional metadata associated with the syntenic regions.

5

References

[1] Warren C Lathe, Berend Snel, and Peer Bork. Gene context conservation of a higher order than operons. *Trends in biochemical sciences*, 25(10):474–479, 2000. doi: 10.1016/s0968-0004(00)01663-7.

[2] Ivan Junier and Olivier Rivoire. Synteny in bacterial genomes: inference, organization and evolution. *arXiv preprint arXiv:1307.4291*, 2013. doi: 10.48550/arXiv.1307.4291.

[3] Eduardo PC Rocha. The organization of the bacterial genome. *Annual review of genetics*, 42:211–233, 2008. doi: 10.1146/annurev.genet.42.110807.091653.

[4] Antoine de Daruvar, Julio Collado-Vides, and Alfonso Valencia. Analysis of the cellular functions of escherichia coli operons and their conservation in bacillus subtilis. *Journal of molecular evolution*, 55:211–221, 2002. doi: 10.1007/s00239-002-2317-1.

[5] Tao Liu, Hao Luo, and Feng Gao. Position preference of essential genes in prokaryotic operons. *Plos one*, 16(4):e0250380, 2021. doi: 10.1371/journal.pone.0250380.

[6] Marit S Bratlie, Jostein Johansen, and Finn Drabløs. Relationship between operon preference and functional properties of persistent genes in bacterial genomes. *Bmc Genomics*, 11:1–22, 2010. doi: 10.1186/1471-2164-11-71.

[7] Dina Svetlitsky, Tal Dagan, Vered Chalifa-Caspi, and Michal Ziv-Ukelson. Csbfinder: discovery of colinear syntenic blocks across thousands of prokaryotic genomes. *Bioinformatics*, 35(10):1634–1643, 2019. doi: 10.1093/bioinformatics/bty861.

[8] Hagay Enav and Ruth E Ley. Syntracker: a synteny based tool for tracking microbial strains. *BioRxiv*, pages 2021–10, 2021. doi: 10.1101/2021.10.06.463341.

[9] Shujiro Okuda and Akiyasu C Yoshizawa. Odb: a database for operon organizations, 2011 update. *Nucleic acids research*, 39(suppl_1):D552–D555, 2010. doi: 10.1093/nar/gkq1090.

[10] Ronghui You, Zihan Zhang, Yi Xiong, Fengzhu Sun, Hiroshi Mamitsuka, and Shanfeng Zhu. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics*, 34(14):2465–2473, 03 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty130. URL https://doi.org/10.1093/bioinformatics/bty130.

[11] Maxat Kulmanov and Robert Hoehndorf. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422–429, 07 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz595. URL https://doi.org/10.1093/bioinformatics/btz595.

[12] Maria Littmann, Nicola Bordin, Michael Heinzinger, Konstantin Schütze, Christian Dallago, Christine Orengo, and Burkhard Rost. Clustering funfams using sequence embeddings improves ec purity. *Bioinformatics*, 37(20):3449–3455, 05 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab371. URL https://doi.org/10.1093/bioinformatics/btab371.

[13] Michael Heinzinger, Maria Littmann, Ian Sillitoe, Nicola Bordin, Christine Orengo, and Burkhard Rost. Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genomics and Bioinformatics*, 4(2), 06 2022. ISSN 2631-9268. doi: 10.1093/nargab/lqac043. URL https://doi.org/10.1093/nargab/lqac043. lqac043.

[14] Aysun Urhan, Bianca-Maria Cosma, Ashlee M Earl, Abigail L Manson, and Thomas Abeel. Safpred: Synteny-aware gene function prediction for bacteria using protein embeddings. *Bioinformatics*, 40(6):btae328, 2024. doi: 10.1093/bioinformatics/btae328.

[15] Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. Gtdb: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794, 09 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab776. URL https://doi.org/10.1093/nar/gkab776.

[16] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006. doi: 10.1093/bioinformatics/btl158.

[17] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/doi/10.1073/pnas.2016239118.

[18] Christian Dallago, Konstantin Schütze, Michael Heinzinger, Tobias Olenyi, Maria Littmann, Amy X. Lu, Kevin K. Yang, Seonwoo Min, Sungroh Yoon, James T. Morton, and Burkhard Rost. Learned embeddings from deep learning to visualize and predict protein sets. *Current Protocols*, 1(5):e113, 2021. doi: 10.1002/

cpz1.113. URL https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpz1.113.

[19] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000. doi: 10.1038/75556.

[20] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 11 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1049. URL 10.1093/nar/gky1049.

[21] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990. doi: 10.1016/s0022-2836(05)80360-2.

[22] Naihui Zhou, Yuxiang Jiang, Timothy R. Bergquist, Alexandra J. Lee, Balint Z. Kacsoh, Alex W. Crocker, Kimberley A. Lewis, George Georghiou, Huy N. Nguyen, Md Nafiz Hamid, Larry Davis, Tunca Dogan, Volkan Atalay, Ahmet S. Rifaioglu, Alperen Dalkıran, Rengul Cetin Atalay, Chengxin Zhang, Rebecca L. Hurto, Peter L. Freddolino, Yang Zhang, Prajwal Bhat, Fran Supek, José M. Fernández, Branislava Gemovic, Vladimir R. Perovic, Radoslav S. Davidović, Neven Sumonja, Nevena Veljkovic, Ehsaneddin Asgari, Mohammad R. K. Mofrad, Giuseppe Profiti, Castrense Savojardo, Pier Luigi Martelli, Rita Casadio, Florian Boecker, Heiko Schoof, Indika Kahanda, Natalie Thurlby, Alice C. McHardy, Alexandre Renaux, Rabie Saidi, Julian Gough, Alex A. Freitas, Magdalena Antczak, Fabio Fabris, Mark N. Wass, Jie Hou, Jianlin Cheng, Zheng Wang, Alfonso E. Romero, Alberto Paccanaro, Haixuan Yang, Tatyana Goldberg, Chenguang Zhao, Liisa Holm, Petri Törönen, Alan J. Medlar, Elaine Zosa, Itamar Borukhov, Ilya Novikov, Angela Wilkins, Olivier Lichtarge, Po-Han Chi, Wei-Cheng Tseng, Michal Linial, Peter W. Rose, Christophe Dessimoz, Vedrana Vidulin, Saso Dzeroski, Ian Sillitoe, Sayoni Das, Jonathan Gill Lees, David T. Jones, Cen Wan, Domenico Cozzetto, Rui Fa, Mateo Torres, Alex Warwick Vesztrocy, Jose Manuel Rodriguez, Michael L. Tress, Marco Frasca, Marco Notaro, Giuliano Grossi, Alessandro Petrini, Matteo Re, Giorgio Valentini, Marco Mesiti, Daniel B. Roche, Jonas Reeb, David W. Ritchie, Sabeur Aridhi, Seyed Ziaeddin Alborzi, Marie-Dominique Devignes, Da Chen Emily Koo, Richard Bonneau, Vladimir Gligorijević, Meet Barot, Hai Fang, Stefano Toppo, Enrico Lavezzo, Marco Falda, Michele Berselli, Silvio C. E. Tosatto, Marco Carraro, Damiano Piovesan, Hafeez Ur Rehman, Qizhong Mao, Shanshan Zhang, Slobodan Vucetic, Gage S. Black, Dane Jo, Erica Suh, Jonathan B. Dayton, Dallas J. Larsen, Ashton R. Omdahl, Liam J. McGuffin, Danielle A. Brackenridge, Patricia C. Babbitt, Jeffrey M. Yunes, Paolo Fontana, Feng Zhang, Shanfeng Zhu, Ronghui You, Zihan Zhang, Suyang Dai, Shuwei Yao, Weidong Tian, Renzhi Cao, Caleb Chandler, Miguel Amezola, Devon Johnson, Jia-Ming Chang, Wen-Hung Liao, Yi-Wei Liu, Stefano Pascarelli, Yotam Frank, Robert Hoehndorf, Maxat Kulmanov, Imane Boudellioua, Gianfranco Politano, Stefano Di Carlo, Alfredo Benso, Kai Hakala, Filip Ginter, Farrokh Mehryary, Suwisa Kaewphan, Jari Björne, Hans Moen, Martti E. E. Tolvanen, Tapio Salakoski, Daisuke Kihara, Aashish Jain, Tomislav Šmuc, Adrian

Altenhoff, Asa Ben-Hur, Burkhard Rost, Steven E. Brenner, Christine A. Orengo, Constance J. Jeffery, Giovanni Bosco, Deborah A. Hogan, Maria J. Martin, Claire O'Donovan, Sean D. Mooney, Casey S. Greene, Predrag Radivojac, and Iddo Friedberg. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20 (1):244, 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1835-8.

6

Intrinsic and extrinsic antimicrobial resistance in enterococcus genus

"Defence, however, is of much more importance than opulence."

— Adam Smith, *Book IV: On Systems of Political Economy*

Abstract

Enterococci are widespread in the guts of all animals, from insects to mammals. Due to their hardiness and ability to adapt, they evolve with their host and the changing environment, and they can especially survive in hospital settings leading to many hospital-acquired infections. With more than 60 enterococcal species identified, *Enterococcus faecalis* and *Enterococcus faecium* are prevalent in the human microbiome and have become the leading causes of multidrug-resistant hospital-associated infections. Currently, most of the research on enterococci is focused on clinical isolates with human influence, confined to a specific drug class or resistance mechanism, and lacking in technical rigor due to the urgency of studying antimicrobial resistance for medicinal applications and improving current treatments. In this work, we present the largest collection of *Enterococcus* genomes, expanding the number of known species and the composition of clades in the genus. We propose a standardized approach to define intrinsic antibiotic resistance separately from extrinsic and sporadic resistance. Our broad view of the genus reveals that resistance traits are more likely to be clade-specific as opposed to species-specific. In addition, we observed resistance genes follow similar trajectories as they become intrinsic after being acquired by the bacteria at first. Our findings confirm the known intrinsic resistance in enterococci, and expand with additional genes conferring resistance to existing antimicrobial agents as well as species carrying the genes. Finally, we identify genes that could potentially become intrinsic in the future.

6.1 Introduction

Enterococci are lactic acid bacteria with the rare ability to populate the guts of a variety of animals from insects to mammals [1]. They are incredibly resilient to changing environmental conditions; believed to have originated in the guts of arthropods, they have transitioned to land, evolving over hundreds of millions of years to adapt to changing hosts and diets [2]. Enterococci, being opportunistic pathogens that exhibit resistance to several antibiotics currently in use today, are often associated with nosocomial infections [3]. As the number of infections and related hospital outbreaks increase, there is also a growing interest in studying antimicrobial resistance (AMR) in *Enterococcus* to improve treatments [4].

In broad terms, antimicrobial resistance in bacteria is described in three phenotypes: bacteria can be susceptible to an antibiotic, it can exhibit resistance intrinsically or the resistance phenotype can be acquired from the environment, the latter of which is also called extrinsic resistance. Intrinsic resistance is generally defined as a resistance phenotype inherent to all strains of a bacterial species. It is usually mediated by a gene or it can stem from the physical attributes of the bacterial cell. For instance, gram-negative bacteria are intrinsically resistant to conventional macrolides due to the physicochemical properties of their cell wall [5]. Similarly, beta-lactam resistance in enterococci stems from mutations in their penicillin-binding proteins [6].

In addition to beta-lactams, previous studies on *Enterococcus* had established its intrinsic resistance to aminoglycosides, vancomycin, cephalosporins, sulphonamides and lincosamides [7, 8]. Moreover, due to their extensive survival capabilities, *Enterococcus* can easily acquire resistance to several antibiotics, including chloramphenicol, tetracyclines,

erythromycin, rifampicin, ampicillin and glycopeptides [9, 10]. There is a growing number of research in parallel with the increased awareness of AMR, investigating *Enterococcus* outbreaks using new sequencing technologies and applying more sophisticated analysis [11].

Most of these studies, in particular those in clinical settings are concerned with healthcare-associated outbreaks; they focus on the acquired resistance since it is more likely to be transferred to other bacteria in the same environment, leading to massive hospital outbreaks or the emergence of pathogens showcasing multidrug resistance (MDR). While these studies are valuable as a tool to control outbreaks, and minimize their impact on the healthcare system, they are limited in their scope; due to their clinical urgency, the goal of these studies is to address outbreaks immediately and prevent future disasters [12].

Similarly, a large fraction of the existing work on intrinsic resistance is performed to develop new drugs or design new treatments and thus confined to environments that are heavily influenced by human interventions [13, 14]. However, at the most fundamental level, intrinsic resistance predates human use of antibiotics, and it is shaped through the evolutionary constraints imposed on bacteria in the environment, i.e. all other living species and the physical conditions of the environment. Although this view of intrinsic resistance is already known and established, it has not led to the formation of a standard definition for intrinsic resistance.

In addition to the gap in studies about AMR in the absence of human intervention, when we closely examined the current literature in terms of its technical rigor we found there is great potential for improvement. For instance, bioinformatics tools and pipelines adopted in the field can be updated to their latest versions. Similarly, since we have access to more methods today, we can make use of tools and approaches that are tailored to work on bacteria [15, 16]. Similarly, we should promote releasing the source code and describing the technical details of analyses for reproducible research, thereby accelerating the technical advances in future work. Thus, the field of microbiology and medicine can greatly benefit from improving the applications of bioinformatics and computational methods.

In this work, we aim to address the shortcomings of research on AMR traits in *Enterococcus*; we leverage the largest collection of public genome data to conduct a robust, unbiased study of AMR gene patterns in enterococci and provide a starting point to define intrinsic resistance in bacteria. Our study is unique in being inclusive of all genome samples available, and not confined to a specific environmental setting, resistance mechanism, drug class or bacterial species. We establish genotypically clear boundaries between the *Enterococcus* species already known, and we expand the species labels to include 127 species in the genus in total. We provide a broad view of the genus where the four clades established in previous studies are maintained and expanded with the addition of new species. We propose a systematic approach to study intrinsic and acquired resistance in bacteria and demonstrate its use first on our extensive *Enterococcus* dataset. Based on our proposed definitions of AMR genotypes, we show that clade-specific patterns are more prominent than species-specific ones, especially for the species that tend to occupy similar niches. We also observed that AMR genes, regardless of the drug class they confer resistance to or the resistance mechanism, follow similar trajectories from when they are initially acquired from the environment by the bacteria, and to the final point where it becomes intrinsic to the entire species. Finally, to promote reproducible science and best practices

in microbial genomics research, we provide python scripts and relevant documentation to both reproduce our analyses or implement our approach to study other bacterial organisms in Github[1].

6.2 Materials and methods

6.2.1 Enterococci data

To build our enterococci dataset, we retrieved all entries in the Assembly database of NCBI that were labeled as Enterococcus, Melisococcus, Tetragenococcus, Vagococcus, Catelicoccus or Pilibacter, with no restrictions on assembly quality [17]. We expanded the dataset with the in-house collection of 805 assemblies from our collaborators at Broad Institute. The in-house collection was generated for collaborative projects of Broad Institute and Harvard Medical School throughout the years, and it consists of only *Enterococcus* and *Vagococcus* assemblies. Once obtained, the entire dataset went through the same steps regardless of their source of retrieval.

