Flinders University

Metagenomics and the role of prophages in the human microbiome

By

Laura Kate Inglis
Bachelor of Science (Marine Biology) (Honours)

Thesis
Submitted to Flinders University
for the degree of

Doctor of Philosophy
College of Science and Engineering
31/10/2025

Table of contents

Thesis summary

There is still much to be discovered about the virome, the ecosystem of viruses that make up the viral component of the microbiome, especially the lysogenic phages that have integrated themselves into their bacterial hosts. Online databases are a huge boon to scientific discovery, allowing researchers to work with more data from more places than a whole lab could gather; however, freeform metadata fields mean that extensive manual curation can be required, depending on the questions being asked. I took advantage of these large repositories of bacterial DNA sequences to collect bacterial genomes from across the world to examine the prophages found within the human microbiome and what genes they may be providing to their hosts while they are sequestered away within their genomes.

This thesis aims to answer three main questions:
First, how the lysogenic prophages of the human microbiome vary across different areas of the human body. Second, to catalogue which antimicrobial resistance (AMR) genes are found in the prophages of the human microbiome and whether AMR genes are common in prophages or not, and lastly, whether the functional or taxonomic profiles can be used to train a machine learning algorithms to sort metagenomes by their isolation environments and alleviate some of the metadata holes in online databases.

To achieve these aims, I gathered genomes and whole-genome metagenomes from two online genomic databases, GenBank and MGnify, and manually curated tens of thousands of genomes. I analysed them with PhiSPy and AMRfinder+, ran statistical analyses in SPSS, wrote code in R and Python, used the university's High-performance computing resources, trained and tested several machine learning models on multiple sets of features, and extracted meaning from terabytes of data. I described the analysis in four papers, two of which have been published and two of which are soon to be submitted.

I found that the functional profile was more informative than the taxonomic profile for training a machine learning algorithm and that phage genes were some of the most important functional genes for differentiating isolation environments. Phage genes were found in every environment at different abundances. In the human microbiome, these phages varied on a smaller scale than the metagenomic dataset could show, with even areas of the body that were physically linked having very different amounts of prophage DNA in the bacterial genomes I analysed. The average amount of prophage DNA in a bacterial genome was also affected by more specific aspects of their environment, such as the health of the human host or the geographical region of the world where the sample was isolated from the human host. These factors also impacted the kinds of genes that the prophages were providing to their hosts. While I found that the presence of AMR genes was rarer in the human microbiome than some other studies claim, a large variety and in patterns that suggest that phages are transferring these genes between species.

Declaration

I certify that this thesis:
1. does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university
2. and the research within will not be submitted for any other future degree or diploma without the permission of Flinders University; and
3. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.
4. if generative artificial intelligence has been used in my thesis it has been duly acknowledged with details to identify the extent to which generative artificial intelligence formed the final thesis.

Date: 31/10/25
Signed: Laura Inglis

Acknowledgements

To start I'd like to thank my supervisor, Professor Rob Edwards, for all his support and advice over the course of this PhD. I'd also like to give thanks to everyone in the FAME lab, particularly Mike Roach, Bhavya, Susie, and Vijini for helping me learn to code and troubleshoot my first attempts. Thanks also to Jess, Abbey, Anita, Sarah and the rest of the lab for the more general advice and morale support.

I would also like to acknowledge the resources provided by Flinders University, particularly the Flinders HPC, which was crucial in completing this work.

Lastly, I'd like to thank my parents, sister and pets, for their love and support through the last few years, particularly the pets.

Funding

Awards from the NIH NIDDK RC2DK116713 and the Australian Research Council DP220102915 to R.A.E. supported this research.

The Australian Government Research Training Program (RTP) fee offset supported L.K.I.

Word count justification

Word count: 22,251

My thesis is composed of published or soon-to-be published work that I produced over the course of my PhD. This method of presenting a thesis was suggested by my supervisor as it is become a more common practice and having several published papers this early into my career would be a massive boost. However, many journals have word limits for papers, and even in ones that dont have strict limits, succinct writing is the standard. This leads to a much shorter overall thesis as it is made up of four papers of 4000-5000 words as was the norm for the journals I published in.

Publications list

First author papers

How metagenomics transformed our understanding of bacteriophage research

> Inglis, Laura K., and Robert A. Edwards. 2022. “How Metagenomics Has Transformed Our Understanding of Bacteriophages in Microbiome Research.” Microorganisms 10 (8): 1671.

Prophages: an integral but understudied component of the human microbiome

> Inglis, L.K., Roach, M.J. & Edwards, R.A., 2024. Prophages: an integral but understudied component of the human microbiome. Microbial genomics, 10(1). Available at: http://dx.doi.org/10.1099/mgen.0.001166

Submitted or soon to be

Prophages as a source of antimicrobial resistance in the human microbiome

Inferring microbial habitat from function and taxonomy: A machine-learning approach to metagenome classification

Submitted to Microbial Genomics

Minor author papers

The Promise and Pitfalls of Prophages

Philympics 2021: Prophage Predictions Perplex Programs

The human gut virome: composition, colonisation, interactions, and impacts on human health

Phables: from fragmented assemblies to high-quality bacteriophage genomes

Koverage: Read-coverage analysis for massive (meta)genomics datasets

Who bit the boat? New DNA collection and genomic methods enable species identification in suspected shark-related incidents

Introduction

The world is home to countless microbes. They are found everywhere, from the soil to the oceans, and even in the human body. These microbes include bacteria, fungi, viruses, and archaea and come together to form a microbial ecosystem called the microbiome.

Microbiomes play a crucial role in regulating their environments. Every environment, even our own body, has a microbiome. Our microbiome usually lives in symbiosis with us; the flora in our mouth and digestive tract help us digest food and produce useful chemicals (Guarner and Malagelada 2003), When healthy and in balance, they also keep the populations of potentially harmful bacteria in check (Guarner and Malagelada 2003).

Microbes were originally required to be cultured to isolate a useable sample ((Staley and Konopka 1985)). However, only a small percentage of species can currently be cultured, and learning to culture new species takes much time and effort (Lok 2015). This left large gaps in our knowledge of the microbiome known as 'the great plate count anomaly' ((Staley and Konopka 1985)). Since then, improvements in genomic and metagenomic techniques have allowed us to see more of the immense microbial diversity that is present in the world.
One of the largest components of the microbial world is one of its smallest members, the viruses. With approximately 10^{31} viruses on Earth, they outnumber the bacteria in the microbiome by a large margin (Breitbart and Rohwer 2005a).

The majority of these viruses are bacteriophages (hereafter called phages), which infect bacteria instead of eukaryotes. Phages have two main lifecycles, the lytic and the lysogenic life cycles. During both lifecycles, the phage begins by infecting its bacterial host; the next step is where they begin to diverge. Lytic phages act much like viruses most people are familiar with; they hijack their host's replication mechanisms and make more virions before lysing their host and releasing the new viruses into the environment to find new hosts.

Lysogenic phages, also referred to as temperate phages, instead integrate into their host's DNA. As the bacteria grow and replicate, so too are they replicated. While integrated into their host, they are referred to as prophages, and the whole cell becomes a lysogen. The prophage remains within its host until conditions change, such as the host becoming stressed, at which point, it reverts to the lytic lifecycle.

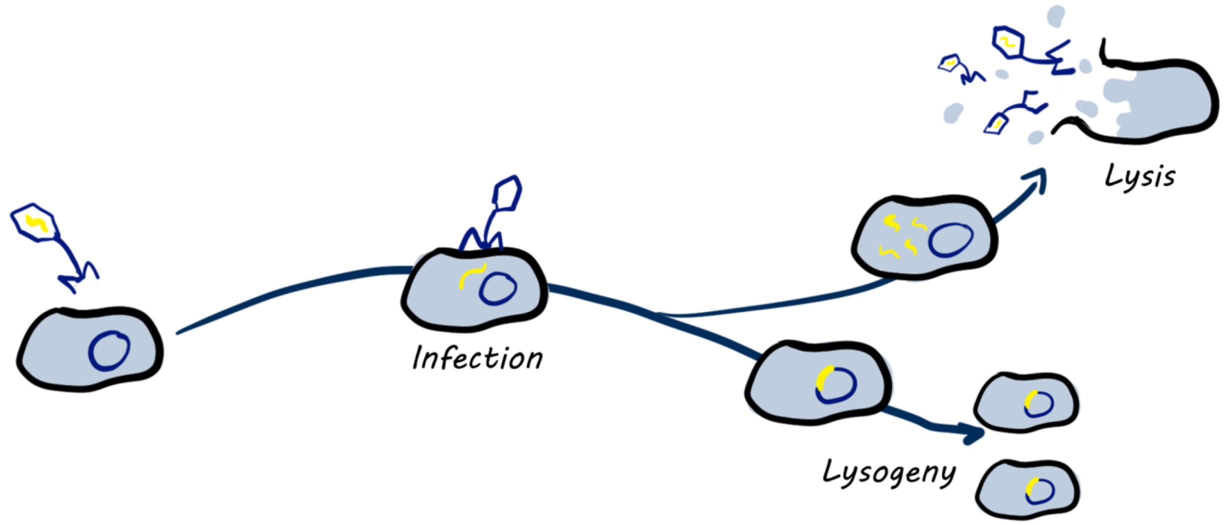

Figure 1: The lytic and lysogenic lifecycles of a phage

While temperate phages are within their hosts, they can provide genes that the bacteria can make use of. Much attention has been focused on the toxins that can be provided to otherwise harmless bacteria. Escherichia coli is a common member of the human microbiome, but when infected by a specific prophage, it can produce a Shiga-like toxin and cause food poisoning. Diseases such as diphtheria and cholera are also caused by phage-encoded toxins; without their phages, the host bacteria cannot cause the devastating diseases they are known for.

Phages can also facilitate the transfer of genes between bacteria. Recombination events can allow phages to acquire genes from their hosts that they can in turn, pass onto other bacteria, slowly moving genes around the microbiome (Morris et al. 2008; Hendrix et al. 1999). There are thought to be up to 10^{24} phage-mediated transfer events per second (Chibani-Chennoufi et al. 2004). While it is common knowledge that bacteria such as *Escherichia coli* and *Vibrio cholerae* get their toxins from the prophage integrated into their genome, there is less agreement on whether phages can provide antimicrobial resistance genes to their hosts, with many studies saying they are common among prophages (Kondo, Kawano, and Sugai 2021; Eugen Pfeifer, Bonnin, and Rocha 2022; Colomer-Lluch, Jofre, and Muniesa 2011) and many saying they are not, or that previous estimates were overstated (Enault et al. 2017; Billaud et al. 2021).

Antibiotic resistance is a major threat to public health, and a lot of work is put into understanding how the resistance is spreading and what must be done to curb it. Phages are often touted as a potential solution to the problem of antibiotic resistance, but evidence also suggests that they may be contributing to the spread of the problem.

In the 70+ years since the discovery of the first prophage (Lederberg and Lederberg 1953), our knowledge of phage biology and their place in the microbiological community has grown, however, there is still much that is unknown. My thesis aims to delve into some of those unknown regions.

First, I aim to focus on the human microbiome and the temperate phages found within bacterial genomes from different regions of the human body, determining how they vary in abundance across the human microbiome. I also aim to learn what genes they may provide to their hosts. Due to the significant changes in the environment across the human body, I hypothesize that the amount of prophage DNA in the bacterial genomes will change significantly as the conditions of some body sites will be more conducive to lysogeny than others.

To do this, I used genomes from the online database GenBank. GenBank is a genomic database where scientists from around the world deposit their sequence data. This allows us to get a wide variety of bacterial genomes from a wide variety of people. However, genomic data uploaded to databases such as GenBank, both genomes and metagenomes, can be lacking in accompanying metadata. Information on where the samples were collected, in particular, can be vague or incomplete, making it difficult to use. It requires a lot of manual sorting through metadata and curating the genomes to be able to use them in a study such as this.

This leads to my third aim, to train a machine learning model to automatically sort environmental metagenomes by their isolation source, lessening the need for manual curation in the future. Metagenomes, particularly whole genome metagenomes have an advantage over genomes for this as they capture the majority of the microbiome. Microbiomes vary within environments due to many factors, but there is a ‘core microbiome’ that varies by environment (Turnbaugh et al. 2009). This core microbiome is usually thought of in terms of taxonomy, but there may be a core set of functional genes as well that could be used to differentiate environments. I hypothesise that both the taxonomic and functional profiles can be used to train an algorithm to classify metagenomes by isolation environment. However, functional genes are known to be highly conserved (Rodriguez-Brito et al. 2010; Turnbaugh et al. 2009; Lloyd-Price, Abu-Ali, and Huttenhower 2016), so I aim to compare the functional and taxonomic profiles to determine which core can be best used to train the algorithm.

My work provides a greater understanding of the human phageome, how it varies with common factors such as human health and area of the body. Currently, most of the focus in the phageome space is on the lytic phages, so this will help to fill that knowledge gap. My work will also help alleviate the problem of lacking isolation source metadata for metagenomes. This should allow for more of the deposited metagenomes to be used for future work and lessen the amount of manual curation required for studies that use large amounts of data from genomic databases.

Review Chapter: How metagenomics transformed our understanding of bacteriophage research

Paper declaration

Laura K Inglis planned the paper, found and analysed the literature and wrote the document.

Robert A Edwards provided editing and assistance with the overall concepts and provided advice on its contents.

This paper was published in MDPI Microorganisms on August 19th 2022.
DOI: 10.3390/microorganisms10081671

Abstract

The microbiome is an essential part of most ecosystems. It was originally studied mostly through culturing but relatively few microbes can be cultured so much of the microbiome was left unexplored. The emergence of metagenomic sequencing techniques changed that and allowed for research into the microbiomes of all sorts of habitats. Metagenomic sequencing also allowed for a more thorough exploration of prophages, viruses that integrate into bacterial genomes, and how they benefit their hosts. One issue with using open-access metagenomic data is that sequences added to databases often have little to no metadata to work with, so finding enough sequences to work with can be difficult. Many metagenomes have been manually curated but this is a time-consuming process and relies heavily on the uploader to be accurate and thorough when filling in metadata fields and the curators to be working with the same ontologies. Using algorithms to automatically sort metagenomes based on either the taxonomic profile or the functional profile may be a viable solution to the issues with manually curated metagenomes, but it requires that the algorithm is trained on carefully curated datasets and using the most informative profile possible in order to minimize errors.

Introduction

The microbiome is the microbial component of an ecosystem. It includes bacteria, archaea and viruses and is an essential part of most ecosystems. It can greatly influence human and environmental health, and it can be heavily impacted by human activities.

Microbiomes are usually very diverse, and a microbiome dominated by a few species, a dysbiosis, is often a cause or symptom of a disease state (Mirsepasi-Lauridsen et al. 2018; de la Calle 2017). Similar to macro-scale food webs, bacteria and viruses operate under complex predator/prey or symbiotic relationships. Studying these dynamics can provide insight into disease and provide potential solutions for pressing issues such as antimicrobial resistance.

Historical approaches to studying the microbiome relied on in vitro culture as the first step in isolating a workable sample. However, only a small fraction of microbes could be cultivated, leading to a massive underestimation of the diversity of the microbiome. This source of error is known as the great plate count anomaly (Staley and Konopka 1985). Nowadays, the main approach to studying the microbiome is through metagenomic sequencing, which does not require laboratory culture.

This review will briefly summarise the history of metagenomics and describe microbiome research made possible through the availability of metagenomic sequencing. Then it will explore potential approaches for categorising metagenomes based on their isolation environment and compare the pros and cons of manual versus algorithmic methods for this research. Finally, the review will consider how metagenomic sequencing has improved research into bacteriophages, particularly for determining how these viruses operate within microbiomes.

A Brief Overview of Metagenomics

The term metagenome refers to the sum of the genomes found in a tested sample and was first coined in 1998 (Handelsman et al. 1998). Outside of examining the few microbes that could be cultured, the first attempts to examine the wider metagenome were performed by isolating total DNA from fresh samples, then cloning large DNA fragments into plasmids or bacterial artificial chromosomes (BACs) maintained in microbes like E. coli. High-throughput screening was then used to examine the chemical diversity produced by the clones (Handelsman et al. 1998) or, later, sequencing the cloned fragments (DeLong et al. 2006). The first published metagenomes were sampled from acid mine drainage biofilms in early 2000 (Tyson et al. 2004). Since then, the methods for analysing metagenomes have been consistently improving to the current day. Now we perform metagenomic sequencing using massively parallel sequencing or deep sequencing. This involves sequencing millions of small fragments of DNA, and then recreating the genome by connecting the fragment sequences using bioinformatics analyses (Papudeshi et al. 2017).

There are two main methods for metagenomic sequencing: amplicon sequencing and whole-genome sequencing. Amplicon sequencing detects only the target gene, and usually focuses on the highly conserved 16S gene because of its ubiquity in bacterial genomes. This approach is relatively cheap and simple; however, it is limited in scope: the common 16S gene primers only target some bacteria, and depending on what primers, reference databases, and bioinformatics settings are used different genera can be underrepresented or missed entirely (Abellan-Schneyder, Matchado, and Reitmeier 2021). Viruses, Archaea, and Eukarya go undetected by this method. Whole-genome sequencing, as the name implies, sequences all the DNA in a sample. This approach has the benefit of being able to detect viruses and has also been shown to be more accurate in detecting bacteria to the species level (Thompson et al. 2021).