6.2.2 Data preprocessing

We used CheckM (v1.1.3) and PhyloSift (v1.0.0) to calculate several assembly statistics to assess the assembly quality [18, 19]. While the output of these two tools overlap to some extent, they differ in the set of marker genes they use and they provide different metrics, hence they complement one another. We ran CheckM using their marker genes selected specifically for firmicutes, and extracted the "# of scaffolds", "contamination", "completeness" and "N50" statistics from the output file. We downloaded the PhyloSift Reference Marker Genes (v4) which is an HMM database of 37 housekeeping genes that are single copy core to all bacteria. We used HMMER (v3.3.2) to search against the PhyloSift database with e-value cutoff selected for enterococci specifically [20]. Assemblies that carried less than 30 of the PhyloSift housekeeping genes were marked as "incomplete", and the ones that had multiple copies of any gene were flagged as "contaminated"; both incomplete and contaminated assemblies were removed to obtain 18,015 assemblies in total.

To identify the isolation source of assemblies, we adopted an approach similar to Pradier and Bedhomme in [21]. We cross-referenced the assemblies we downloaded from NCBI to collect sample metadata from the NCBI BioSamples database. The in-house assemblies from Broad Institute were accompanied by detailed metadata on the sampling location and source already. Based on these metadata, we categorized the assemblies into three main groups: human, nonhuman and NA. The nonhuman category was further broken down to designate samples collected from animals, environment, or food.

6.2.3 Genome annotation

All 18,015 assemblies in our collection went through the same annotation process. First, we ran Prokka (v1.14.6) [22] with a list of manually curated *Enterococcus*-specific reference genes. Next, we used two additional HMM-based databases to supplement Prokka annotations: the Pfam database (release 32.0) and the KOfam database (release 94.0), which is

[1] You can find the python scripts and relevant documentation on github: https://github.com/aysunrhn/Intrinsic-antimicrobial-resistance/

based on KEGG Orthology [23, 24]. We used HMMer [20] to search the Pfam database, and KofamScan (v1.3.0) [25] for the KOfam database.

6.2.4 Prediction of mobile genetic elements

To account for as many mobile genetic elements (MGEs) as possible, we included plasmids, insertion sequences (IS), transposons and prophage sequences in our list of MGEs.

To predict plasmids, we used three different tools: PlasmidFinder (v2.1.1) [26], MOBsuite (v3.0.3) [27] and DeepMicrobeFinder (v1.0.1) [28]. We used BLAST (v.2.12.0) to search the PlasmidFinder database and retained the significant hits (e-value < 1e-3). We used the `MOB-typer` command in MOBsuite to predict and type plasmids: `MOB-typer` returns replicon family predictions and auxiliary information about the predicted family for every contig, but we extracted only the "mob" field, which states if the contig was classified as "non-mobilizable", "mobilizable" or "conjugative"; we further grouped the "mobilizable" or "conjugative" labels as *mobile*. Finally, we ran DeepMicrobeFinder with four different parameter settings: single mode at lengths 500, 1000 and 2000, and hybrid mode. DeepMicrobeFinder classifies contigs into give classes: eukaryote, eukaryote virus, plasmid, prokaryote and prokaryote virus. The output file contains prediction scores for each class, where the contig is assigned the class with the maximum score. To interpret the scores, we transformed the values through a softmax function, and we retained the label with maximum score only if it was greater than 0.5. We aggregated the predictions from four models by majority voting.

Since we used three different tools, PlasmidFinder, MOBsuite and DeepMicrobeFinder namely, not only employ different algorithms but they also have their own separate databases, we treated their predictions as independent outcomes, and summarized them by majority voting as well. Our *Enterococcus* dataset, however, is made up of assemblies of varying quality because we removed assemblies only if they were severely contaminated or incomplete. Thus, there are many fragmented assemblies with several contigs, which inevitably exacerbate false positive rates of prediction tools. To reduce the false positive rate, we re-labeled contigs predicted to be plasmids as chromosomal if

1. The contig is longer than 500 kbp, or
2. There are more than five genes on the contig and 60% of these genes are core to the *Enterococcus* genus, and the entire assembly contains more than 10 contigs

To identify transposable elements and insertion sequences, we used the TnCentral [29] and ISFinder database (last updated in October 2020) [30], respectively. We searched both databases using BLAST with e-value cutoff 1e-6. And finally, to predict the presence of phage associated sequences in the assemblies, we used ProphET (v0.5.1) [31].

6.2.5 Prediction of antimicrobial resistance genes

We ran rgi (v5.2.0) [32] with the option `-include_loose` to identify antimicrobial resistance genes. Following the guidelines of its developers, we filtered out the resulting hits to remove the nudged hits if the alignment identity to the reference gene sequence was lower than 50%. We then classified the AMR genes as *intrinsic* if (i) the gene was located on a contig we predicted to be nonmobile and (ii) the gene was not associated with any other

mobile element, i.e. there were no transposoble elements, insertion sequences or phage sequences within its 5000bp vicinity.

6.2.6 Assigning species labels and orthogroup clustering

We employed a three-step procedure to identify species labels of all assemblies. First, we started with an initial set of reference genomes used as part of a previous study on *Enterococcus* in [33]. Schwartzman et al. compiled 103 assemblies within the family Enterococcaceae to use in their analyses. We calculated the average nucleotide identity (ANI) of all the assemblies in our dataset to their comparator genomes using FastANI [34] and we used their species labels as our starting point. We then expanded this initial comparator set with assemblies that had ANI values lower than 95%. Next, we calculated all-vs-all mash distance using Mash (v2.3) [35], and we clustered the assemblies using a distance threshold of 0.05, which corresponds to approximately 95% ANI. As a final step, we refined the mash clusters based on the ANI values: we split clusters if there were assemblies with different initial species labels assigned to them.

To cross-check our species assignments and obtain a new, up-to-date phylogeny to describe the Enterococcaceae family we built a phylogenetic tree. We picked representative genomes for each of the species giving priority to the least contaminated and the most complete assemblies, with fewer contigs and the highest N50 value. Only for the species *E. faecalis* and *E. faecium*, we picked 2 representatives. Among the representatives, we removed 2 severely contaminated assemblies (CheckM v1.1.3, contamination value higher than 2) and 3 incomplete assemblies (less than 90% complete). We ran OrthoFinder (v.2.5.4) [36] to find orthologous clusters of genes and identify single copy core (SCC) genes in the representative genomes. We ran IQ-TREE (v.2.1.4 beta) [37] in the `model finder plus` mode on the nucleotide alignment of SCC genes.

6.2.7 Pairwise syntenic distance between gene contexts

In our work, to quantify the difference of two gene contexts, we defined pairwise syntenic distance. Inspired by Teixeira et al., we modified the 2-break distance to obtain a standardized metric, which takes the presence of inserted and deleted genes into account as well. We consider these differences in gene content as an additional penalty on top of the generic 2-break distance. We simply calculate the 2-break distance, which essentially measures the syntenic difference in the shared gene content, and add the number of genes inserted and/or deleted. To standardize, we divide the final value by the total number of unique genes in the two gene contexts.

Following this procedure, if there are no genes shared by the two contexts, the pairwise syntenic distance is 1.0, whereas if there is at least 1 gene in common the distance increases as the number of inserted and/or deleted genes increases. Our distance metric penalizes contexts that share a core set of genes, but have diverged significantly through inserted and/or deleted genes compared to the size of their shared gene content more harshly than completely unrelated contexts that contain entirely different sets of genes. We consider conserved synteny only when the distance is less than 1.0, and a smaller value indicates higher similarity. The minimum value of our distance metric is 0.0 when the two contexts are identical. Figure 6.1 shows pairwise syntenic distance of different gene contexts that can be found on the genome.

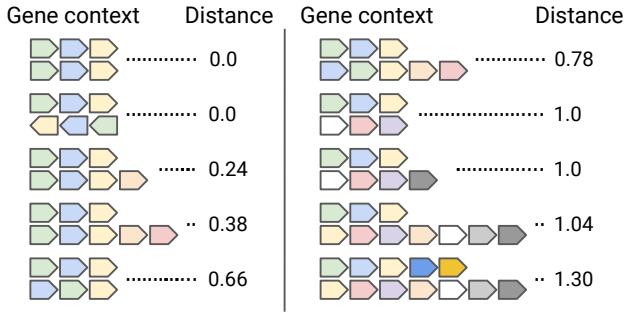


Figure 6.1: Our modified 2-break distance can quantify conservation of synteny between two gene contexts: different combinations of possible gene contexts are shown under "Gene context" along with the corresponding pairwise syntenic distance under "Distance".

6.3 Results and Discussion

6.3.1 Most extensive view of the Enterococcaceae family

As part of our effort to study intrinsic and acquired resistance in *Enterococcus*, we have amassed the largest, most extensive collection of the Enterococcaceae family. To our knowledge, our work is also unique in our unbiased, bottom-up approach to identify species, and perform taxonomic assignment. Thus, our dataset provides the most up-to-date, comprehensive and clear view of the genus *Enterococcus* in the literature.

Due to the clinical importance of enterococcal species and their implications for human health, with 11,700 genomes in total (65% of the entire dataset), most of the assemblies in public repositories are isolated from humans directly, or they were collected in locations with high levels of human activity such as hospitals, laboratories, or locations occupied by humans (ex. hotel rooms ad public restrooms) (Figure 6.2). The largest fraction of nonhuman samples (65%) were associated with animals, as expected. Enterococci are commonly found in the guts of most land animals as part of the microbiome. Since enterococci are a type of lactic acid bacteria, they also play an important role in fermentation, and they are present in several fermented food products, such as dairy, olives and cured meat. Thus, the remainder of the nonhuman samples in our dataset are either collected from environmental settings or food products.

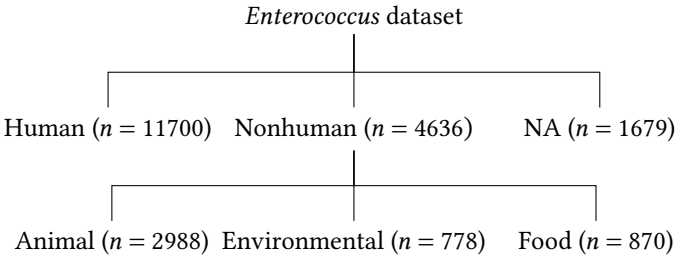


Figure 6.2: Breakdown of assemblies in the *Enterococcus* dataset based on their isolation source.

Table 6.1: The most frequently found species in our *Enterococcaceae* dataset: we have more than 25 genome assemblies for each species.

Species	Count
E. faecium clade A	11095
E. faecalis OG1RF	4243
E. faecalis V583-GB	703
E. hirae	390
E. faecium clade B	380
E. lactis	232
E. casseliflavus	139
E. durans	132
E. gallinarum	74
E. mundtii	57
E. pernyi	56
E. cecorum	55
E. avium	43
E. thailandicus	28

To obtain robust species labels, we started with the comparator set of 103 genomes from Schwartzman et al., and we added assemblies in our dataset that were released after their study was published and that had ANI less than 95% to any of their genomes to obtain 137 species. Guided by this initial list of 137 species, we performed hierarchical clustering on the entire dataset to obtain 115 clusters using the conventional species cutoff value of 0.05 mash distance. Mash distance is a rough approximation of ANI, and using it as a basis led to underclustering since we observed multiple genomes from the initial comparator set grouped into a single cluster. We refined mash clusters by breaking them up, and forming new, smaller clusters based on the ANI values. Ultimately, we had 129 unique species, including *Enterococcus*, *Vagococcus*, *Tetragenococcus* and *Melissococcus*.

From the 129 species, we found 5 species with assemblies of low quality (either incomplete or severely contaminated) and there was only 1 instance of each in our dataset, which suggests issues in either sequencing or assembly. Thus, we removed these 5 species before proceeding with analyses. For the remaining 124 species, we picked one representative with the exceptions of *E. faecium* and *E. faecalis* where we retained 2 species for both to differentiate their commensal subspecies (*E. faecalis* OG1RF and *E. faecium* clade B) from the hospital-associated one (*E. faecalis* V583-GB and *E. faecium* clade A). To ensure robustness and reliability, we prioritized higher-quality assemblies, and we obtained 126 genomes to represent the entire Enterococcacea family.

For the whole family, we found 11,433 orthogroups in total, 120 of which were single copy core (SCC). When we considered the genus *Enterococcus* only, there were 10,879 orthogroups in total where 438 were core and 198 were SCC, consistent with the previous studies [7, 33]. Based on the nucleotide alignment of SCC genes, we built a SCC phylogenetic tree, which revealed a structure with 5 distinct clades: 4 for *Enterococcus* (blue, yellow, red and green clades in Figure 6.3) and 1 for *Vagococcus* (gray clade in Figure 6.3), along with 4

outgroup species. The outgroup species included *Catelicoccus* and *Pilibacter*, as well as 2 species that were most likely mislabeled as *Enterococcus* on NCBI. All major branches in the final phylogenetic tree have bootstrap values larger than 95%. Thus, with this final view of the Enterococcaea family, we have not only expanded the known phylogeny significantly but also produced a high-resolution, robust representation of the family that will prove a valuable resource for future studies (Figure 6.3).

6.3.2 Intrinsic resistance patterns in the *Enterococcus* phylogeny.

To study the intrinsic and acquired antibiotic resistance as two distinct features in enterococci, we developed a systematic framework to define and assess intrinsic AMR. We build on top of Miller et al.'s review of AMR and Hollenbeck and Rice's study. Given that our extensive collection of *Enterococcus* genomes covers the entire sequencing effort available publicly, we are justified in using our representative set as a valid sample of the population. In our work, we recommended the following conditions that define intrinsic resistance:

1. The AMR gene is present in at least 85% of the instances of a species. For the over-represented species, or if there are clear, evolutionarily supported sub-speciation within the species then this threshold should be applied at the subspecies level.
2. At least 85% of the instance of the gene is found on the chromosome, and it is not associated with any mobile element.
3. The genomic context of the AMR gene is largely conserved across the genomes.
4. Gene tree correlates with the species phylogeny.

Note that the final point on the correlation of trees, while adding strong evidence to the assessment, is a challenging one to measure. Although there have been many attempts at quantifying pairwise distance between phylogenetic trees, there is no clear consensus on which metric to use. In addition, the choice of the metric also determines how the measurements should be interpreted, if it is all possible to interpret. In our work, we suggest using a combination of visual checks and tree distance metrics: tanglegrams are useful for small trees where the leaf sets are consistent, and multiple metrics should be calculated to assess the difference between trees [39]. It should be up to the user to decide whether the findings from these two approaches are strong enough to conclude such a correlation between the trees.

Acquired resistance, on the other hand, is associated with genes located on a plasmid, transposon or near a mobile element such as an IS. Lastly, we define sporadic resistance as AMR genes carried on the chromosome, not associated with any mobile elements, but also do not meet the threshold of intrinsic resistance.

All the AMR genes present in *Enterococcus* belong to only 4 of the gene families designated in the CARD database based on the resistance mechanism: antibiotic efflux, inactivation, target alteration and target protection. In addition, we found most of the AMR genes to be either exclusively on the chromosome or a mobile element. It is very rare to see an AMR gene on both, except for the two most studied species *E. faecium* and *E. faecalis*. These two species, often considered separately in our work, are unique in the way they can easily acquire new AMR genes and develop resistance [4].

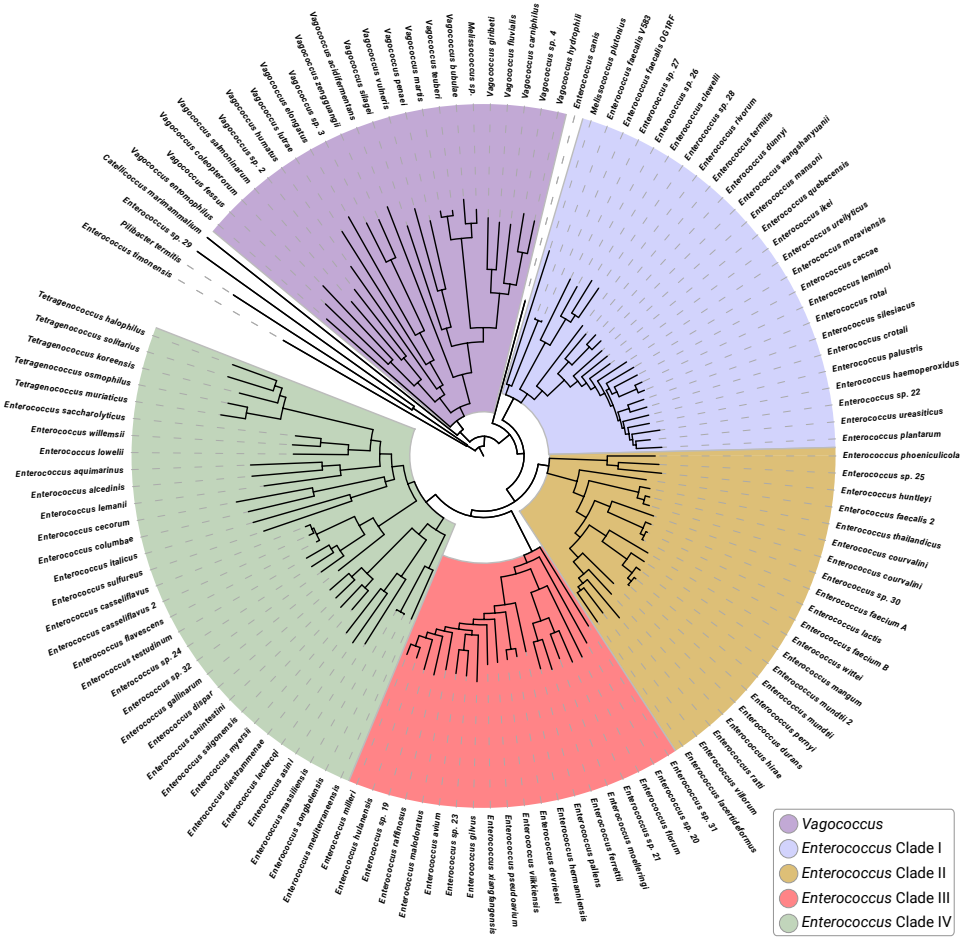


Figure 6.3: A single copy core phylogenetic tree of representative *Enterococcus* genomes. In our work, we expand the known Enterococcaea family phylogeny, and we provide a robust view of the clades with higher resolution.

6.3.3 Clade-specific AMR patterns are more prevalent than species-specific patterns

At the genus level, clade-specific patterns appeared more prominently than species-specific ones. For instance, one of the defining characteristics of *Enterococcus* as a genus is the presence of antibiotic efflux resistance mechanism through the AMR gene *efrA*. As we expected, *efrA* is absent in *Vagococcus*, and other outgroup species (Figure S6.1), as well as *Tetragenococcus* although the latter is branching in our tree within clade IV of *Enterococcus*.

However, we found clade III species to be lacking not only *efrA* but also any AMR gene with this mechanism. Making up for the lack of efflux mechanism, some clade III species have developed alternative means of resistance. For instance, *E. raffinosus* and *E. avium*, two species that are most commonly isolated from hospital infections, can readily acquire resistance genes present in the environment. However, there are no reports of these species displaying intrinsic resistance [40].

Apart from these two species, members of clade III in general are the least likely to carry any AMR gene, despite having a larger genome size on average (Table 6.2). While it might be a trait specific to this clade, it could also be an artifact of this clade being undersampled compared to the rest of the genus. Although our dataset is well balanced in terms of the number of species per clade, we have fewer assemblies of clade III species in total and the average number of genomes per species is the lowest for this clade (Table 6.2).

Table 6.2: Statistics of four *Enterococcus* clades in our genus dataset.

Clade	No. of species	No. of genomes per species	No. of genomes per species (excl. *E. faecalis* and *E. faecium*)	Mean genome size (bp)
I	26	120.8	3.1	2982827
II	21	581.3	44.6	2984041
III	19	4.6	4.6	4252120
IV	32	11.1	11.1	3145746

In addition to clade III species, we found 3 species in clade IV, *E. cecorum*, *E. columbae* and *E. aquamarinus*, that do not carry the *efrA* gene. In the literature, these species are considered to be *specialized* since they are found in very specific environments, similar to *Tetragenococcus* species which are their immediate neighbors on the tree. These species, being specialized to their niche, are unique in the genus.

Finally, we note that both clade II and clade IV carry several resistance genes with antibiotic inactivation mechanism, however the specific AMR genes differed between these two clades. In clade II, the intrinsic resistance was due to chromosomal-encoded AAC(6') and APH(3") gene families, whereas clade IV carried genes from the lincosamide nucleotidyltransferase (LNU) and chloramphenicol acetyltransferase (CAT) gene families which confer resistance to lincosamides and phenicols, respectively (Figure 6.5).

Among all species, we found *E. faecalis* to carry AMR genes not only intrinsic but also unique to the species, namely, *emeA*, *efrB* and *lsaA* genes, all of which are efflux-associated (Figure S6.1). Together with *efrA*, *efrB* forms the efrAB multidrug ATP binding cassette (ABC), and in our dataset, almost all *E. faecalis* genomes (3056 out of 3068) carried it on the chromosome. In previous studies, the efrAB cassette and the *lsaA* genes were reported

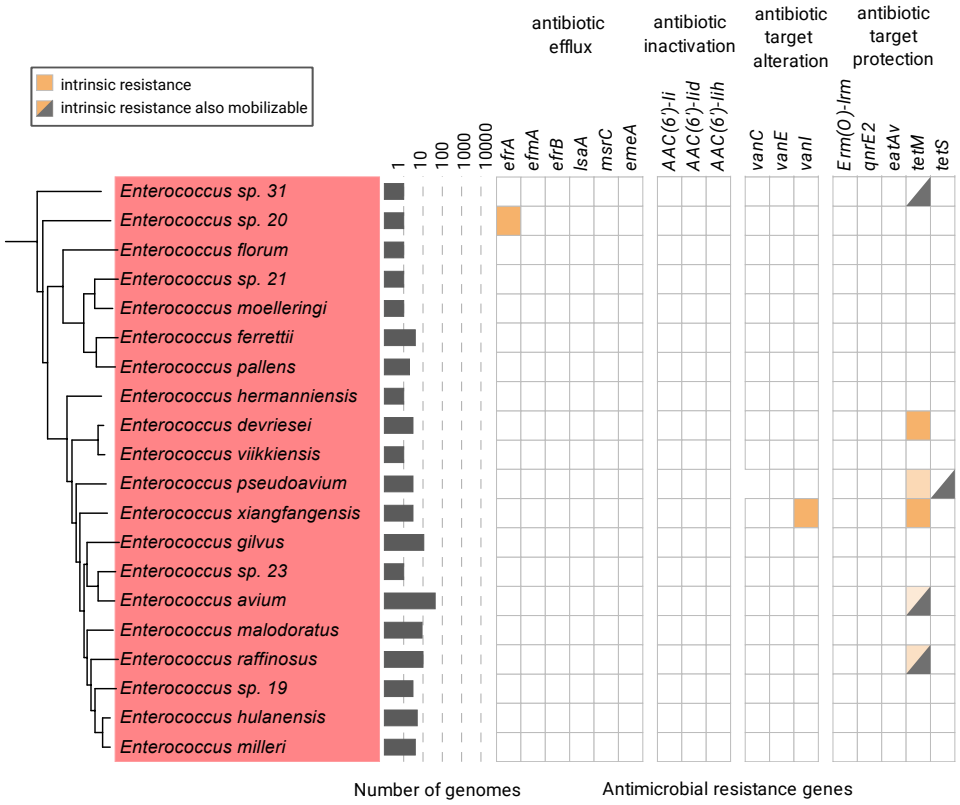


Figure 6.4: Clade III is unique in the *Enterococcus* for their lack of intrinsic AMR, in particular the *Enterococcus* core gene, efrA.

to be intrinsic to *E. faecalis*, while *E. faecium* carried a different intrinsic efflux pump that includes the gene *msrC* [10]. Our work also supports their findings, and we also suggest that the efrAB pump is unique to *E. faecalis*, as we have not detected it intrinsically in any other enterococci.

6.3.4 Acquired resistance genes follow similar trajectories as they become intrinsic

Regardless of the gene family, AMR mechanism or the species, we found common patterns among AMR genes in our dataset. Once acquired, an AMR gene goes through an evolutionary trajectory that is preserved among a few notable examples. We investigate the evolution of AMR genes in detail by observing the phylogenetic gene trees. We found evolutionary patterns common to multiple genes, these patterns suggest genes are in different stages of becoming intrinsic.

AAC(6')-Ii gene conferring resistance to aminoglycosides was initially reported to be chromosomally encoded in *E. faecium A* only [41]. In our work, we confirm this, and we also report three other species that are immediate neighbors to *E. faecium A*; *E. thailandicus*,

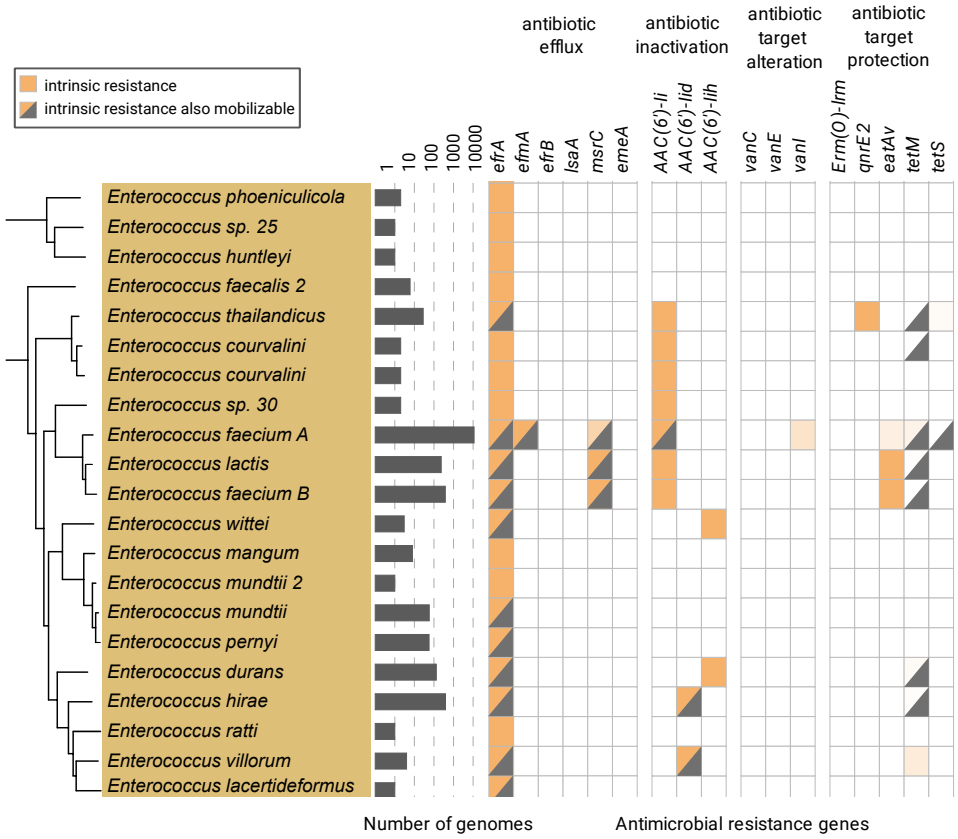


Figure 6.5: Intrinsic aminoglycoside resistance is the main source of AMR with antibiotic inactivation mechanism in clade II

E. courvalini, and an un-named species *E. sp. 30* (Genbank accesion GCA_011038845.1) to be intrinsically resistant to aminoglycosides (Figure 6.6). In addition to carrying *AAC(6')-Ii* gene on a chromosomal contig, we found these genes within a conserved genomic neighborhood of approximately six genes. The *AAC(6')-Ii* gene tree is in agreement with the species tree, and the gene is most commonly observed within the gene context numbered 6 in Figure 6.7 found on the chromosome. The chromosomal *AAC(6')-Ii* context is conserved among the neighboring species we deemed to be intrinsically resistant. *E. faecium A*, the pathogenic subspecies, can readily acquire and transfer resistance genes, and we also found some of the *E. faecium A* genomes in our dataset to carry *AAC(6')-Ii* gene within the genomic context 1 that is associated with mobile elements.