Technological improvements have led to metagenomic sequencing becoming cheaper and more accessible, and it is now being used in many areas of research, from environmental to medical sciences. Since metagenomic sequencing was first applied to diagnose infection in a human patient in 2014 (Wilson et al. 2014), its use as a clinical tool has slowly increased (Lim et al. 2014; Brown, Bharucha, and Breuer 2018). The fact that metagenomic sequencing targets all a sample's genetic material at once means the approach has many advantages over traditional diagnostic methods, which are generally limited to microbes that are well studied (de Vries et al. 2021; Winter and Hegde 2020; Chiu and Miller 2019; Gupta et al. 2019; Hematian et al. 2016; Schmieder and Edwards 2012) and to testing for only a couple of potential pathogens at a time. Metagenomic sequencing is also useful for monitoring pathogens in non-human settings. Recently it has been applied to monitor sewage for the SARS-CoV-2 virus (Crits-Christoph et al. 2021; Landgraff et al. 2021), and has proved to be an effective monitoring tool, detecting SARS-Cov-2 in a community before any pa-tients tested positive (NSW Health, 2022b). Metagenomic sequencing is also applied more broadly to monitor pathogens in human communities, including both bacterial and viral pathogens (Rothman et al. 2020), and to examine or monitor antibiotic resistance genes (Hendriksen et al. 2019; Schmieder and Edwards 2012).

The wider microbiome is also explored with metagenomics (Mirsepasi-Lauridsen et al. 2018; Vandenkoornhuyse et al. 2015; Cho and Blaser 2012). Only 1% of microbes can be cultured (de la Calle 2017), limiting our capacity to understand the entire microbiome using traditional methods. Metagenomics has enabled access to the remaining 99% that was almost entirely unknown to us 20 years ago (Cho and Blaser 2012) and may be used to quickly screen microbiomes for useful functions (Dinsdale et al. 2008). For example, these previously uncharacterized microbes could be important indica-tors of host health (Mirsepasi-Lauridsen et al. 2018; Cho and Blaser 2012; Vandenkoornhuyse et al. 2015) or reveal opportunities to develop new antimicrobials (de la Calle 2017; Hover et al. 2018; Tortorella et al. 2018).

The Human Microbiome Project began with the intention of documenting the microbiome of healthy humans (Cho and Blaser 2012; Turnbaugh et al. 2007), but that is proving to be a challenging endeavour as the taxonomic composition of the human microbiome varies between people (Lloyd-Price, Abu-Ali, and Huttenhower 2016).

Curating Metagenomes

Data curation is essential when working with metagenomes sourced from online databanks. Curation in this context refers to collecting the data and sorting it into useable categories. Metagenomes can either be manually curated by the researchers looking to use the data, or by an algorithm trained to sort the metagenomes based on different features of the sequences. Ontologies are systems for categorising data. Numerous ontologies designed for many different types of data can be found on the online repository www.bioontology.org (NCBO Bioportal, 2022). One ontology for categorising metagenomic data is FOAM, Functional Ontology Assignments for Metagenomes which uses Hidden Markov Models to classify gene functions (Prestat et al. 2014). Some other biological ontologies include the Biological Collections Ontology (Walls et al. 2014), or Interlinking Ontology for Biological Concepts (Kushida et al. 2017). One of the better-known ontologies for categorising biological samples is ENVO, the environment ontology. ENVO uses a directed acyclic graph (DAG) to categorise meta-genomes based on the environment from which samples are sourced (Buttigieg et al. 2013).

Because “environment” is a vaguely-defined concept that is often made up of many smaller factors, ENVO uses multiple descriptors in its categorization process. For example, the surface of a stone at the bottom of a lake is both a rock and a lake environment. Depending on factors such as the depth of the lake and the water clarity, the microbiome of the rock surface may vary significantly. While the capability to factor in such details is important, categories can end up being split incredibly finely. In extreme cases, any ontology that attempts to account for all environmental details risks becoming overly complex and difficult to work with in terms of manual sorting of samples. Many ontologies avoid this by being more specialised for different research fields such as marine environments (Blumberg et al. 2021). Training an algorithm to sort samples into the finer categories of an ontology would alleviate this difficulty. However automatic curation has its own set of issues that need to be considered, as discussed below.

Issues with Curating Metagenomes

Manually curated genomes are susceptible to human error and rely on the data generator to provide sufficient and accurate information while uploading the sequences. As a result, manual curations often lack critical information and contain mistakes in the metadata provided.
Over the last few years, the number of metagenomes uploaded to databases such as the Sequence Read Archive (SRA) or MGnify has increased exponentially (Mitchell et al. 2020). While this does provide more data for anyone to use the amounts of low quality, contaminated, or mislabelled sequences have also increased. This adds to the amount of work required to manually curate enough genomes. While automated curation of metagenomes could address these issues, this approach requires that the categories have little overlap in the data being analysed by the algorithm, which can prove difficult. The datasets used to train such an algorithm would also need to be carefully curated to avoid contaminated or low-quality sequences, or sequences collected using very different methods. For example, in the SRA database both amplicon sequencing and whole genome shotgun sequencing can be categorised as metagenomic data (Torres, Edwards, and McNair 2017). This can be an issue as they provide different information that could cause issues for an algorithm trying to compare them. An algorithm can successfully differentiate between the two types of data, however, as the partition engine, PARTIE, has sorted many sequences from the SRA into amplicon and WGS datasets (Torres, Edwards, and McNair 2017). Any attempts to automatically curate

metagenomic data from the SRA database will need to take the broad definition of metagenomic data into account when collecting the training data.

Categorising Metagenomes

Distinguishing distinct environment categories for microbiome ontologies is difficult for a number of reasons. For human microbiomes, sequences vary greatly between individuals (Turnbaugh et al. 2009; Yi et al. 2014; Lassalle et al. 2018), between different locations on a single individual (Yi et al. 2014), and even within the same location on a single individual over time (Turnbaugh et al. 2009; Yi et al. 2014; Man, de Steenhuijsen Piters, and Bogaert 2017).

The metagenome of an external environment also changes over time (Fernández et al. 1999) or in response to mi-nor variations (Mahoney, Yin, and Hulbert 2017) in features such as water depth (DeLong et al. 2006). Furthermore, large variations within environments may warrant subdividing the environmental category further. For example, previous attempts to test the ability of algorithms to assign metagenomes to environments had high rates of incorrect assignments in some categories, particularly the “human respiratory” category (Burke, 2019). This category included the lungs, mouth, nose, throat, and a few dental plaque samples. The large error makes sense, as both anatomically and biologically the mouth and the lungs are entirely different environments and cultivate different microbiomes (Man, de Steenhuijsen Piters, and Bogaert 2017).

Depending on how specific the ontology needs to be, separate categories for healthy and unhealthy sample environments should also be considered, as the microbiomes of healthy and unhealthy humans and animals can vary significantly (Yi et al. 2014; Cuthbertson et al. 2020; Yatera, Noguchi, and Mukae 2018).

Using AI to Curate Metagenomes

Most existing genomic databases do not require uploaders to provide comprehensive sample information (metadata) during the submission process. This results in many sequences being submitted to databases with little to no context. Manual curation often relies on this contextual information. The impacts of these missing details can be broad, as they translate into reducing the pool of sequences that can be used for other projects. For example, a water sample without metadata detailing the depth at which the sample was taken will be unable to be used to answer a question about the difference in microbiomes at various water depths as it cannot be assigned to any relevant category.

Artificial intelligence (AI) algorithms hold considerable promise in metagenomics, including for curation. For example, differences between metagenomes could be used to automatically and accurately curate based on the environment from which samples are isolated. However, this sort of decisionmaking activity will only happen once enough existing data is categorised and labelled appropriately.

It is also important to consider hierarchy of labelling: for example, how should the choice of labelling categories be managed for samples collected from different places on the same person? We previously used a random forest algorithm to sort metagenomes into categories based on their environmental source with 78.5% accuracy. There is potential to improve that accuracy by either refining the training data or categories, and/or by changing the type of information the algorithm uses to curate the metagenomes.

Automated curation could radically change how researchers do microbiome research. Sequences that are unusable for large scale data integrations due to insufficient metadata could become available if a pipeline for automatic curation became available.

Functional diversity vs taxonomic diversity

Different types of data can be used for curation. For example, we used the taxonomic profile of metagenomes to curate them. This approach is tempered because taxonomic make-up can vary greatly in different variations of the same environment. For example, the human microbiome varies greatly between people and can be influenced by diet (Lassalle et al. 2018), weight (Turnbaugh et al. 2009), medication (Falony et al. 2016), and many other factors(Lassalle et al. 2018; Camarillo-Guerrero et al. 2021).

Another approach to curation is to use the functional profile. These genes are often relatively conserved across microbiomes within similar environments, with more similarity in the relative abundance of functional genes compared to bacterial phyla in the microbiome of different environments (Rodriguez-Brito et al. 2010; Turnbaugh et al. 2009; Lloyd-Price, Abu-Ali, and Huttenhower 2016). It is possible a ‘core microbiome’ of functional genes that is similar across groups of environments may exist.

Overall, while taxonomic differences between people’s microbiomes can differentiate between samples on a personal scale, use of the functional profile for larger-scale metagenome curation could be more accurate. However, further research is required to compare the two curation methods to consolidate this hypothesis.

Phages and the Microbiome

Access to metagenomics has transformed how we study the microbiome, including research focusing on bacteriophages. Commonly referred to as phages, these viruses infect bacteria and are an important part of almost every microbiome. Individual phages often have a very limited host range, although almost every species of bacteria have phages that prey on them (Aziz et al. 2015) Phages can occupy either a predatory or a symbiotic role in the microbiome. Phages that predate bacteria are known as lytic phages, and those that play a more symbiotic role are referred to as lysogenic phages.