In addition to *AAC(6')-Ii*, intrinsic aminoglycoside resistance in *Enterococcus* exists through two other variants of the gene *AAC(6')-Ii* in our dataset that we observed to be occasionally mobilizable. *AAC(6')-Iid* is intrinsic to both *E. durans* and *E. wittei*, within similar gene contexts (contexts 2 and 3 in Figure 6.6. *AAC(6')-Iih* is intrinsic to *E. hirae* and *E. villorum*, although we also report instances associated with mobile elements in both

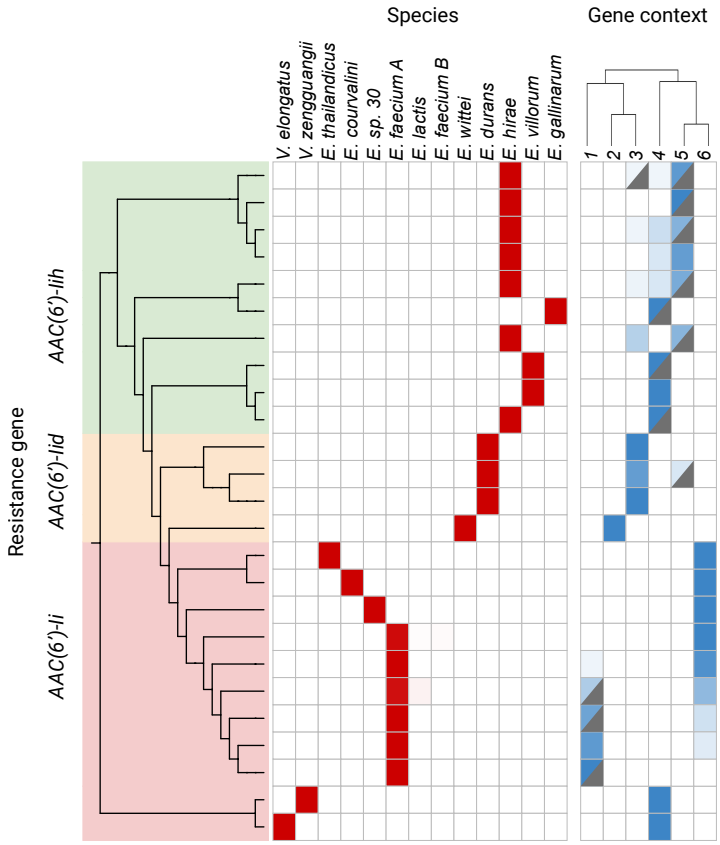


Figure 6.6: *Enterococcus* species are intrinsically resistant to aminoglycosides through three variants of the gene *AAC(6')-Ii*: *AAC(6')-Ii* (red tree), *AAC(6')-Iid* (orange tree), *AAC(6')-Iih* (green tree). Red cells are a heatmap for the proportion of the gene presence/absence within each species listed in the columns and ordered w.r.t. their position in the genus tree. Heatmap rows are ordered according to the gene trees shown on the far left for each gene. Blue cells on the right panel are the proportion of gene sequences found in one of the six gene contexts, and the concordance between different gene contexts is displayed with a dendrogram positioned on top of the column names.

these two species as well as *E. gallinarum*. Given that *E. gallinarum* is phylogenetically distant, positioned within clade IV, this species likely acquired *AAC(6')-Iih*, and it is not intrinsically resistant yet.

Since it is not possible to set an accurate root for the gene tree, the root was placed at the midpoint in Figure 6.6, and hence it is tricky to infer the earliest variant of the gene. However, we speculate that *AAC(6')-Ii*, being chromosomally encoded the most among all variants, is the earliest example of this gene. This is also supported by the two *Vagococcus* species (*V. elongatus* and *V. zengguangii*) that carry the same gene on the chromosome; *Vagococcus* being ancestrally related to *Enterococcus* (Figure 6.3, it is more likely the earliest variant of *AAC(6')-Ii* emerged in *Vagococcus*.

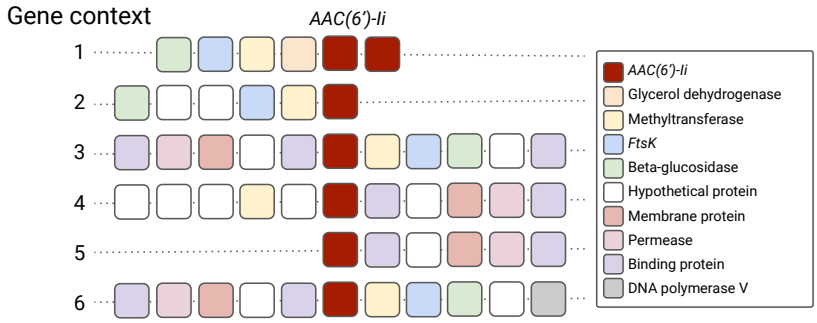


Figure 6.7: *AAC(6')-Ii* gene and its two variants, *AAC(6')-Iid* and *AAC(6')-Iih*, are found in six different gene contexts where the synteny is conserved on the genomes. See Table S6.1 for the same contexts listed in more detail.

6.3.5 Patterns of intrinsic resistance correlates with subspecies differentiation in *E. faecium*

Due to its clinical significance, AMR in *E. faecium* has been extensively studied; species are known to be intrinsically resistant to lincosamides and fluoroquinolone. Chromosomally encoded *msrC* and *eatAv* confer resistance to lincosamides, and fluoroquinolone resistance is associated with the *efmA* gene. In our work, we confirm these three genes to satisfy our definitions of intrinsic resistance as well, and we find that they correlate with the subspecies differentiation in *E. faecium*.

E. faecium A carries *efmA* gene on the chromosome, not associated with any mobile element, while *eatAv* and *msrC* genes are intrinsic in *E. faecium B* and *E. lactis*. We note that both of these genes also appear sporadically in *E. faecium A* as we found that 57% of the *E. faecium A* genomes in our dataset carry at least one copy of *msrC* gene, and 19% carry at least one copy of *eatAv* on the chromosome. Compared to the commensal *E. faecium B*, the *E. faecium A* genome is highly malleable and it can readily acquire new resistance genes; thus we also see more AMR genes appear sporadically in *E. faecium A* (Tables S6.3-S6.5).

Finally, we note that we also found *tetM*, a wide-spread AMR gene conferring resistance to tetracyclines, to be sporadic in *E. faecium*, not associated with any mobile elements (Table S6.5). According to previous studies, *tetM* gene is the most prominent cause of tetracycline in *Enterococcus*, however, there are no reports of it being intrinsic [14]. In our dataset, four novel species recently isolated appeared to be intrinsically resistant to tetracycline; namely *E. devriesei*, *E. saigonensis*, *E. songbeiensis* and *E. xiangfengensis* (Table S6.5). All four of these species, being identified only recently, have only a handful of genomes available publicly. In their study, the authors Li and Gu isolated the four species from pickle juice and reported them to be sensitive to aminoglycosides [42]. While collateral sensitivity between aminoglycosides and tetracyclines has been shown in *K. pneumoniae*, there are no systematic studies performed on *Enterococcus* [43]. We speculate that a similar mechanism can be found in enterocci as well; it is possible that in the absence of aminoglycoside resistance, enterococci are more likely to develop resistance to tetracyclines. Thus, we suggest it should be investigated in future studies. Finally, given the sporadic appearance of *tetM* in multiple species, one could hypothesize that *tetM* is intrinsic to a few *Enterococcus*

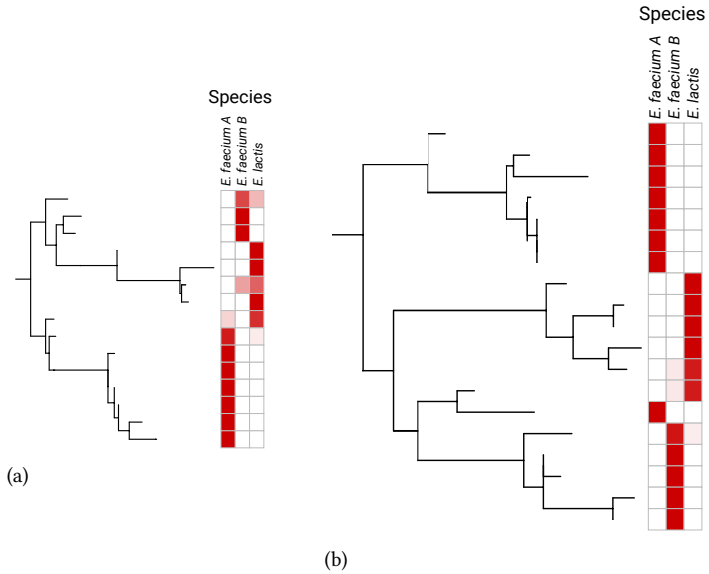


Figure 6.8: Patterns of intrinsic resistance correlate with subspecies differentiation in *E. faecium*: Both *eatAv* (a) and *msrC* (b) are intrinsic to *E. faecium B* and *E. lactis*, but they appear sporadically in *E. faecium A* as well.

species, and it is in the process of becoming intrinsic to several others as well since it is easily circulated within the genus [14].

6.3.6 Clade IV species exhibit distinct intrinsic resistance patterns

Unlike the clinically relevant species of *Enterococcus*, e.g. *E. faecium*, *E. faecalis* and *E. hirae*, clade IV species have been studied to a lesser extent. In our work, we found this clade, *E. casseliflavus* and its immediate neighbors, *E. flavescens* and *E. gallinarum* in particular, to stand out for their distinct resistance traits. This prominent feature is likely due to the ecological niche of the species. According to recent studies, clade IV species are the most commonly found members of enterococci in the guts of insects [33]. As insects encounter an abundance of microorganisms in diverse environments, the *Enterococcus* species residing in these organisms require a large repertoire of intrinsic resistance traits for survival.

We report the *Erm(O)-lrm* gene (CARD database short name) that confers lincomycin resistance, is intrinsic to *E. flavescens* and *E. casseliflavus 2*. It should be noted that *E. casseliflavus 2* species name is a placeholder we use throughout this study because we found the species to be closely related to both *E. flavescens* and *E. casseliflavus*, but well below the sequence identity threshold we defined especially for these two species (97% ANI). Since these genomes, on average, resemble *E. casseliflavus* more than they do *E. flavescens*, we decided to use the placeholder name *E. casseliflavus 2*. Despite having a higher ANI, *E. casseliflavus* lacks the *Erm(O)-lrm* gene (we found only 1 instance of the gene in *E. casseliflavus*).

Although *Erm(O)-lrm* gene is usually found in a similar context in both *E. flavescens* and

E. casseliflavus 2, there is a clear separation of these genes at the aminoacid level (Figure 6.9. The absence of *Erm(O)-lrm* gene in *E. casseliflavus*, and the differences in aminoacid sequence between the instances carried by *E. flavescens* and *E. casseliflavus 2* genomes suggest this gene would be particularly useful to differentiate among these three species. It has been hypothesized that speciation of *E. flavescens* and *E. casseliflavus* coincides with the subspecies split in *E. faecium* with *E. faecium A* and *E. faecium B*, in terms of the evolutionary timeline. Using *Erm(O)-lrm* gene as a marker, the parallel speciation within these species can be investigated.

In addition to lincomycin resistance, we found that clade IV species are intrinsically resistant to vancomycins through the *vanC* gene cluster. In all instances, the vanC operon is located on the chromosome, but the ligase gene itself differs at the aminoacid level. The ligase in *E. gallinarum* is at most 91% similar to the ligase sequences we found in other species. Similarly, the *vanC* gene in *E. sp. 24* forms a separate branching in the gene tree, although the aminoacid similarity can be as high as 95% in this case (Figure 6.9 (b)). However, the remaining ligase genes are not significantly different from one another as we found the aminoacid identity to be more than 97%. The most common variants carried by the species *E. flavescens* and its neighbors *E. casseliflavus* and *E. casseliflavus 2* are highly similar.

Interestingly, the overall gene context, both the composition and the order of genes, in the vanC operon is highly conserved among all species. This is consistent with the recent studies on vanC operons in enterococci: currently, there are three known variants of this operon that have been established based on the amino acid similarity of both the ligase genes as well as the components of the vanC operon. Despite the aminoacid difference in the individual genes, the operon structure is known to be highly conserved [13]. We note that in our dataset, the aminoacid identity between the genes ranged from 61% to 97%, and hence the genes diverge enough to be annotated as different genes, resulting in three distinct gene contexts in Figure 6.9 (b).

While vancomycin resistance in *E. gallinarum* and *E. flavescens* was observed previously, the vanC operon in *E. sp. 24*, an unnamed species from NCBI, is unknown. Although *E. sp. 24* satisfies all the criteria for intrinsic resistance, we note that there is only one instance of this species in our dataset and the immediate neighbors, *E. testudinum* and *E. sp. 32*, lack the VanC operon. Since all three of these species are recently discovered *Enterococcus*, we have fewer than 5 samples in total. Given their proximity to species already known to be intrinsically resistant to vancomycin, we suggest these new species warrant further study.

Finally, we note that *vanC* gene, and its context, has been studied to a much lesser extent compared to the other vancomycin resistance patterns in *Enterococcus*, hence there is limited data available in AMR gene databases. In our work, we found the *vanC* gene was often missing in most databases, or misannotated as *vanB*, another AMR gene associated with vancomycin resistance. When searching for the *vanC* gene in our dataset, we also checked for the genes *vanXYC* and *vanTC*, which are required for the vancomycin resistance phenotype, as well as the regulating genes *vanRC* and *vanSC* [44]. Since the regulators are not necessary for the ligase activity, they provide additional evidence when one of the required components of the operon is missing.

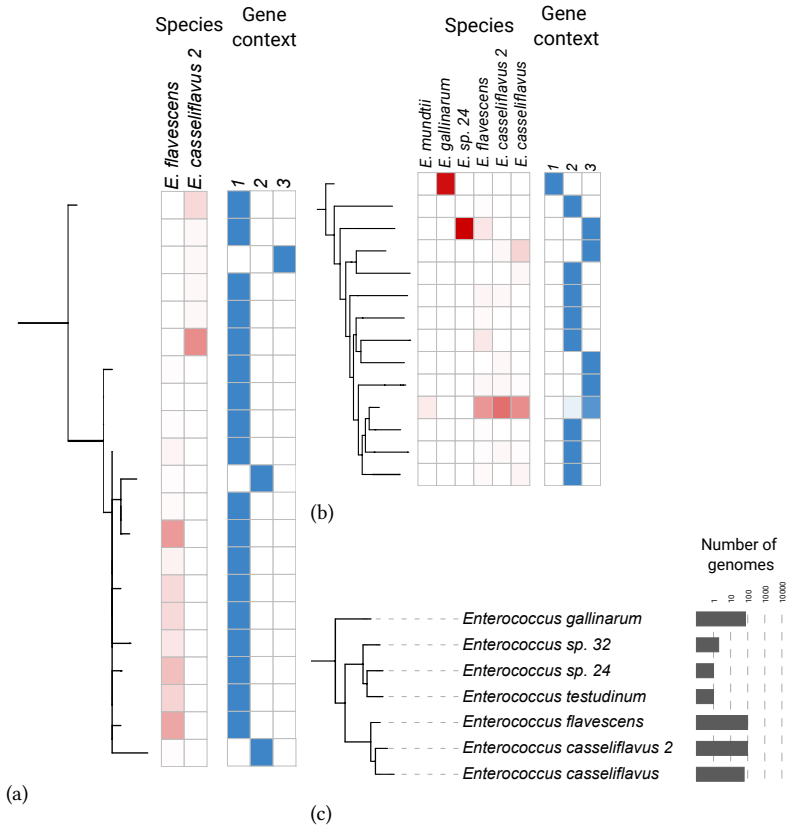


Figure 6.9: Clade IV species exhibit distinct resistance traits within the genus: (a) *E. flavescens* and *E. casseliflavus* variant 2 are inherently resistant to lincomycin: chromosomally-encoded *Erm(O)-lrm* gene differs at the aminoacid level among the species, while the genomic context of the gene is similar. (b) Evolutionary relationship of the species that exhibit intrinsic resistance to vancomycin and lincomycin: a close-up view obtained from the complete genus phylogenetic tree in Figure 6.3. (c) Evolutionary relationship of the species that exhibit intrinsic resistance to vancomycin and lincomycin: a close-up view obtained from the complete genus phylogenetic tree in Figure 6.3.

6.4 Conclusion

Enterocci are among the most widespread microbes in the guts of animals, and they are often associated with AMR in clinical settings leading to massive outbreaks. In recognition of their significance and the growing rate of novel species discovered recently, we have designed a robust and unbiased study in this work to investigate AMR traits in *Enteroccus*. We have amassed the largest collection of publicly available genomic data which allowed us to (i) define a systematic approach to identify AMR, (ii) distinguish between intrinsic and resistance types of resistance and (iii) provide a starting point to study AMR genes from when they are initially acquired by the bacterial species to it becoming intrinsic to all members of a strain, species or even an entire clade.

Our work is the most extensive study on AMR traits in *Enterococcus* in terms of its scope since our data is not limited to any specific sampling location, setting, drug class or species.

In parallel with our fundamental view of intrinsic resistance in bacteria, predating human use of antibiotics, we adopted a systematic approach to investigate intrinsic resistance in enterococci. In addition, we make use of the most recent methods and general conventions of data analysis in bioinformatics, addressing the lack of technical rigor in the field of microbiology. Hence, we provide a valuable pipeline to investigate AMR not only in *Enterococcus* genus but in all bacteria[2].

In addition to the technical value of our work, we also established clear genomic boundaries to distinguish enterococcal species. Since we accounted for the recently discovered species as well, we have successfully expanded the known species labels to 127 in total. With a bigger genus tree, we solidified the four clades of *Enterococcus* to facilitate future work on the genus. Our unique approach combining DNA sequence with synteny, and phylogenetics is a solid starting point to map out the progress of resistance traits as they become intrinsic within a strain, species or a clade of bacteria.

The main limitation of our study is that we only use genomic data, genome assemblies to be specific. While we took the most conservative approach in gene detection, we can assess AMR only in terms of the presence and absence of a gene. We validated AMR gene content based on genomic and phylogenetic analysis. In the absence of experimental work, our study provides a fully data-driven tool, bounded by the publicly available genome assemblies. Future work should first supplement our findings with not only clinical experiments and new samples, but also additional sources of genomic data, such as gene expression, and protein structures.

6.5 Supplementary material

6.5.1 Supplementary Tables and Figures

[2]You can find the python scripts and relevant documentation to reproduce our analyses, or adopt the same approach as we did to study other bacterial organism on github: https://github.com/aysunrhn/Intrinsic-antimicrobial-resistance/

Table S6.1: *AAC(6')-Ii* gene and its two variants are found in six different gene contexts where the synteny is conserved on the genomes.

Gene order	Gene context					
	1	**2**	**3**	**4**	**5**	**6**
1	beta-glucosidase	beta-glucosidase	ABC transporter ATP-binding protein	hypothetical protein		ABC transporter ATP-binding protein
2	FtsK	hypothetical protein	ABC transporter permease OppC	hypothetical protein		ABC transporter permease OppC
3	ribosomal RNA methyltransferase N	hypothetical protein	ABC transporter membrane protein	hypothetical protein		ABC transporter membrane protein
4	glycerol dehydrogenase	FtsK	hypothetical protein	ribosomal RNA methyltransferase N		hypothetical protein
5	GNAT family acetyltransferase	ribosomal RNA methyltransferase N	ABC transporter binding protein	hypothetical protein		ABC transporter binding protein
6	GNAT family acetyltransferase	GNAT family acetyltransferase	GNAT family acetyltransferase	GNAT family acetyltransferase	GNAT family acetyltransferase	GNAT family acetyltransferase
7	GNAT family acetyltransferase		ribosomal RNA methyltransferase N	ABC transporter binding protein	ABC transporter binding protein	ribosomal RNA methyltransferase N
8			FtsK	hypothetical protein	hypothetical protein	FtsK
9			beta-glucosidase	ABC transporter membrane protein	ABC transporter membrane protein	beta-glucosidase
10			hypothetical protein	ABC transporter permease OppC	ABC transporter permease OppC	hypothetical protein
11			extracellular solute-binding protein	ABC transporter ATP-binding protein	ABC transporter ATP-binding protein	DNA polymerase V

Table S6.2: Intrinsic AMR genes with antibiotic inactivation mechanism.

Antibiotic	Gene family	Gene name	Intrinsic	Sporadic		Acquired
				Species	%	
Aminoglycosides	*AAC(6')*	*AAC(6')-Ii*	*E. faecium A, E. faecium B, E. lactis E. thailandicus, E. sp. 30, E. sp. 61, E. sp. 91*			
Aminoglycosides	*AAC(6')*	*AAC(6')-Iid*	*E. hirae, E. villorum*			
Aminoglycosides	*AAC(6')*	*AAC(6')-Iih*	*E. durans, E. sp. 13*			

Table S6.3: Intrinsic AMR genes with antibiotic efflux mechanism.

Antibiotic	Gene family	Gene name	Intrinsic	Sporadic		Acquired
				Species	%	
Fluoroquinolone	major facilitator superfamily antibiotic efflux pump	*efmA*	*E. faecium A*			
	ATP-binding cassette antibiotic efflux pump	*efrA*	Absent in *E. hulanensis* and clade III (except *E. sp. 201*)			
	ATP-binding cassette antibiotic efflux pump	*efrB*	*E. faecalis O and E. faecalis V*			
Lincosamides	lsa-type ABC-F protein	*lsaA*	*E. faecalis O and E. faecalis V*			
	msr-type ABC-F protein	*MsrC*	*E. faecium B, E. lactis*	*E. faecium A*	56	
disinfecting agents and antiseptics	multidrug and toxic compound extrusion transporter	*emeA*	*E. faecalis*			

Table S6.4: Intrinsic AMR genes with antibiotic target alteration mechanism.

Antibiotic	Gene family	Gene name	Intrinsic	Sporadic		Acquired
				Species	%	
Vancomycin	Van ligase	*vanC*	*E. gallinarum, E casseliflavus, E casseliflavus 2* and *E. sp. 241*	*E. flavescens*	50	*E. flavescens*
		vanE	E. caccae and E. sp. 5			
		vanI	*E. massiliensis, E. songbeiensis*1 and *E. xiangfangensis*			
Lincosamides	Erm 23S ribosomal RNA methyltransferase	*Erm(O)-lrm*	*E. casseliflavus 2* and *E. flavescens*	*E. cass*	54	

Table S6.5: Intrinsic AMR genes with antibiotic target protection mechanism.

Antibiotic	Gene family	Gene name	Intrinsic	Sporadic		Acquired
				Species	%	
Fluoroquinolone	quinolone resistance	*qnrE2*	*E. thailandicus*			
Lincosamide	ribosomal protection	*eatAv*	*E. faecium B* and *E. lactis*	*E. faecium A*	19	
Tetracyclines	ribosomal protection	*tetM*	*E. devriesei E. saigonensis, E. songbeiensis E. xiangfengensis*	*E. avium*	28	*E. asini, E. avium, E. cecorum, E. durans, E. faecalis, E. faecium A, E. faecium B, E. flavescens, E. gallinarum, E. hirae, E. lactis, E. lemanii, E. raffinosus, E. sp. 31, E. sp. 9, E. thailandicus*
				E. cecorum	53	
				E. durans	7	
				E. faecalis O	31	
				E. faecalis V	20	
				E. faecium A	15	
				E. gallinarum	10	
				E. lemanii	50	
				E. pseudoavium	50	
				E. raffinosus	40	
				E. villorum	25	
		tetS	*E. sp. 10*	*E. cass*	5	*E. asini, E. faecium A, E. gallinarum, E. pseudoavium*
				E. flavescens	7	
				E. gallinarum	14	
				E. massiliensis	25	
				E. thailandicus	6	

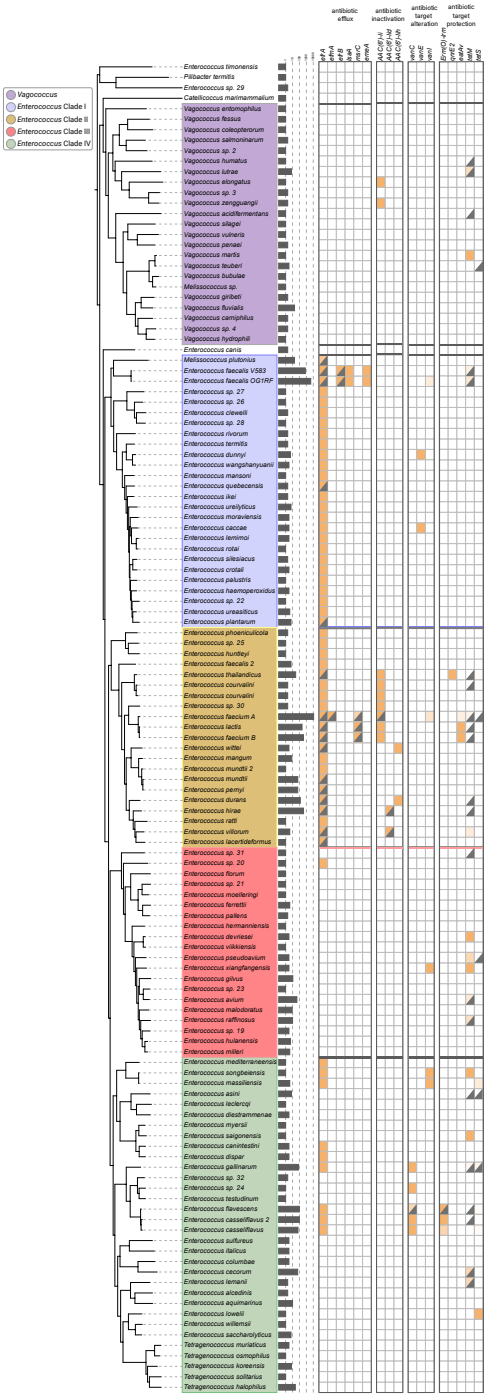


Figure S6.1: Clade-specific patterns of intrinsic AMR are more prominent than species-specific patterns.

References

[1] J Orvin Mundt. Occurrence of enterococci in animals in a wild environment. *Applied microbiology*, 11(2):136–140, 1963. doi: 10.1128/am.11.2.136-140.1963.

[2] Anthony O Gaca and José A Lemos. Adaptation to adversity: the intermingling of stress tolerance and pathogenesis in enterococci. *Microbiology and Molecular Biology Reviews*, 83(3):10–1128, 2019. doi: 10.1128/mmbr.00008-19.

[3] Kathy E Raven, Sandra Reuter, Theodore Gouliouris, Rosy Reynolds, Julie E Russell, Nicholas M Brown, M Estée Török, Julian Parkhill, and Sharon J Peacock. Genome-based characterization of hospital-adapted enterococcus faecalis lineages. *Nature Microbiology*, 1(3):1–7, 2016. doi: 10.1038/nmicrobiol.2015.33.

[4] François Lebreton, Willem van Schaik, Abigail Manson McGuire, Paul Godfrey, Allison Griggs, Varun Mazumdar, Jukka Corander, Lu Cheng, Sakina Saif, Sarah Young, et al. Emergence of epidemic multidrug-resistant enterococcus faecium from animal and commensal strains. *MBio*, 4(4):10–1128, 2013. doi: 10.1128/mbio.00534-13.

[5] Patrick F Mc Dermott, Robert D Walker, and David G White. Antimicrobials: modes of action and mechanisms of resistance. *International journal of toxicology*, 22(2): 135–143, 2003. doi: 10.1080/10915810305089.

[6] R Fontana, P Canepari, MM Lleo, and G Satta. Mechanisms of resistance of enterococci to beta-lactam antibiotics. *European Journal of Clinical Microbiology and Infectious Diseases*, 9:103–105, 1990. doi: 10.1007/bf01963633.

[7] François Lebreton, Abigail L Manson, Jose T Saavedra, Timothy J Straub, Ashlee M Earl, and Michael S Gilmore. Tracing the enterococci from paleozoic origins to the hospital. *Cell*, 169(5):849–861, 2017. doi: http://dx.doi.org/10.1016/j.cell.2017.04.027.

[8] Carmen Torres, Carla Andrea Alonso, Laura Ruiz-Ripa, Ricardo León-Sampedro, Rosa Del Campo, and Teresa M Coque. Antimicrobial resistance in enterococcus spp. of animal origin. *Antimicrobial resistance in bacteria from livestock and companion animals*, pages 185–227, 2018. doi: 10.1128/9781555819804.ch9.

[9] William R Miller, Jose M Munita, and Cesar A Arias. Mechanisms of antibiotic resistance in enterococci. *Expert review of anti-infective therapy*, 12(10):1221–1236, 2014. doi: 10.1586/14787210.2014.956092.

[10] Brian L Hollenbeck and Louis B Rice. Intrinsic and acquired resistance mechanisms in enterococcus. *Virulence*, 3(5):421–569, 2012. doi: 10.4161/viru.21282.

[11] Vanja Piezzi, Nasstasja Wassilew, Andrew Atkinson, Stéphanie D'Incau, Tanja Kaspar, Helena MB Seth-Smith, Carlo Casanova, Pascal Bittel, Philipp Jent, Rami Sommerstein, et al. Nosocomial outbreak of vancomycin-resistant enterococcus faecium (vre) st796, switzerland, 2017 to 2020. *Eurosurveillance*, 27(48):2200285, 2022. doi: 10.2807/1560-7917.ES.2022.27.48.2200285.

[12] Abdulhakim Abamecha, Beyene Wondafrash, and Alemseged Abdissa. Antimicrobial resistance profile of enterococcus species isolated from intestinal tracts of hospitalized patients in jimma, ethiopia. *BMC research notes*, 8:1–7, 2015.

[13] Ireena Dutta and Peter E Reynolds. Biochemical and genetic characterization of the vanc-2 vancomycin resistance gene cluster of enterococcus casseliflavus atcc 25788. *Antimicrobial agents and chemotherapy*, 46(10):3125–3132, 2002. doi: 10.1128/aac.46.10.3125-3132.2002.

[14] Yingjie Tian, Hui Yu, and Zhanli Wang. Distribution of acquired antibiotic resistance genes among enterococcus spp. isolated from a hospital in baotou, china. *BMC research notes*, 12:1–5, 2019. doi: 10.1186/s13104-019-4064-z.

[15] Aysun Urhan, Bianca-Maria Cosma, Ashlee M Earl, Abigail L Manson, and Thomas Abeel. Safpred: Synteny-aware gene function prediction for bacteria using protein embeddings. *Bioinformatics*, 40(6):btae328, 2024. doi: 10.1093/bioinformatics/btae328.

[16] Yannick Mahlich, Chengsheng Zhu, Henri Chung, Pavan K Velaga, M Clara De Paolis Kaluza, Predrag Radivojac, Iddo Friedberg, and Yana Bromberg. Learning from the unknown: exploring the range of bacterial functionality. *Nucleic Acids Research*, 51 (19):10162–10175, 2023. doi: 10.1093/nar/gkad757.