Lytic phages infect bacteria and hijack their replication machinery to make more copies of themselves, eventually killing the bacteria (Howard-Varona et al. 2017). This form of predation is thought to kill approximately 20% of the ocean's microbial biomass each day (Suttle 2007). The targeted application of lytic phages has been considered as a potential alternative or complement to antibiotics (Romero-Calle et al. 2019; Tagliaferri, Jansen, and Horz 2019). Lysogenic phages integrate into the bacterial host’s DNA and replicate with the host. Referred to as prophages, they can remain within hosts for thousands of generations (Clokie et al. 2011).

Host bacteria can utilise genes provided by prophages to gain a variety of benefits. For example, many prophages provide protection against infection by other phages (Zinder 1958; Mavrich and Hatfull 2019), including prevention of predation by lytic phages. Prophages can also provide genes conferring resistance to antimicrobial drugs, protection from immune responses, or production of toxins (Castillo et al. 2018; Sweere et al. 2019; Bielaszewska et al. 2012; von Wintersdorff et al. 2016). For example, a prophage encodes the toxin genes that enable Vibrio cholerae to cause the life-threatening disease cholera (Waldor and Mekalanos 1996).

Piggyback-the-Winner and Other Hypotheses

While some phages are obligate lytic phages, many prophages are temperate phages – this means they can shift between lytic and lysogenic life cycles when conditions change. Three main hypotheses have been proposed to explain why temperate phages switch between lytic and lysogenic life cycles: Piggyback-the-Winner (PTW) (B. Knowles et al. 2016), Piggyback-the-Loser (PTL) (aka refugium hypothesis), and Kill-the-Winner (KTW) (Thingstad 2000). Each hypothesis is characterized by different environmental conditions (Table 1).

Table 1: A brief summary of the different hypotheses for explaining lysogeny rates. Both Piggyback-the-Winner (PTW) and Kill-the-Winner hypotheses (KTW) occur in similar environmental conditions.

Lysogeny hypothesis	Environmental conditions	Level of lysogeny
Piggyback-the-winner	High energy, cells growing	High
Kill-the-winner	High energy, cells growing	Low
Piggyback-the-Loser	Low energy, little growth	High

The Piggyback-the-Winner hypothesis proposes that while host bacterial populations are replicating quickly, the phage uses a lysogenic lifecycle to take advantage of the host's success (B. Knowles et al. 2016). This hypothesis also suggests the phage's lytic lifecycle is triggered when bacteria start to become stressed.

The Kill-the-Winner hypothesis looks at the problem from a predator-prey dynamics perspective. It proposes that a lytic phage is more likely to come across a bacterium it can infect if that bacterium is more abundant. Therefore, abundant bacteria are encountering and being infected by lytic phages more often than bacteria that are not growing as robustly (Thingstad 2000).

The Piggyback-the-Loser hypothesis suggests the opposite of the PTW hypothesis. It argues that when the host is growing slowly it is more beneficial to remain in a lysogenic state, as it is less likely that the phages produced through lysis will find a new host (Silveira, Luque, and

Rohwer 2021). This hypothesis has also been called Piggyback-the-Persistent (Paterson et al. 2019) as the prophages were found in host species that persistently occurred at low abundance.

Each of these hypotheses explains phage-host dynamics in studied environments. For example, PTL occurs in some regions of the deep sea or polar oceans (Silveira, Luque, and Rohwer 2021; Brum et al. 2016) but PTW dynamics occur in some coral reef regions (Silveira and Rohwer 2016). However, each hypothesis is an incomplete explanation of microbi-ome dynamics on its own. PTW and KTW also appear to be conflicting hypotheses (Figure 2) as they both have been shown to occur in high-energy environments but have the opposite results (Silveira, Luque, and Rohwer 2021; Silveira and Rohwer 2016; Thingstad 2000). However, a recent study has suggested a way for these opposing hypotheses to co-exist (Silveira, Luque, and Rohwer 2021).

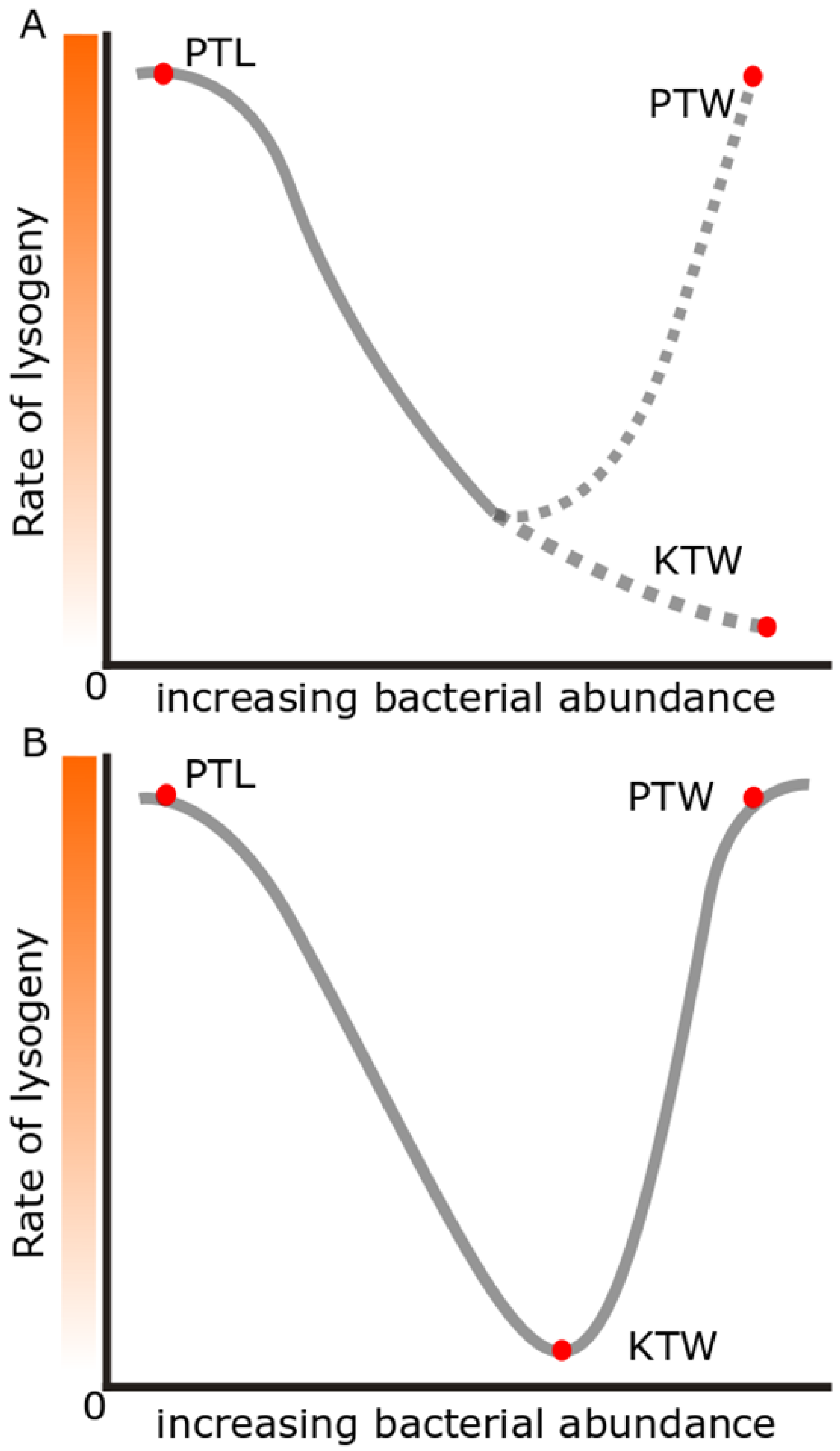

Figure 2: Plotting the hypothetical rates of lysogeny between the PTL and KTW/PTW dynamics shows how conflicting the hypotheses appear to be (A). However, a recent hypothesis [68]

suggests that the previous hypotheses can coexist and PTW overtakes KTW dynamics at high enough bacterial abundances (B).

Silveira et al. propose that a combination of different mechanisms results in lysogeny being promoted at both very high and very low host abundances (Silveira, Luque, and Rohwer 2021). Their new hypothesis suggests that the effects of microbial density, community diversity, host-virus interactions, host metabolism, and the environment combine to either promote or repress lysis depending on which factors are more prominent. Overall, the main factors at play in determining whether a phage will be lytic or lysogenic are thought to be (1) chance of co-infections, and (2) host metabolism.