[17] National center for biotechnology information (ncbi)[internet, 1988. URL `https://www.ncbi.nlm.nih.gov/`. Accessed 1 Jan 2020.

[18] Donovan H. Parks, Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7):1043–1055, 2015. doi: 10.1101/gr.186072.114. URL `http://genome.cshlp.org/content/25/7/1043.abstract`.

[19] Aaron E Darling, Guillaume Jospin, Eric Lowe, Frederick A Matsen IV, Holly M Bik, and Jonathan A Eisen. Phylosift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, 2:e243, 2014. doi: 10.7717/peerj.243.

[20] Sean R Eddy. Accelerated profile hmm searches. *PLoS computational biology*, 7(10): e1002195, 2011. doi: 10.1371/journal.pcbi.1002195.

[21] Léa Pradier and Stéphanie Bedhomme. Ecology, more than antibiotics consumption, is the major predictor for the global distribution of aminoglycoside-modifying enzymes. *bioRxiv*, 2022. doi: 10.1101/2022.01.07.475340. URL `https://www.biorxiv.org/content/early/2022/01/07/2022.01.07.475340`.

[22] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30 (14):2068–2069, 2014. doi: 10.1093/bioinformatics/btu153.

[23] Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, et al. Interpro in 2022. *Nucleic Acids Research*, 51(D1):D418–D427, 2023. doi: 10.1093/nar/gkac993.

[24] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000. doi: 10.1093/nar/28.1.27.

[25] Takuya Aramaki, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. Kofamkoala: Kegg ortholog assignment based on profile hmm and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252, 2020. doi: 10.1093/bioinformatics/btz859.

[26] Alessandra Carattoli, Ea Zankari, Aurora García-Fernández, Mette Voldby Larsen, Ole Lund, Laura Villa, Frank Møller Aarestrup, and Henrik Hasman. In silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrobial agents and chemotherapy*, 58(7):3895–3903, 2014. doi: 10.1128/aac.02412-14.

[27] James Robertson and John HE Nash. Mob-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial genomics*, 4(8), 2018. doi: 10.1099%2Fmgen.0.000206.

[28] Shengwei Hou, Siliangyu Cheng, Ting Chen, Jed A Fuhrman, and Fengzhu Sun. Deepmicrobefinder sorts metagenomes into prokaryotes, eukaryotes and viruses, with marine applications. *BioRxiv*, pages 2021–10, 2021. doi: 10.1101/2021.10.26.466018.

[29] Karen Ross, Alessandro M Varani, Erik Snesrud, Hongzhan Huang, Danillo Oliveira Alvarenga, Jian Zhang, Cathy Wu, Patrick McGann, and Mick Chandler. Tncentral: a prokaryotic transposable element database and web portal for transposon analysis. *MBio*, 12(5):10–1128, 2021. doi: 10.1128/mbio.02060-21.

[30] Patricia Siguier, Jocelyne Pérochon, L Lestrade, Jacques Mahillon, and Michael Chandler. Isfinder: the reference centre for bacterial insertion sequences. *Nucleic acids research*, 34(suppl_1):D32–D36, 2006. doi: 10.1093/nar/gkj014.

[31] João L Reis-Cunha, Daniella C Bartholomeu, Abigail L Manson, Ashlee M Earl, and Gustavo C Cerqueira. Prophet, prophage estimation tool: A stand-alone prophage sequence prediction tool with self-updating reference database. *PLoS One*, 14(10):e0223364, 2019. doi: 10.1371/journal.pone.0223364.

[32] 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*, 51(D1):D690–D699, 2023. doi: 10.1093/nar/gkac920.

[33] Julia A Schwartzman, Francois Lebreton, Rauf Salamzade, Terrance Shea, Melissa J Martin, Katharina Schaufler, Aysun Urhan, Thomas Abeel, Ilana LBC Camargo, Bruna F Sgardioli, et al. Global diversity of enterococci and description of 18 previously unknown species. *Proceedings of the National Academy of Sciences*, 121(10):e2310852121, 2024. doi: 10.1073/pnas.2310852121.

[34] Chirag Jain, Luis M Rodriguez-R, Adam M Phillippy, Konstantinos T Konstantinidis, and Srinivas Aluru. High throughput ani analysis of 90k prokaryotic genomes reveals clear species boundaries. *Nature communications*, 9(1):5114, 2018. doi: 10.1038/s41467-018-07641-9.

[35] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome biology*, 17(1):1–14, 2016. doi: 10.1186/s13059-016-0997-x.

[36] David M Emms and Steven Kelly. Orthofinder: phylogenetic orthology inference for comparative genomics. *Genome biology*, 20:1–14, 2019. doi: 10.1186/s13059-019-1832-y.

[37] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt Von Haeseler, and Robert Lanfear. Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, 37(5):1530–1534, 2020. doi: 10.1093/molbev/msaa131.

[38] Marco Teixeira, Stephanie Pillay, Aysun Urhan, and Thomas Abeel. Ship: identifying antimicrobial resistance gene transfer between plasmids. *Bioinformatics*, page btad612, 2023. doi: 10.1093/bioinformatics/btad612.

[39] Tomasz Goluch, Damian Bogdanowicz, and Krzysztof Giaro. Visual treecmp: comprehensive comparison of phylogenetic trees on the web. *Methods in Ecology and Evolution*, 11(4):494–499, 2020. doi: 10.1111/2041-210X.13358.

[40] Dan Alexandru Toc, Stanca Lucia Pandrea, Alexandru Botan, Razvan Marian Mihaila, Carmen Anca Costache, Ioana Alina Colosi, and Lia Monica Junie. Enterococcus raffinosus, enterococcus durans and enterococcus avium isolated from a tertiary care hospital in romania—retrospective study and brief review. *Biology*, 11(4):598, 2022. doi: 10.3390%2Fbiology11040598.

[41] Gerard D Wright and Pakizah Ladak. Overexpression and characterization of the chromosomal aminoglycoside 6'-n-acetyltransferase from enterococcus faecium. *Antimicrobial agents and chemotherapy*, 41(5):956–960, 1997. doi: 10.1128/aac.41.5.956.

[42] Yu Qin Li and Chun Tao Gu. Enterococcus pingfangensis sp. nov., enterococcus dongliensis sp. nov., enterococcus hulanensis sp. nov., enterococcus nangangensis sp. nov. and enterococcus songbeiensis sp. nov., isolated from chinese traditional pickle juice. *International journal of systematic and evolutionary microbiology*, 69(10):3191–3201, 2019. doi: 10.1099/ijsem.0.003608.

[43] Xinqian Ma, Wen Xi, Deqing Yang, Lili Zhao, Wenyi Yu, Yukun He, Wentao Ni, and Zhancheng Gao. Collateral sensitivity between tetracyclines and aminoglycosides constrains resistance evolution in carbapenem-resistant klebsiella pneumoniae. *Drug Resistance Updates*, 68:100961, 2023. doi: 10.1016/j.drup.2023.100961.

[44] Patrice Courvalin. Vancomycin resistance in gram-positive cocci. *Clinical infectious diseases*, 42(Supplement_1):S25–S34, 2006. doi: 10.1086/491711.

7

Discussion

"I met a lot of things on the way that astonished me."

— J.R.R Tolkien explaining to W. H. Auden
why he spent 12 years writing *Lord of the Rings*

The microbial world grows larger in our eyes each day. The rapid evolution of sequencing technologies, coupled with progress in bioinformatics tools and computational resources, led to new methodological developments that drove significant advances in microbial genomics [1]. With more genomic data than we could even dream of before, comparative genomics has revolutionized our understanding of microbial life. From unraveling the evolutionary relationships between different organisms to deciphering the genetic basis of microbial pathogenicity and antibiotic resistance, microbial genomics has emerged as a cornerstone of modern biological research. In this chapter, I will reflect on the progress I have made to elucidate the hidden world of microorganisms through the lens of comparative genomics. I will end up on a bittersweet note, reminding everyone how much more there is still out there to discover and the enduring challenges that lie ahead in our quest to unlock the full potential of microbial genomes.

From the pioneering efforts of the Human Genome Project to the current era of large-scale genomic data generation and analysis, the trajectory of microbial genomics reflects a journey marked by innovation, collaboration, and the relentless pursuit of knowledge[2]. Although I have had the opportunity to witness only a small fraction of this journey, I am proud to be part of this ongoing endeavor. In the context of these advances, my work is among the pioneering efforts, a massive undertaking to bridge the computational gap between eukaryotic and prokaryotic genomics. By applying novel methods and techniques originally developed for eukaryotic organisms to prokaryotic organisms, I wanted to push the boundaries of microbial genomics and have a peak into the microbial dark matter. This work not only expands our understanding of microbial biology but also demonstrates the versatility and adaptability of genomic methodologies across different domains of life. If done rightly, that is.

Overall, my five-year-long journey traversed the intricate landscapes of viruses and bacteria, leveraging cutting-edge computational methods to unravel hidden traits and illuminate fundamental aspects of microbial biology, evolution, and antimicrobial resistance. By focusing on microbial organisms, I hope I have done I have achieved my goal of highlighting the importance of studying microbial genomes, which have often taken a backseat compared to their eukaryotic counterparts. Understanding microbial life, their biology, population dynamics, and ecological roles, have implications far beyond the seemingly small world of microbial genomics; it transcends the boundaries of microbial genomics, resonating with the broader understanding of ourselves as a species. Our lives are intertwined with these microscopic organisms, and remember: we are as much microbial as we are human.

7.1 With big data comes great insight, but also great responsibility

One of the several main themes in my work has been the power of large genomics data to provide insight. Right from the start in Chapter 2 when we delved into the realm of viruses, focusing on the COVID-19 pandemic and the genomic dynamics of SARS-CoV-2 in the Netherlands. What set our work apart from most of the studies published at that time was the large collection of SARS-CoV-2 genomes. By aggregating this collection, we could gain insight into the local landscape of COVID-19, tracing its spread and monitoring emerging

variants; our findings are of great value for implementing targeted measures to control outbreaks and informing strategies for treatment and vaccine development, showcasing the utility of large-scale genomics data in addressing public health challenges.

Moving on to Chapter 3, our attention shifted to bacteria, where we confronted the issue of inadequate representations of bacterial population diversity, particularly in the case of *Acinetobacter baumannii*. Although microbiologists were well aware of the vastness of genetic diversity harbored by microbial genomes, it has never been as pronounced as it is today with public access to a large collection of genomes. Thus, the traditional linear reference sequences prove insufficient to capture this intricate genetic diversity. To overcome this limitation, we explored the emerging field of bacterial pangenomics and developed practical guidelines for its application. Our ensemble method for building pangenome graphs was effective in detecting structural variants, shedding light on clinically significant genetic elements such as MDR plasmids, with implications for human health.

In Chapter 4, we expanded our view of bacterial functional diversity when we developed our tool SAFPred to predict gene function in all bacteria. In SAFPred we combined two ideas: using alternative representations for protein sequences and bacterial synteny. SAFPred exploits bacterial synteny using SAFPredDB, a comprehensive database of bacterial operons and syntenic regions we have built in Chapter 5. Large-scale genomics data served as the foundation for building SAFPredDB, our database, providing a valuable genomic resource for comparative analyses and functional genomics research across diverse bacterial taxa. We designed a computational algorithm based on our proposed synteny model to build SAFPredDB, independent of any experimental validation. SAFPredDB is first and foremost significant since it addresses the need for a comprehensive catalog of bacterial synteny. In addition, by mining the largest collection of bacterial genomes available we could bypass the need for experiments to identify operons, demonstrating the power of large genomics data. This power was showcased thoroughly when we used SAFPredDB to improve gene function prediction in bacteria in Chapter 4 with our tool SAFPred. SAFPred, relying heavily on SAFPredDB, outperformed the state-of-the-art in the field to prove that conserved structural patterns in bacteria hold clues to their function.

The added value of big data is perhaps best observed in Chapter 6 in our study of AMR in *Enterococcus*. We expanded the scope of data analysis to encompass the most extensive range of enterococcal species, collected from a wide range of environments and hosts. Our work shed light on species boundaries, evolutionary relationships, and AMR traits within the genus that were not possible previously with limited data. Large-scale genomics data enabled the systematic exploration of AMR traits across *Enterococcus* populations, contributing to a deeper understanding of bacterial pathogenicity and AMR mechanisms. The comprehensive analysis of *Enterococcus* genomes underscored the importance of large-scale genomics data in uncovering hidden traits and fundamental aspects of bacterial biology and evolution in Chapter 6.

Although the work in this thesis highlights the transformative potential of large-scale genomics data in microbial research, our bioinformatics analyses have been limited to genomic data. There is a growing interest in our field in adopting a multi-omics approach which has proven useful for numerous applications in life sciences [3]. Similarly, within the scope of this thesis, incorporating other omics data, such as transcriptomics, proteomics, and metabolomics, could provide a more comprehensive understanding of microbial biology. A

multi-omics approach can reveal functional insights, regulatory mechanisms, and metabolic pathways, enhancing the depth of our analyses [4, 5]. In addition to providing further insight into microbial life, utilizing multiple sources and types of biological data can serve as a sanity check, and add support to our findings. This final point is particularly relevant for our work in Chapter 6 where we relied heavily on genomic data and defined AMR through the presence and absence of a gene. Our approach was based on previous work that established links between the genotypes we identified and AMR traits. However, experimental validation of predicted functions and genetic elements is essential to confirm their biological relevance. By diversifying our data types to include mutli-omics data, as well as functional validation experiments, such as gene knockout studies, functional assays, and phenotypic characterization, we can validate our predictions and enhance the credibility of our findings. While it is fairly easy to comment on ways we can overcome these limitations, the real challenge is to figure out *how* we can implement them. In practice, this will require new methodologies to analyze disparate data types and novel techniques to integrate them within our existing pipelines. I expect one direction of research will be focused on developing new techniques to model data and generate alternative representations to facilitate data integration in a unified framework. Already in Chapter 4 we explored this new perspective by replacing aminoacid sequences with protein embeddings. Since protein embeddings are essentially numerical vectors, they can seamlessly be incorporated into analyses and combined with any numerical data, allowing us to build on top of our tool SAFPred in the future. A final point I need to address is the practical consequences of big data, and the great responsibilities that it endows upon us. As the size of each data type continues to grow in depth, a multi-omics approach brings an orthogonal direction of growth where we expand the breadth of our data. I presume the added dimension will result in a compounding effect, placing an extra burden on our computing resources and once again bringing algorithmic scalability to our attention, if it has ever left it, that is. Many bioinformatics algorithms and analytical methods are not inherently designed to scale to large multi-omics datasets and thus developing scalable algorithms and computational frameworks capable of processing massive amounts of data efficiently is essential for enabling timely and cost-effective analyses.

7.2 You say eukaryote I say prokaryote: the perpetual need to catch up with eukaryotic genomics

Throughout this thesis, I have done my best to emphasize the importance of being up to date with novel methodological approaches in the field of genomics, as well as following the best practices in computational comparative genomics and data analysis. These two concepts are more interlinked than they have ever been as we borrow more ideas and techniques from different fields, especially those outside the realm of genomics, blurring the lines in between. The boundaries between disciplines are becoming increasingly porous, allowing for the cross-pollination of ideas and methodologies. And yet, novel methods in machine learning, network analysis, or new deep learning models still make their way into eukaryotic genomics research first and sadly remain within their confinements [6]. Although some researchers have been let down by the difficulties in applying newer techniques, or unimpressed by the marginal gains from exceedingly more intricate approaches

developed for larger and more complex organisms such as humans, in microbial genomics we have a lot to learn from these advancements.

My goal in this thesis was to avoid blind adoption and advocate for assimilating these methods into our work, *tailoring* them to fit the metaphorical *body* of microbial genomics. The vastness and diversity of genomic data provide a fertile ground for methodological innovation, enabling the development and refinement of computational tools and analytical frameworks to address complex biological questions in microbial genomics. I have demonstrated this in Chapter 4 where I developed SAFPred, influenced by the impressive progress made in protein language models, a concept adopted from NLP into bioinformatics [7]. Protein language models are a perfect example of this interdisciplinary fluidity. Proteins also speak a language of their own; just as sentences and paragraphs convey meaning in human language, the arrangement of amino acids in a protein sequence dictates its structure, function, and interactions within biological systems. Protein language models exploit the *linguistic* patterns learned through their deep learning architectures trained on millions of protein sequences, to capture the complex relationships between amino acids and predict various properties of proteins, ranging from their secondary and tertiary structures to their biological functions [7]. Given the propensity to prioritize eukaryotic genomes, I address the large gap in research where we know very little about how these models perform on bacteria in Chapter 4. Our work is especially significant in our experimental setup and approach to evaluating SAFPred; we presented a rigorous benchmark that, despite being limited in size, allowed us to make inferences that extend beyond the scope of the benchmark. It was through this systematic approach that we confirmed that we can use protein language models to extract new representations of bacterial proteins that are more meaningful than aminoacid sequences to infer functional attributes. Our work follows the established conventions in the field of automated function prediction, modified to suit bacterial organisms, and thus serves as a framework for future studies.

Chapter 4 stands out as a significant achievement in combining new methodologies with our knowledge and experience in bacterial genomics to develop new methods and pipelines tailored for microbial organisms. In SAFPred, we harnessed bacterial synteny, a widely established concept in bacterial genomics, and protein embeddings to present a new tool that improves bacterial function prediction. Although we have achieved greater prediction accuracy, our approach can be expanded to account for more aspects of bacterial genomes. For instance, in SAFPredDB we treat all conserved genomic structures as equal, i.e. we do not differentiate between operons, mobilizable genomic islands or MGE associated regions. All of these different structures might appear similar *on paper*, however they could have considerably different implications for their functional role in a bacterial genome. Genomic islands and mobile regions are associated with MDR and virulence, and they mediate the spread of AMR in pathogens [8]. Since a large fraction of microbial genomics is devoted to understanding AMR, integrating information about MGE in our function prediction pipeline would allow us to understand the dynamics of AMR, in addition to aiding in the surveillance and management of AMR bacteria. Apart from the clinical significance of MGE, accounting for HGT and the dynamic nature of bacterial DNA would help identify foreign genes acquired through horizontal transfer. In Chapter 6 we have emphasized the need to differentiate functional traits that are intrinsic to an organism from those that are acquired from its environment. This difference has implications for not just our understanding of

the evolution of a gene across different taxa, but also for deciphering the evolution of an organism as it adapts to its niche through the acquisition of new traits. By identifying HGT events, we can also trace the origin of these traits and understand how they contribute to microbial adaptation and ecological success.

7.3 The isolating effects of isolate genomics

The conventional approach to genome sequencing is based on cultured organisms. Isolate genomes, obtained through the cultivation and sequencing of individual microbial strains in laboratory settings, serve as foundational resources for microbial research, enabling the characterization of key features such as genome structure, gene content, and functional potential [9]. Isolate genome assemblies provide valuable insights into the genetic makeup, physiological traits, and metabolic capabilities of specific microbial taxa. Metagenomics, on the other hand, offers a complementary approach by directly sequencing DNA from environmental samples, providing a holistic view of microbial communities without the need for cultivation. Metagenomic data capture the genetic diversity and functional potential of an entire ecosystem, including both cultured and uncultured taxa, shedding light on ecosystem dynamics, biogeochemical processes, and microbial interactions [5]. It has been recognized for a long time now that isolate genomes represent only a fraction of microbial diversity; since many microorganisms resist laboratory cultivation or exist in natural environments as part of complex communities, it is currently estimated that we can culture less than 1% of microbial species[10]. Uncultured bacteria, which cannot be grown using traditional laboratory techniques, represent a significant portion of *microbial dark matter*, the ever so elusive realm of microbial diversity that remains largely inaccessible through culture-dependent methods.

Metagenomics gives us a holistic view of all microbial organisms sharing the same environment. For instance, this "beyond the isolate genome" view allows us to study HGT with more detail and accuracy. HGT plays a crucial role in shaping microbial interactions within ecosystems. By acquiring genes involved in niche specialization, resource utilization, and competitive advantage, microbes can exploit diverse ecological niches and establish complex relationships with other organisms [11]. Understanding HGT-mediated interactions between microbial organisms and their environment provides valuable insights into ecosystem dynamics, microbial community structure, and the coevolution of host-microbe relationships. It is incorrect to view microbes in isolation, their interactions with their environment contribute to their remarkable diversity and plasticity. By exchanging genetic material with other organisms, bacteria can explore a vast genetic reservoir and access a wide range of functional traits. Similarly, new functional pathways emerge in bacteria as they adapt to their environment. The holistic view will allow us to gain insight into the mechanisms driving bacterial diversification and the evolutionary forces shaping microbial communities.

A crucial point in studying microbial communities is taxonomic classification [12]. Taxonomic profiling of metagenomic samples remains a core challenge despite the impressive improvements in the accuracy, resolution and the scalability of taxonomic classification methods. [13] Hidden under such practical issues, lies a more philosophical inquiry: *what is a species*? I personally find it difficult to answer this question in the context of microbial organisms. Defining species boundaries in microbial organisms is far from being

straightforward due to their high levels of genetic diversity, extensive HGT, and frequent recombination events. These complexities often blur the traditional concept of species, making it difficult to apply conventional species definitions based on morphology or genetic similarity. Phylogenetic studies can supplement the species identification, however this approach has severe limitations as well. MGEs do not abide by any rules of species boundaries or taxonomic nodes, leading to mosaic genomes that defy traditional notions of vertical inheritance and complicate species delineation. Similarly, recombinant genomes are chimeras made up of shuffled variants, confounding efforts to reconstruct evolutionary relationships based on sequence similarity alone. HGT and recombination both lead to phylogenetic discordance, where different genomic regions or genes exhibit conflicting evolutionary histories. Hence the assumption of a single, bifurcating tree of life falters. Methods that account for reticulate evolutionary processes, such as phylogenetic networks, may be more appropriate for capturing the complex evolutionary relationships among microbes, although they have found limited applications.

Perhaps, it is a good opportunity to change our perspective when we are talking about microbial organisms. Given the challenges of defining species based on genetic relatedness alone, some researchers advocate for an "ecological species concept" (ESC), which defines species based on shared ecological niches and functional traits rather than genetic similarity [14]. This approach considers factors such as habitat preference, metabolic capabilities, and ecological interactions to delineate microbial species boundaries, acknowledging the importance of environmental adaptation and ecological context in microbial speciation [15]. The ESC framework emphasizes the role of ecological niche differentiation, and defines species based on their habitat preferences, metabolic capabilities, and interactions with other organisms. Although the ESC concept was put forth almost 40 years ago today, it remains useful as a point of view in certain applications to interpret research findings [14]. For instance, the concept has been brought up again in the last decade following several metagenomics surveys that suggest microbial communities are functionally redundant despite carrying immense taxonomic diversity [16]. By switching our focus from discriminating based on genotype, we can elucidate the functional roles of novel microbial taxa in diverse ecosystems. This will help us to reveal the hidden functional diversity and ecological significance of microbial dark matter within microbial communities. Sometimes all it takes is to look at things through *a new lens*.

References

[1] Alice Maria Giani, Guido Roberto Gallo, Luca Gianfranceschi, and Giulio Formenti. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*, 18:9–19, 2020. doi: 10.1016/j.csbj.2019.11.002.

[2] US DOE Joint Genome Institute: Hawkins Trevor 4 Branscomb Elbert 4 Predki Paul 4 Richardson Paul 4 Wenning Sarah 4 Slezak Tom 4 Doggett Norman 4 Cheng Jan-Fang 4 Olsen Anne 4 Lucas Susan 4 Elkin Christopher 4 Uberbacher Edward 4 Frazier Marvin 4, RIKEN Genomic Sciences Center: Sakaki Yoshiyuki 9 Fujiyama Asao 9 Hattori Masahira 9 Yada Tetsushi 9 Toyoda Atsushi 9 Itoh Takehiko 9 Kawagoe Chiharu 9 Watanabe Hidemi 9 Totoki Yasushi 9 Taylor Todd 9, Genoscope, CNRS

UMR-8030: Weissenbach Jean 10 Heilig Roland 10 Saurin William 10 Artiguenave Francois 10 Brottier Philippe 10 Bruls Thomas 10 Pelletier Eric 10 Robert Catherine 10 Wincker Patrick 10, Institute of Molecular Biotechnology: Rosenthal André 12 Platzer Matthias 12 Nyakatura Gerald 12 Taudien Stefan 12 Rump Andreas 12 Department of Genome Analysis, GTC Sequencing Center: Smith Douglas R. 11 Doucette-Stamm Lynn 11 Rubenfield Marc 11 Weinstock Keith 11 Lee Hong Mei 11 Dubois JoAnn 11, Beijing Genomics Institute/Human Genome Center: Yang Huanming 13 Yu Jun 13 Wang Jian 13 Huang Guyang 14 Gu Jun 15, et al. Initial sequencing and analysis of the human genome. *nature*, 409(6822):860–921, 2001. doi: 10.1038/35057062.

[3] Justin P Shaffer, Louis-Félix Nothias, Luke R Thompson, Jon G Sanders, Rodolfo A Salido, Sneha P Couvillion, Asker D Brejnrod, Franck Lejzerowicz, Niina Haiminen, Shi Huang, et al. Standardized multi-omics of earth's microbiomes reveals microbial and metabolite diversity. *Nature microbiology*, 7(12):2128–2150, 2022. doi: 10.1038/s41564-022-01266-x.

[4] Shirley Bikel, Alejandra Valdez-Lara, Fernanda Cornejo-Granados, Karina Rico, Samuel Canizales-Quinteros, Xavier Soberón, Luis Del Pozo-Yauner, and Adrián Ochoa-Leyva. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. *Computational and structural biotechnology journal*, 13:390–401, 2015. doi: 10.1016/j.csbj.2015.06.001.

[5] Jason Lloyd-Price, Cesar Arze, Ashwin N Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W Poon, Elizabeth Andrews, Nadim J Ajami, Kevin S Bonham, Colin J Brislawn, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, 2019. doi: 10.1038/s41586-019-1237-9.

[6] Irene van den Bent, Stavros Makrodimitris, and Marcel Reinders. The power of universal contextualized protein embeddings in cross-species protein function prediction. *Evolutionary Bioinformatics*, 17, 2021. doi: 10.1177/11769343211062608.

[7] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, 20:1–17, 2019. doi: 10.1186/s12859-019-3220-8.

[8] Elizabeth J Klemm, Vanessa K Wong, and Gordon Dougan. Emergence of dominant multidrug-resistant bacterial clades: Lessons from history and whole-genome sequencing. *Proceedings of the National Academy of Sciences*, 115(51):12872–12877, 2018. doi: 10.1073/pnas.1717162115.

[9] Nicholas J Loman and Mark J Pallen. Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology*, 13(12):787–794, 2015. doi: 10.1038/nrmicro3565.

[10] Corie Lok. Mining the microbial dark matter. *Nature*, 522(7556):270, 2015. doi: 10.1038/522270a.

[11] Maureen A O'Malley. 'everything is everywhere: but the environment selects': ubiquitous distribution and ecological determinism in microbial biogeography. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 39(3):314–325, 2008. doi: 10.1016/j.shpsc.2008.06.005.

[12] Stephen Nayfach, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, David Paez-Espino, I-Min Chen, Marcel Huntemann, et al. A genomic catalog of earth's microbiomes. *Nature biotechnology*, 39(4):499–509, 2021. doi: 10.1038/s41587-020-0718-6.

[13] Eric A Franzosa, Lauren J McIver, Gholamali Rahnavard, Luke R Thompson, Melanie Schirmer, George Weingart, Karen Schwarzberg Lipson, Rob Knight, J Gregory Caporaso, Nicola Segata, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nature methods*, 15(11):962–968, 2018. doi: 10.1038/s41592-018-0176-y.

[14] Lennart Andersson. The driving force: species concepts and ecology. *Taxon*, 39(3): 375–382, 1990. doi: 10.2307/1223084.

[15] Stilianos Louca, Saulo MS Jacques, Aliny PF Pires, Juliana S Leal, Diane S Srivastava, Laura Wegener Parfrey, Vinicius F Farjalla, and Michael Doebeli. High taxonomic variability despite stable functional structure across microbial communities. *Nature ecology & evolution*, 1(1):0015, 2016. doi: 10.1038/s41559-016-0015.

[16] Stilianos Louca, Martin F Polz, Florent Mazel, Michaeline BN Albright, Julie A Huber, Mary I O'Connor, Martin Ackermann, Aria S Hahn, Diane S Srivastava, Sean A Crowe, et al. Function and functional redundancy in microbial systems. *Nature ecology & evolution*, 2(6):936–943, 2018. doi: 10.1038/s41559-018-0519-1.

Acknowledgments

It takes 2 to tango, a village to raise a kid, but to complete a PhD? I don't think we'll ever know the exact number since PhD is a chaotic process. Almost as chaotic as winemaking; every day brings a new blend of spring frost, drought, and who knows what else climate change surprises us with. The concept of *terroir* in wine, *a sense of place*, echoes in academia as well. Just as wine is not just a product of the grape variety or winemaking techniques but also a reflection of the land and its specific conditions where the grapes are cultivated, PhD is a product of its terroir, which is the unique and chaotic combination of the environment and the people I have encountered with during my journey. This is the moment where I reflect on the past 5 years, recognizing the invaluable support, guidance, and inspiration that have shaped my path to this vintage.