Co-infection is one of the best predictors for lysogenization (Silveira, Luque, and Rohwer 2021) as it increases the abundance of phage repressors in the host cell, preventing lysis (Cheng et al. 1988). Slower host growth increases the window for co-infections to occur, which increases the potential for co-infections to promote lysogeny, supporting the PTL hypothesis. Conversely, hosts that are growing quickly have a shorter window for co-infection. However, eventually, the abundance of phages in the environment will increase to the point where co-infections happen quickly, leading to increased lysogeny at high host abundances which supports the PTW hypothesis. Co-infection is also affected by community diversity. When the community is less diverse phages are more likely to encounter bacteria that they can infect, which increases the probability of coinfections (compared to the lower probability of coming across an appropriate host if it is a relatively rare species).

Like co-infection, cell metabolism favours lysogeny at the extreme ends of the spectrum. In a low-energy environment where the hosts are growing slowly, cell starvation causes a signalling cascade. This increases the abundance of phage repressors (Cheng et al. 1988), favouring lysogeny. When the hosts are not starving, they increase their stores of ATP. This allows the host proteases to degrade the phage repressors, and increase the chance of lysis (Cheng et al. 1988). However, in a very high-energy environment with limited oxygen, bacteria switch to a less efficient form of metabolism (e.g. fermentation) leading to low ATP/high NADPH levels. This prevents the host proteases from destroying the phage repressors and thereby limiting lysis.

As influencing factors, co-infection and metabolism work together to favour lysogeny at both very high and very low bacterial abundances. Silveira and colleagues suggest the rate of lysis peaks in environments with a bacterial concentration of 10^6 per millilitre (Silveira, Luque, and Rohwer 2021). They propose that starvation and large windows for co-infection to occur favour lysogeny in low energy environments, while high bacterial concentrations and inefficient metabolism favour lysogeny in higher energy environments.

Silveira et al. suggests approximate bacterial concentrations for different virus-host ratios. In theory, this means that this hypothesis could be tested by counting the prophages, viruses, and bacteria in different environmental metagenomes and modelling environmental carrying capacity from the number of prophages in each sample (Silveira, Luque, and Rohwer 2021). A further test is to experimentally manipulate the virus-to-microbe ratio (VMR), for example by adding phages or removing bacteria (e.g. by filtration).

Prophage Abundances in Different Environments

Environments vary greatly in the amount of bacterial growth they can sustain, called the carrying capacity. Each situation exists at a unique point on Silveira et al model (Silveira, Luque, and Rohwer 2021). For example, there are likely to be relatively few prophages per cell in low-energy environments such as the deep ocean, though likely still enough to trigger lysogeny, because prophages take resources to maintain. In higher energy environments — such as the ocean's surface or a coral reef — lytic growth is also the dominant lifestyle, so very few prophages would be expected in bacterial genomes. In highly productive environments, such as coastal waters and estuaries, the environments favour lysogeny. However here, the energy cost to maintain the prophages is not as big of a limiting factor (because energy is plentiful), so the number of prophages per cell should theoretically be much greater than in low energy environments.

Numerous studies have examined phages across many different environments, including in various animals (Kim and Bae 2018), in humans (Cazares et al. 2021; Edwards et al. 2019; Dutilh et al. 2014; Oh et al. 2019), from most aquatic environments (Castillo et al. 2018; Motlagh et al. 2017; L. McDaniel et al. 2008; L. D. McDaniel et al. 2014; Parmar et al. 2018), and from soil (Stokar-Avihail et al. 2019; Gómez et al. 2015; Braga et al. 2020). While many of these papers examine phages from single environments, there are comparatively few studies that compare phages from different environments in one analysis. Of these, many only compare a couple of environments at once (Parmar et al. 2018; Rezaei Javan et al. 2019; Oh et al. 2019; Castillo et al. 2018). As different prophage detection methods can produce slightly different results (Roach et al. 2021), comparing phages from environmental samples that have been collected and analysed with various methods could produce inaccurate results.

Identifying Prophages

The first step in analysing prophages is to identify their sequences in the bacterial genome. The approaches for doing so broadly separate into two main categories: experimental induction, or computational examination of a bacterial genome for prophage-like regions.
Induction involves causing lysis in the bacteria, which releases the phage into solution. It is one of the oldest methods for detecting phages and uses chemicals that signal to the phage to switch to a lytic lifecycle, destroying the host bacteria and releasing the phages into the solution. The issue with this method is that it can only be used on culturable bacteria and incomplete induc-tion will cause the phage-host ratio to be skewed. Although we have discovered many chemical inducers, there are also plenty of prophages where we have not identified the signals they are sensing, and so cannot induce them.

Computational examination of bacterial genomes first requires that the bacterial DNA is sequenced, then the sequences can be examined for genes that belong to prophages either manually or by using programs such as PhiSpy (Akhter, Aziz, and Edwards 2012) and many others (Roach et al. 2021). The accuracy of these predictions is affected by the quality of the sequence assembly. As genomes are being assembled small, overlapping segments of DNA sequences are pieced together to form larger sequences, or contigs. A genome with fewer contigs is more complete than one with more contigs. A genome with many contigs generally has a lot more predicted prophages than genomes with few contigs and the rate of false positives increases as the number of contigs in a genome increases as it is more likely for a smaller segment of DNA to appear the same as a prophage by chance alone. A comparison of 8 different prophage prediction tools showed that each package has its strengths and weaknesses (Roach et al. 2021). Some sacrificed speed and efficiency for accuracy while others tried to find a balance. Some had good precision but that necessitates an increased chance of false-positive results, while others increased recall performance which comes with more false positives. The pros and cons of each program need to be evaluated by users before deciding which is most appropriate for each project.

Conclusions

Metagenomic sequencing is useful for finding prophages from many different environmental conditions, but many genomes are added to databases without the inclusion of comprehensive metadata. Being able to automatically sort these sequences into an environmental ontology would allow for these sequences to be useful in future projects, but we need considerably more high-quality data to determine how best to sort these sequences.

Prophages play significant roles in the microbiomes of many species and in different environments. Phages can protect their host from deadly infections, and give their host access to beneficial genes such as antimicrobial resistance or toxin production. The way phages interact with their host changes depending on the environment. Researchers have conducted many studies on phages from various environments, and developed hypotheses regarding what factors influence survival strategies, such as the lytic/lysogenic decision. However, there is still much more to learn about how prophages interact with their hosts under different conditions. Learning more about metagenomes and prophages could provide many insights into human and environmental health, while getting a better understanding of what a healthy microbiome should be may enable us to detect changes more quickly or accurately in microbiomes that could be a sign of disease.

Results Chapter 1: Prophages: an integral but understudied part of the human microbiome

Paper declaration

Laura K Inglis assisted in conceiving the manuscript. They also generated and analysed the data and wrote the manuscript.

Michael J Roach provided technical advice on the programs and coding languages used for data analysis. They also edited the final document and provided advice on its layout and contents.

Robert A Edwards conceived the study. They also provided the initial genomes and prophage predictions, assisted and provided advice regarding data analysis, edited the final document, and provided advice on its contents.

This paper was published in Microbial Genomics in January 2024.
DOI: 10.1099/mgen.0.001166

Abstract

Phages integrated into a bacterial genome–called prophages–continuously monitor the health of the host bacteria to determine when to escape the genome, protect their host from other phage infections, and may provide genes that promote bacterial growth. Prophages are essential to almost all microbiomes, including the human microbiome. However, most human microbiome studies focus on bacteria, ignoring free and integrated phages, so we know little about how these prophages affect the human microbiome. To address this gap in our knowledge, we compared the prophages identified in 11,513 bacterial genomes isolated from human body sites to characterise prophage DNA in the human microbiome. Here, we show that prophage DNA is ubiquitous, comprising an average of 1-5% of each bacterial genome. The prophage content per genome varies with the isolation site on the human body, the health of the human, and whether the disease was symptomatic. The presence of prophages promotes bacterial growth and sculpts the microbiome. However, the disparities caused by prophages vary throughout the body.

Introduction

The human microbiome is a complex ecosystem of microbes that inhabit every part of the human body. Most body sites typically contain a multitude of microbes resulting in diverse ecosystems. In contrast, sites in the human body dominated by one or a few species—dysbiosis—are often an indicator of disease (Mirsepasi-Lauridsen et al. 2018; Inglis, Roach, and Edwards 2024; Brüssow 2023).

While the term ‘human microbiome’ may evoke the mental image of the human body as a single environment, the body contains many different niches. Environments such as the skin, stomach, lungs, and mouth are so different from each other that combined, they have an extensive range of bacterial concentrations, very high species richness, and widely varying species diversities. Most of the total microbial biomass in humans and other mammals resides in the gut, and that organ’s metabolism contributes to the animal’s overall thermogenic energy expenditure (Riedl et al. 2021). Other body areas have orders of magnitude lower bacterial concentrations than the gut (Sender, Fuchs, and Milo 2016). The gut microbiome is also highly diverse (A. N. Shkoporov and Hill 2019), while others, such as the lung microbiome, are dominated by only a few groups (Yagi et al. 2021).

Bacteriophages (phages) are viruses that infect bacteria found in almost every environment (Ben Knowles et al. 2017). In the ocean, they kill approximately 20% of the microbial biomass daily (Suttle 2007), but their role in sculpting and controlling most microbiomes, including the human microbiome, is underestimated. A few clades dominate gut phages, but phages infect almost every bacteria in the human gut (Dutilh et al. 2014; Camarillo-Guerrero et al. 2021). There are two main kinds of phages: virulent phages, where the phage infects the host bacteria, replicates, and lysis the bacteria to release phage progeny, and temperate phages, which may either choose a lytic lifecycle or choose to integrate into the host’s DNA and replicate alongside the host until the phage senses suitable conditions for the switch to lytic replication (Lwoff 1953). Prophages are temperate phages integrated into their host’s genome, and the resulting host bacteria is a ‘lysogen’. Almost every bacterial species has temperate phages, although much is still unknown about both lytic and temperate phages.

Prophages confer various benefits to their host through lysogenic conversion. The most common is superinfection exclusion: the protection of the lysogen against other phage infections (Zinder 1958; Mavrich and Hatfull 2019). Many prophages also express virulence genes or toxins that promote the growth of the lysogen (Castillo et al. 2018; Sweere et al. 2019; Bielaszewska et al. 2012; von Wintersdorff et al. 2016; Waldor and Mekalanos 1996). Some examples of prophages providing the toxins that allow their bacterial host to cause human disease include Shiga toxin-producing E. coli, cholera and diphtheria.

The genetic switch that controls the decision to integrate into the host or replicate and kill the host has been at the centre of many molecular biology breakthroughs (Ptashne 2004), such as the Nobel prizes in Physiology and Medicine in 1965 and 1969, which were awarded for discoveries regarding the viral synthesis and replication mechanisms respectively.

Many factors affect the outcome of that decision, including the concentration of bacteria, the diversity of bacterial and phage species, the redox potential of the cell (i.e. the metabolic efficiency of the bacteria), the presence of other phages, and signalling peptides that phages produce to communicate with each other (Erez et al. 2017; Silveira, Luque, and Rohwer 2021). Here we explored the variation in prophage composition across the human body, investigated how much of the bacterial DNA in the human microbiome is provided by prophages, demonstrated how this diverges across the different areas of the human body, and we quantified whether diseased microbiomes and disease-causing bacteria have different prophage abundances than the microbiomes of healthy people.

Methods

All 949,935 publicly accessible bacterial genomes (as of 1st June 2022) listed in the dataset "NCBI Genome Assemblies Summary Archive 20220601" (key resources) were downloaded from GenBank for analysis on Flinders University HPC cluster (University 2021). PhiSpy (Akhter, Aziz, and Edwards 2012; McNair et al. 2019) was used to analyse the genomes and detect prophage genomes in the bacterial DNA (McKerral et al. 2023) as it is currently the best-performing prophage prediction tool (Roach et al. 2021). All the predicted prophages are available from FigShare (Key Resources "Prophage predictions").

The data was filtered to remove metagenome-assembled genomes, low-quality genomes with more than 50 contigs, duplicate genome sequences, and genomes not isolated from humans using the NCBI Genome Assemblies Summary (Key Resources "NCBI Genome Assemblies Summary Archive 20220601"). We manually sorted the remaining samples into categories and subcategories based on the area of the body from where they were isolated and the human host's health according to the PATRIC metadata (Key Resources "Archive of the PATRIC Metadata from 20220601") (Inglis and Edwards, 2023).

After filtering, 20,573 unique genome accessions remained. Over half--11,513 genomes--came from bacteria associated with different human body areas. We separated those into 32 categories, with 3-2,970 samples per category. Approximately half of the genomes from human-associated bacteria, 6,844, could be categorised by the human host's health. We provide this data as Key Resources "Prophages in humans".

SPSS was used for statistical analysis. The Kruskal-Wallis test was used to compare the categories of genomes associated with different human body areas.

We found bacterial concentrations for some of the categories by searching the literature. If we found multiple different estimates, we used an average of the concentrations to compare the prophage abundances with the bacterial concentrations.

For the genomes we could categorise by human health, we analysed multiple groups using SPSS. We compared the healthy, symptomatic, and asymptomatic groups within each category with either a Kruskal-Wallis test for categories with genomes in all three groups or Mann-Whitney-U tests for the skin and gut samples that only had two variables. We combined the categories and compared healthy, symptomatic, and asymptomatic groups with a Kruskal-Wallis test. Once we identified significant differences between healthy and symptomatic groups, we reanalysed the categories using only the healthy samples to determine whether the relative prophage abundances changed.

Results

The GenBank genome assembly database contains almost 1 million publicly accessible bacterial genomes, but most are highly fragmented. However, we identified 11,513 genomes from bacteria that could be associated with different areas of the human body. These samples came from various people with different geographical locations, lifestyles, ages, diets, and conditions. We identified prophages in these genomes and calculated the percentage of the prophage sequence genomes for each sample source location (figure 3).

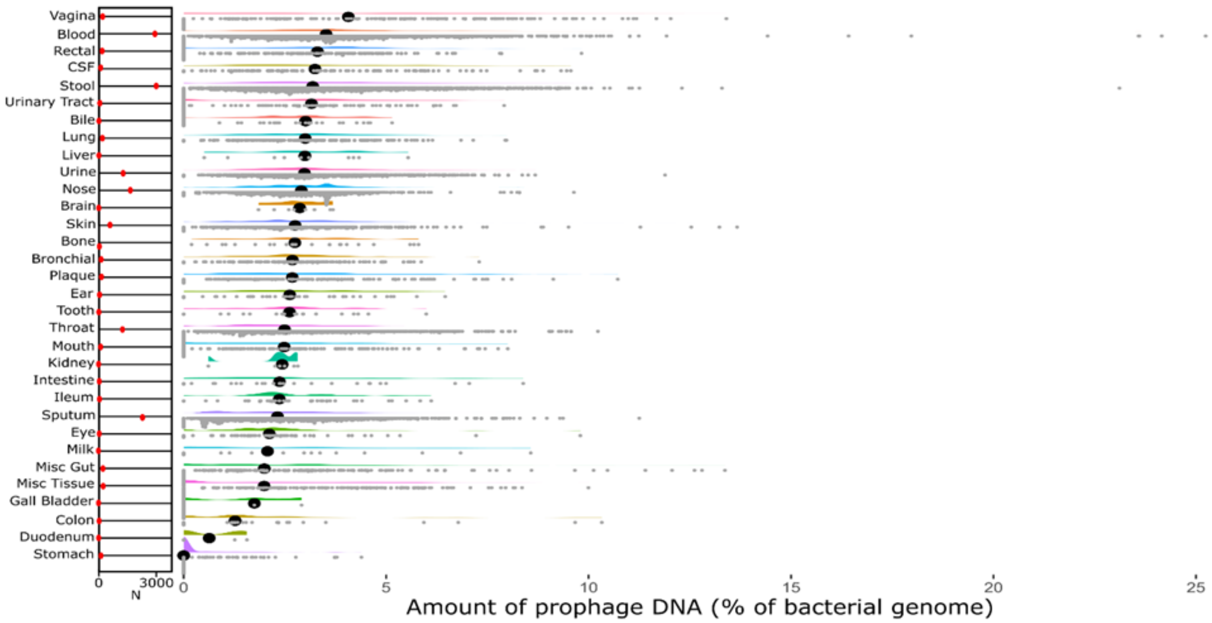

Figure 3: Raincloud plot of all 32 categories ordered by the median percentage of phage DNA from highest to lowest. The red markers on the left show the number of genomes in each category.

There is a large variation in the proportion of prophage DNA within and between different body sites. The median prophage DNA content ranges from 0-5% of the bacterial genome. While many areas have a median prophage DNA content closer to 2-3%, there is a sizeable difference between body sites, especially at the extremes--vagina and blood at the high end, with 4-5% prophage content, and duodenum, and stomach at the lower end, with close to 0% prophage content.

The vaginal samples had the highest median proportion of prophage DNA. A single genus of bacteria dominates the healthy vaginal microbiome—Lactobacillus—which produces antimicrobial compounds that control other bacterial populations (Chen Chen et al. 2017). The vaginal microbiome is also dense, containing 10^{10}-10^{11} bacterial cells (Chen Chen et al. 2017). Both high bacterial concentrations and a microbiome dominated by a few species are two factors previously shown to correlate with higher rates of lysogeny (Silveira, Luque, and Rohwer 2021).

Conversely, the stomach had the lowest average proportion of prophage DNA. No prophages could be detected in most (76.67%) of the genomes from bacteria isolated in the stomach. The stomach is significantly different from almost every other body site and is one of the most extreme environments in the human body. A handful of genera dominate, and the bacterial concentrations are relatively low, in the order of 10^3-10^4 bacteria (Nardone and Compare 2015). Overall, it is quite the opposite of the vaginal microbiome.

Respiratory and gastrointestinal tracts

Narrowing our focus to the respiratory and gastrointestinal tracts allows us to examine how the microbiome changes as the conditions change in transit from mouth to anus. We juxtaposed the distributions of prophage DNA with the GI and respiratory systems for visual assessment (figure 4), and we performed Kruskal-Wallis statistical tests to determine if these distributions were significantly different.