First and foremost, I want to thank my parents for bringing me up and instilling me with values; they have planted the seeds of my love of lifelong learning and science. Their infinite support, care, and love are the main reasons for my academic achievements, and for my seemingly insane courage and will to chase after my goals. It pains me that it took me this long to realize how fortunate I am to be your kid; you were my first teachers and you'll remain as my guide until the end of life. I can only hope I will be a person worthy of your love and support. *Her seyden once beni yetistirdikleri ve bana degerler asiladiklari icin anne ve babama tesekkur etmek istiyorum; hayat boyu ogrenmeye ve bilime olan sevgimin tohumlarını ektiler. Onların sonsuz destegi, ilgisi ve sevgisi, akademik basarilarimin ve gorunuste cilginca olan cesaretimin ve hedeflerimin pesinden kosma istegimin ana nedeni. Sizin gibi ebeveynlerin oldugu icin ne kadar sansli oldugumu anlamamin bu kadar uzun surmesi bana aci veriyor; sizler benim ilk ogretmenlerimdiniz ve hayatimin sonuna kadar rehberim olarak kalacaksiniz. En buyuk hedefim sizin sevginize ve desteginize layik bir insan olmak.*

I would like to thank my promotor **Marcel Reinders**, I have felt your support throughout my time at Delft, although we did not meet often. Your feedback is always honest and to the point; you see through the shiny storefront right to the core of any problem. I wish I could have expressed more how grateful I am to have you as a promotor, a department head, and a member of my committee.

To **Thomas Abeel**, my supervisor, possibly the biggest reason why I am here writing this sentence right now. You are the guide I did not know I needed 5 years ago, but now realize was essential. Your scientific ambition, and passion to do *amazing*, *awesome* and *cool* science have encouraged me to never give up on my goals. It was only through your pushing me out of my comfort zone that I could find my way to become an independent researcher. Your kind and understanding nature, often overlooked, has allowed me to bounce back from the times I hit the lowest points during my PhD.

I would also like to thank **Burak Alakent**, my master's supervisor, who sparked my interest in statistics and the rest was history. I blame you for getting me trapped into the Ponzi scheme that is academia. Unlike many other professors in your department, you

truly want the best for your students, so you kicked me out of the country to do a PhD and I can't thank you enough for that. Rock on \m/.

The group of superheroes, that I know I could knock on their door anytime I ran into a problem and they had the power to solve every issue. **Azza**, welcome to our group, we are fortunate to have you. Thank you **Ruud** for providing me with all the single-use gadgets the new tech imposes on us, and sorry for bothering you with all my bike issues. **Bart**, I was broken down when I heard you were retiring, but I hope you are enjoying your passion about wildlife and photography. Thank you for taking one of the best headshots of my life to date 5 years ago when I first arrived, I'm still using that picture (with full credit, of course). **Saskia**, you are an incredible person who has shielded us from any HR issues. You are the mastermind who makes everything run smoothly, and there're no scheduling conflicts while I can barely manage my own agenda. **Marunka**, I'm sad we didn't overlap much in the group, but even in the limited time I have spent with you I can see how kind and thoughtful you are as a person. Your intricate gifts show how much you value each individual and it gives me hope for mankind.

The old DBL people have been there for me since the first day, and their place is irreplaceable. **Amelia** and **Ramin**, you are my oldest and dearest friends in Delft and I hope you're OK being part of my family here. I know I'm lucky to have you in my life. Amelia, you kept telling me you don't speak English well when I first met you pero, qué tontería! You've remained as humble and down to earth to this day. Ramin, you're the most selfless person I have ever seen, you're running around helping everyone all the time, I don't know how you could possibly find all that time to spend with your neighbor when she cooked for you. That brings me to the family in law, **Yosra** and **Chirag**. Thank you for hosting boardgame days and dinners, Yosra, you have such a kind heart. Chirag, I've never seen anyone as disciplined and hardworking as you are, but please take care of your health. I have so much more to learn from you about being a better extrovert. **Stavros**, you gave me the DBL 101 crash course and you were so patient with me nagging you about my cluster problems, and claiming that cacik is better than tzatziki. I was lying: tzatziki > cacik, gyros > doner, dolmades > dolma, giouvetsi > guvec. You and **Valentina** are the true power couple. I still have nightmares about that time when you guys came to visit me and I had a pregnant woman fish out a soap dispenser from my toilet. **Alex**, we never overlapped at the office but I heard so many tales about *the Sally*, and I feel like I got to know a lot about you when I had to debug the Ptolemy code. I'm glad you and **Diana** are back in Delft, and we will enjoy many more Cataplana feasts in the future. **Soufiane**, if it wasn't for you I'd still be struggling with understanding French accents. Thank you for hosting me at your place, and I'll never forget **Lisa**'s cheesecake. I'll let you know when I visit you in Cambridge so she has the time to bake it fresh. **Tom (Mokveld)** I enjoyed every conversation we had and all the nice restaurants we tried out in Den Haag. Thank you for hosting our boardgame nights, I'm still waiting for the invite to join one in Geneva. I hope you and **Farah** are having some much needed calm there. **Arlin**, I can't believe our group witnessed a true rockstar! When is your next gig? **Ahmed**, I don't know how you remain so humble being a great scientist. I admire your work ethic, also your willingness to run and bike long distance. **Christine**, our room was never the same without your captivating laughter (and the gossip, of course). **Christian**, I met you through the scary tales of your bike injury, but knowing you in person has changed my mind about German humor. I still

remember when you suggested starting a civil war among bacteria at Christine's biotalk to solve antibiotic resistance. **Meng**, we were separated by office walls and doors but I have great memories having you around or when we were out. I still can't believe how you just casually join every retreat with a fully functioning drone in your luggage. **Lieke**, I still remember your master's defense and now you already have a date for you PhD while I'm still writing the acknowledgments! Keep up the hard work, I hope you'll become a famous scientist one day. **Erik**, you always surprised me with how much you can teach me within a 5 minute coffee break. **Tamim**, you are a great colleague and a friend, I always enjoyed our conversations. **Mo**, I wish I could know you better because I always liked having you around and yet you remain a mystery to me.

The women of DBL, **Joana**, **Jasmijn** and **Jana**. Joana, I will fondly remember our extended chats about anything and everything whenever you gave me a ride to a conference or a retreat. Thank you for bringing the good Tawny port to the Christmas potluck. Jasmijn and Jana, you're always patient and understanding, and I'm glad you didn't want Osman's plants in your office, now I have 3 pots of them growing in my home.

And the new DBL folks, who arrived with such youthful energy but now became the old DBL folks, inevitably. Sassy **Stephanie** and caring **Chengyao**, the society of microbe people is under your responsibility now. **Paul**, you must have an incredibly strong personality to remain positive and cheerful all the time, you made the office more bearable. **Mostafa**, you have such a kind heart and a great friend to have. **Yasin**, the number of Turkish people in our room increased 100% when you joined, but you also started the chain of Joanafication of our room. **Colm**, **Roy** and **Sander**, you guys multiplied and took over the room. But I enjoyed our moments together, especially when you didn't open the windows during winter. **Jasper**, your straightforward tone hides a caring nature underneath. I wish we had more opinionated people like you in the group. Thank you for translating my summary into Dutch as well while recovering from COVID, I'm not even surprised you cared about the quality of the summary way more than I did. When is the next boardgame night? **Gerard** and **Kirti**, it was nice to hang out with you in Lyon. **Gabriel**, you are now the representative chemical engineer in our group. One day we'll beat the electrical engineers!

The students hall of fame: **Marco (Teixeira)** and **Bianca**. I guess it was karma after dealing with terrible students that I finally got to supervise you. Which was less supervising but mostly nodding my head. It can't be a coincidence that you're both countries that make great wine, I am still waiting for the invitation to visit vineyards. I could only wish you were on this list too, **Madelon** but I wasn't supervising you. Instead, I got to know you as a caring person who is as crazy as me to go to extreme lengths to compose a personalized birthday gift to every single person. October 25 better be marked on your calendar.

And of course, the nonbio part of our group. Although we were segregated when we moved to a new floor, you've made an impact. **David** and **Marco**, I've always left our coffee machine chats, be it 2 minutes or an hour long, incredibly enlightened and mindblown. **Haley**, I'm glad you're not exposed to all that noise from our room anymore. **Tom (Viering)**, you are the "work hard, play hard" concept in human form. I enjoyed our time partying in Amsterdam during Pride and King's Day, the latter of which almost killed me. Literally. **Ziqi**, don't worry, I didn't mean *the party*. **Robert-Jan** and **Attila**, I could tell you apart only because one of you is a brunette. **Arman**, stay politically incorrect!

Mahdi, I'm sorry we always interrupted you when all you wanted to do was recite some Persian poems. **Rickard**, I have yet to join you in Sweden for Midsommar so you can disprove the movie. **Osman**, I always saw you as the responsible brother in the group, you even cleaned after us in borrels, you always took care of us. **Yancong**, thank you for teaching me how to do Chinese hot pot properly. **Ombretta**, **Aurora**, **Marian**, **Taylan**, **Nergis** and **Hesam**, and many more people, I appreciated your presence at the office and I'll keep many fond memories from our times together.

During my time in Delft, I've been fortunate to make several friends outside of work that have become a core part of my life now. Starting with the Rotterdammers, the KINO host, **Matteo**. I'm glad I came all the way to Rotterdam just to watch Mission Impossible with a bunch of random strangers. It is rare to find people you not just share common interests with, but also a vision in life, if that makes any sense. We have to watch **Sunil** do standup for a gay Chinese crowd. **Anastasia**, the amazing businesswoman, you will open that bakery and I'll buy my birthday cake from you, OK? You have time until 25 October. **Jonas**, we have to do a festival together, and if audiobooks were a thing for thesis, I would love you to do the reading for mine. Then in Delft, I found my "second home" with **Anastasia (Z.)**. You were still learning English when we first met, but we managed to talk for hours. We connected at a deeper level through literature, art and poetry. I'm so thankful to have you as a friend and to be welcomed into your home. Every time I talk to **Leo**, I learn something new and gain a new perspective in life, he's an incredible kid. Our dinners together, with you, **Vladimir** and Leo are some of my favorite moments. **Eliza**, my Greek neighbor who I can count on to join me for a good night out. **Slavica**, I was sad to see you leave Delft. But you made up for it by being an amazing host in Munich and giving me slippers at your place. Finally the Amsterdammers: When I lost my passion for science, I found **wine**. I would like to raise a glass to the world of wine, where every sip tells a story of masterful winemaking, terroir, and culture. The bottle holds more than just some alcohol; it captures the essence of the land and tradition that goes back thousands of years, also deep friendships as I found out recently. Because soon enough, through wine I found friends whom I can count on to bring good vibes on any occasion. There are just too many names, most of which I don't even remember (that's how you know it was a good night) but we are all united through wine. Thank you for sharing that one extra bottle after the 3 other extra bottles we already had. Cheers!

And last but not least, I want to thank **music**. To the classic riffs, the pulsating rhythms, and the drum beat you feel deep inside your soul that served as the soundtrack to my journey, I extend my deepest gratitude. Music has always been more than just background noise to me, it's a companion that I know will never leave me and a source of boundless inspiration. As a proud metalhead through and through, music means life to me. It's a source of strength, resilience, and unwavering support that has carried me through the darkest of days, and for that, I am eternally grateful.

Curriculum Vitæ

Aysun Urhan

25-10-1994 Born in Istanbul, Turkey.

Education

2019-2024 **PhD. Computer Science**
Delft University of Technology, The Netherlands
Thesis: The Microbial World Magnified Through the Bright Lens of Comparative Genomics
Promotors: Dr. T.E.P.M.F. Abeel, and Prof. dr. ir. M.J.T. Reinders

2016–2018 **Master of Science in Chemical Engineering**
Bogazici University, Turkey
Thesis: Soft Sensor Design in Chemical Processes Using Statistical Learning Methods

2012–2016 **Bachelor of Science in Chemical Engineering**
Bogazici University, Turkey
Thesis awarded Highly Commended Work in The Undergraduate Awards in Dublin, Republic of Ireland

Experience

2021-2023 **MIT and Harvard - Broad Institute, Bacterial Genomics Group**
Cambridge, MA, United States (Remote)
Visiting PhD student

2016-2019 **Bogazici University, Chemical Engineering Department**
Istanbul, Turkey
Research and Teaching Assistant

2015 **Sandoz**
Istanbul, Turkey
Manufacturing Science and Technology Intern

List of Publications

1. J. A. Schwartzman, F. L., R. S., M. J. Martin, K. Schaufler, **A. Urhan**, T. Abeel, I. L.B.C Camargo, B. F. Sgardioli, J. Prichula, A. P. Guedes Frazzon, D. Van Tyne, G. Treinish, C. J. Innis, J. A. Wagenaar, R. M. Whipple,A. L. Manson, A. M. Earl, M. S. Gilmore. Global diversity of enterococci and description of 18 novel species. *Proceedings of the National Academy of Sciences*, 121(10):e2310852121, 2024. doi: 10.1073/pnas.2310852121.

2. **A. Urhan**, B. M. Cosma, A. M. Earl, A. L. Manson, and T. Abeel. SAFPred: Synteny-aware gene function prediction for bacteria using protein embeddings. *Bionformatics*, 40(60), 2024. doi: 10.1093/bioinformatics/btae328.

3. M. Teixeira, S. Pillay, **A. Urhan**, and T. Abeel. SHIP: identifying antimicrobial resistance gene transfer between plasmids. *Bioinformatics*, 39(10), pp. btad612, 2023. doi: 10.1093/bioinformatics/btad612.

4. R. S. H. Zade, **A. Urhan**, A. Assis de Souza, A. Singh, and T. Abeel. HAT: Haplotype Assembly Tool using short and error-prone long reads. *Bioinformatics*, 38(24), pp. 5352–5359, 2022. doi: 10.1093/bioinformatics/btac702.

5. S. Pillay, D. Calderón-Franco, **A. Urhan**, and T. Abeel. Metagenomic-based surveillance systems for antibiotic resistance in non-clinical settings. *Frontiers in Microbiology*, 13, pp. 1066995, 2022. doi: 10.3389/fmicb.2022.1066995.

6. **A. Urhan**, and T. Abeel. Emergence of novel SARS-CoV-2 variants in the Netherlands. *Scientific Reports*, 11(1), pp. 6625, 2021. doi: 10.1038/s41598-021-85363-7.

7. **A. Urhan**, and T. Abeel. A comparative study of pan-genome methods for microbial organisms: *Acinetobacter baumannii* pan-genome reveals structural variation in antimicrobial resistance-carrying plasmids. *Microbial Genomics*, 7(11), 2021. doi: 10.1099/mgen.0.000690.

8. **A. Urhan**, and B. Alakent. Integrating adaptive moving window and just-in-time learning paradigms for soft-sensor design. *Neurocomputing*, 392, pp. 23-37, 2020. doi: 10.1016/j.neucom.2020.01.083.

9. **A. Urhan**, and B. Alakent. An Exploratory Analysis of Biased Learners in Soft-Sensing Frames. *arXiv*, 2019. doi: 10.48550/arXiv.1904.10753.

10. **A. Urhan**, and B. Alakent. Soft-Sensor Design for a Crude Distillation Unit Using Statistical Learning Methods. *Computer Aided Chemical Engineering*, 44, pp. 2269-2274, 2018. doi: 10.1016/B978-0-444-64241-7.50373-6

Included in this thesis.