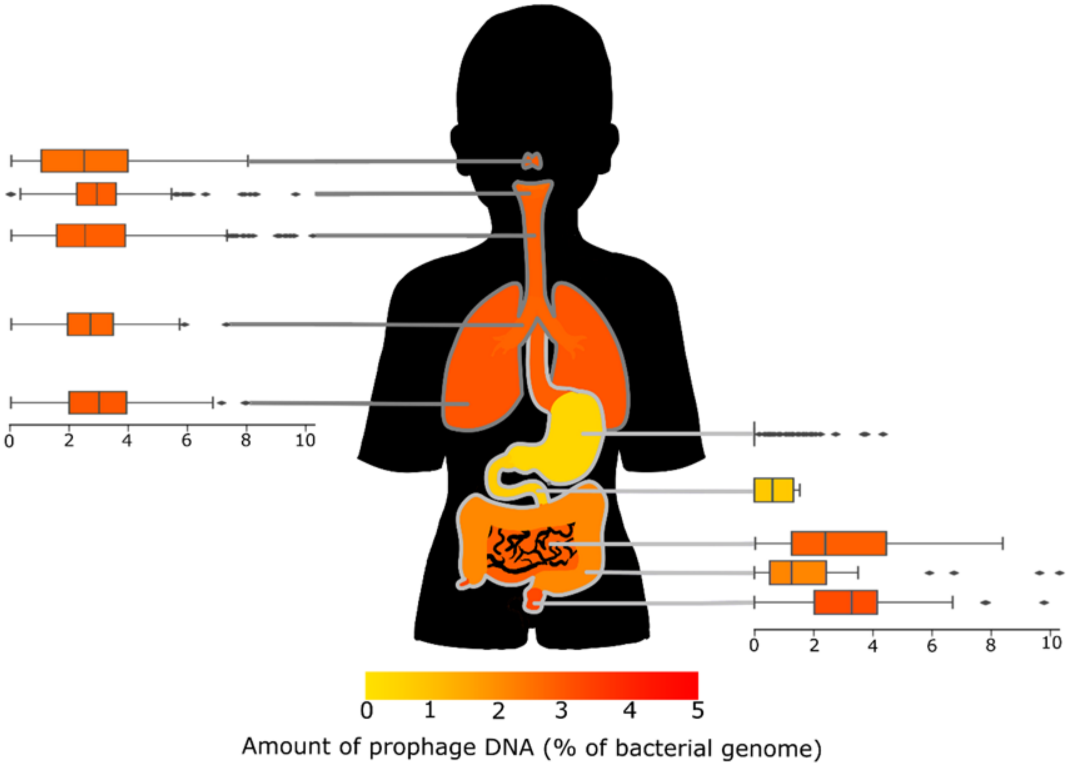

Figure 4: Box plots showing the amount of prophage DNA in each area of the respiratory and gastrointestinal tracts. The figure is coloured by the average proportion of prophage DNA as displayed in the scale below.

The different sections of the respiratory tract have similar distributions of prophage DNA, reflecting similar conditions. The areas also connected to the gastrointestinal tract--the mouth and throat--were significantly different to the lungs (p=<0.05), while the nose was only significantly different to the throat (p=<0.005).

The lungs had the highest prophage DNA, while the mouth had the lowest. The microbiome of the respiratory tract changes with the age of the host, becoming more diverse as the human matures from infant to adult (Kumpitsch et al. 2019). Since the abundance of temperate phages correlates with microbial diversity, there may be fewer temperate phages in the respiratory tract of older people. The overall bacterial concentration estimates suggest the mouth has more bacteria than the lungs (Sender, Fuchs, and Milo 2016; O'Dwyer, Dickson, and Moore 2016), which is generally conducive to higher rates of lysogeny. However, we observe the opposite trend with lung-isolated genomes having higher proportions of prophage DNA.

Conversely, the gastrointestinal tract has a much wider range of bacterial concentrations and does not follow a linear order like the respiratory tract. The distinct areas of the gastrointestinal tract have much more varied environments, and the prophages appear to follow the Piggyback-the-winner model, with more prophages in areas of the body with higher bacterial concentrations, such as the stool or mouth. In contrast, the more hostile environments like the stomach have less prophage DNA per bacteria.

Overall, bacterial concentration alone does not adequately explain the proportion of prophage DNA, and we must look to other factors to explain our results. Generally, bodily fluid samples (e.g. breast milk, urine, and blood) had lower bacterial concentrations (Breitbart and Rohwer 2005b; Païssé et al. 2016; Martín et al. 2012; Kogan et al. 2015) but higher prophage concentrations than the other body sites.

Effects of host health

Many of the samples were clinical samples which likely influenced the results, for instance, bacteria that dominate in dysbiotic microbiomes and specific disease-causing bacteria. To examine if these clinical samples exhibit different lysogenic profiles, we split the samples into groups based on whether the sample metadata listed the human as healthy, having various ailments (including diseases caused by specific bacteria and other ailments involving various bacteria/viruses), or asymptomatic. We independently assessed the samples by body site when investigating differences in the number of prophages per genome and determined significant differences using a Kruskal-Wallis test (figure 5).

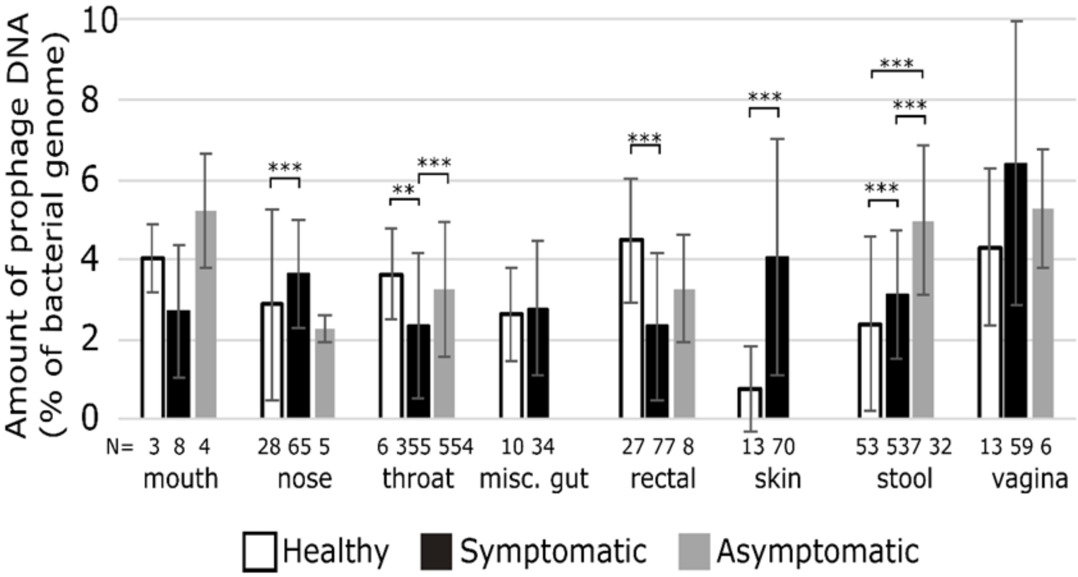

*Figure 5: Each area of the human body with at least three genomes and described as sampled from healthy people. The number beneath each column indicates the number of samples in each group. Error bars represent one standard deviation, while asterisks represent significant differences (**= p=<0.005, ***= p=<0.001).*

Only eight body sites had samples from healthy individuals (figure 5). The nose, skin, and stool samples all had more prophages in samples from symptomatic patients than from healthy individuals, suggesting that prophages may contribute to disease at these sites. In contrast, the throat and rectal samples had fewer prophages in symptomatic individuals than in healthy people.

Both throat and stool samples had significantly more prophages in asymptomatic individuals than in healthy ones. Typically, patients are classified as asymptomatic when they suffer from a disease but are not currently experiencing symptoms. The difference in prophage abundance could suggest that prophages are either decreasing the virulence of their hosts in these areas or providing greater survivability so that once the illness clears, predominantly lysogens remain. Related samples, such as from the lower gastrointestinal tract (stool and rectal) or the respiratory system (nose, mouth, and throat), did not always show similar patterns of prophage abundance. The bacterial species, types of illnesses, or the different types of tests used at different sites could eliminate patterns between body sites.

Discussion

Different body areas had different numbers of bacterial genomes associated with them (figure 3; red markers), and almost every area of the body had a considerable variation in the number of prophages. This variation could primarily be due to the differences in microbiome bacterial compositions between the individuals sampled. While there is evidence for a ‘core microbiome’ of functional genes (Turnbaugh et al. 2009; Lloyd-Price, Abu-Ali, and Huttenhower 2016;

Rodriguez-Brito et al. 2010), the taxonomic makeup of the microbes between individuals varies significantly (Andrey N. Shkoporov et al. 2019). Many factors affect the composition of our microbiomes, including diet, medications, overall health and fitness, and weight (de la Calle 2017; Lassalle et al. 2018; Falony et al. 2016; Turnbaugh et al. 2009; Camarillo-Guerrero et al. 2021; Mirsepasi-Lauridsen et al. 2018).

Some of these bacteria enter our bodies through our mouth and nose as we eat, drink, and breathe, and some of these bacteria find their way down further into the respiratory or gastrointestinal systems to supplement our microbiomes (O'Dwyer, Dickson, and Moore 2016). This results in the connected microbiomes, such as the mouth and throat, having similar compositions, and our results showed that while the prophage concentrations often followed a similar pattern, there were a few outliers. For example, the stomach differed from the rest of the gastrointestinal tract in every way.

There are two main hypotheses regarding lysogeny rates: piggyback-the-winner and piggyback-the-persistent. Piggyback-the-winner suggests that the microbiomes with high bacterial concentrations are more likely to favour lysogeny (B. Knowles et al. 2016), while the piggyback-the-persistent suggests the opposite (Paterson et al. 2019). The lungs have a lower bacterial concentration, yet a relatively high amount of prophage DNA suggests that it follows the persistent strategy of piggyback-the-persistent, while there were less apparent patterns in the gut.

Because different bacteria have different prophages and lower bacterial diversity is associated with higher rates of lysogeny (Silveira, Luque, and Rohwer 2021), external factors that affect the makeup of the human microbiome could affect the number of prophages in each bacterial genome. Human health is perhaps the most critical factor that influences our microbiomes. Illnesses and generally poorer health are often associated with less diverse microbiomes, particularly in the respiratory and gastrointestinal systems (de Steenhuijsen Piters et al. 2016; de la Calle 2017; Mirsepasi-Lauridsen et al. 2018). However, we found wide variation in the prophages between healthy and symptomatic samples in different body sites.
There is wide variation in the amount of prophage DNA in the bacterial genomes of the human microbiome. Categorising the samples by body site revealed patterns in prophage abundance. Areas connected or with similar environments often had a similar prophage distribution. The respiratory tract, which has a lower microbial load, appears to follow the Piggyback-the-Persistent scenario, while the microbially rich gastrointestinal tract follows the Piggyback-the-Winner scenario. The microbiome impacts human health, and vice-versa, and a few body sites showed significant differences in prophage abundance in health and disease.

Conclusions

Metagenomic sequencing is useful for finding prophages from many different environmental conditions, but many genomes are added to databases without the inclusion of comprehensive metadata. Being able to automatically sort these sequences into an environmental ontology would allow for these sequences to be useful in future projects, but we need considerably more high-quality data to determine how best to sort these sequences.

Prophages play significant roles in the microbiomes of many species and in different environments. Phages can protect their host from deadly infections, and give their host access to beneficial genes such as antimicrobial resistance or toxin production. The way phages interact with their host changes depending on the environment. Researchers have conducted many studies on phages from various environments, and developed hypotheses regarding what factors influence survival strategies, such as the lytic/lysogenic decision. However, there is still much more to learn about how prophages interact with their hosts under different conditions.

Learning more about metagenomes and prophages could provide many insights into human and environmental health, while getting a better understanding of what a healthy microbiome should be may enable us to detect changes more quickly or accurately in microbiomes that could be a sign of disease.

Results Chapter 1.5: Prophages and their bacterial hosts

The human microbiome is composed of hundreds of species, with potentially over 500 different species of bacteria just in the oral cavity alone (Zaura et al. 2009), and each bacterial species has their own phages. We know that the taxonomic composition of the microbiome varies by geography, with even the same kind of environment having a different microbial composition (Crump et al. 2007). The microbiome is also affected by factors such as host health, and diet as well (Turnbaugh et al. 2009; Youngblut et al. 2019; Lassalle et al. 2018). This suggests that the change in microbiome composition between the body sites and host health states could be influencing the prophage counts.

To examine this possibility I focused here on specific bacterial species that were common across multiple body sites, region, or host health categories. To do this, I took the data from Results Chapter 1 (Inglis et al. 2024), gathered the taxon metadata for each genome, and condensed the geographical location metadata into geographical regions based on the United Nations Geoscheme groupings. I then compared body sites, geographical locations and host health but split the data by bacterial species instead of body site to determine whether the patterns remained when accounting for species. Stastical analysis was done in SPSS using Kruskal-Wallis tests.

Bodysite

There are some small differences between body sites when looking at specific bacteria, with some species having one or two sites with significantly different amounts of prophage DNA than the rest. For example, *Streptococcus pneumoniae* has next to no prophage DNA when found in the sputum, but contains more prophages when found in other areas, such as the nose and throat (figure 6). Most *Staphylococcus epidermidis* prophages are found in body sites other than the skin, where it is a normal part of the healthy microbiome. Instead, prophages were more common in nose and blood samples where *S. epidermidis* acts as an opportunistic pathogen and is a common hospital-acquired infection (Vuong and Otto 2002).

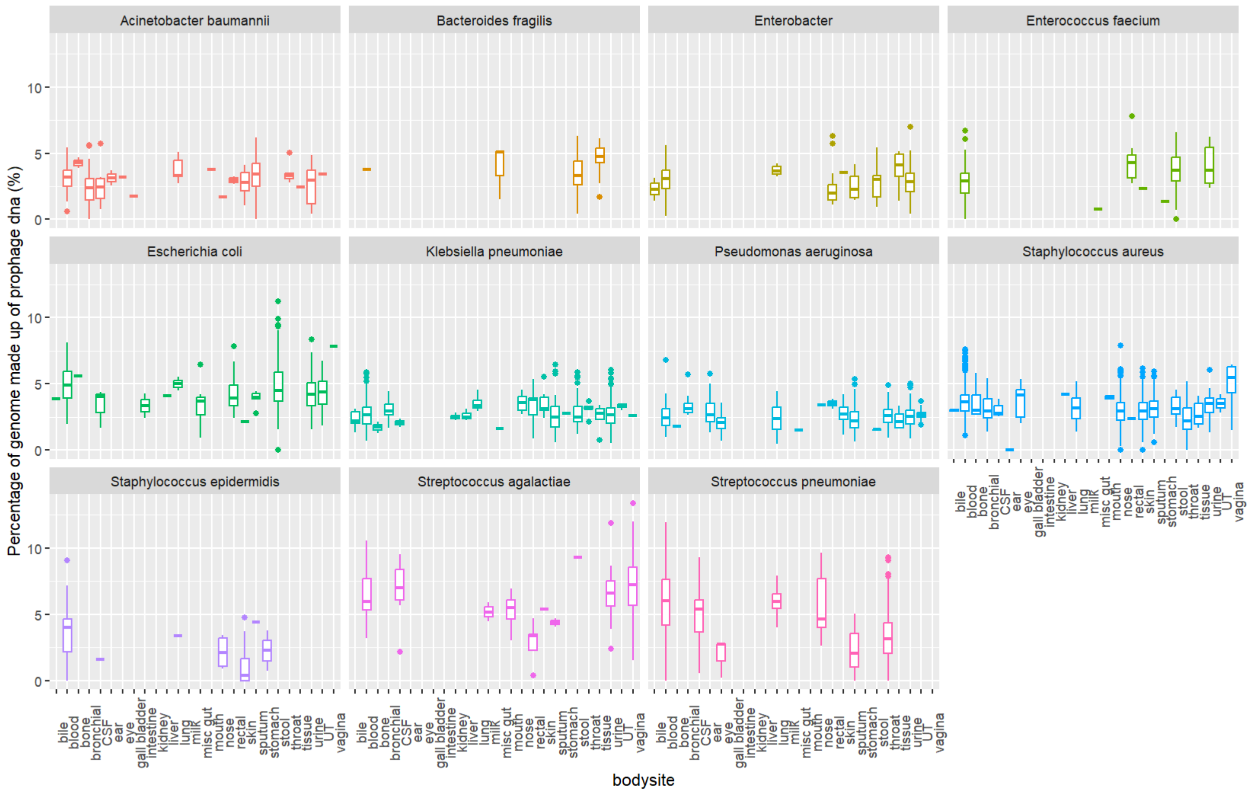

Figure 6: Average amount of prophage DNA in different areas of the human body for 11 different species of bacteria.

Geography

Out of the 22 species with more than 100 genomes in the dataset, 16 had some significant differences between geographic regions. Most of these consisted of one or sometimes two regions being significantly different from the rest. For example, in *Pseudomonas aeruginosa* samples from Asia and Oceania were significantly different from most other regions, including each other, with Asia having a significantly higher average amount of prophage DNA (Figure 7). In *Acinetobacter baumannii*, a dangerous pathogen, genomes from Europe had significantly lower amounts of prophage DNA from other regions except Africa and Central America. However, Central America, Africa and Oceania lacked a large pool of genomes to begin with, with the majority of genomes being collected from North America and Asia. Smaller sample sizes are much harder to draw conclusions from. There were also a number of samples where there was no clear location metadata, further hindering our progress.

Geography has a large overlap with other lifestyle factors such as diet that are known to affect the microbiome, as different foods are more commonplace in different regions. Though, our research shows that the phageome varies within the same bacterial species across different regions, suggesting that some species have either gained or lost prophages over time, while species such as *S. aureus* remain relatively the same across regions and only significantly differ across some areas of the body.

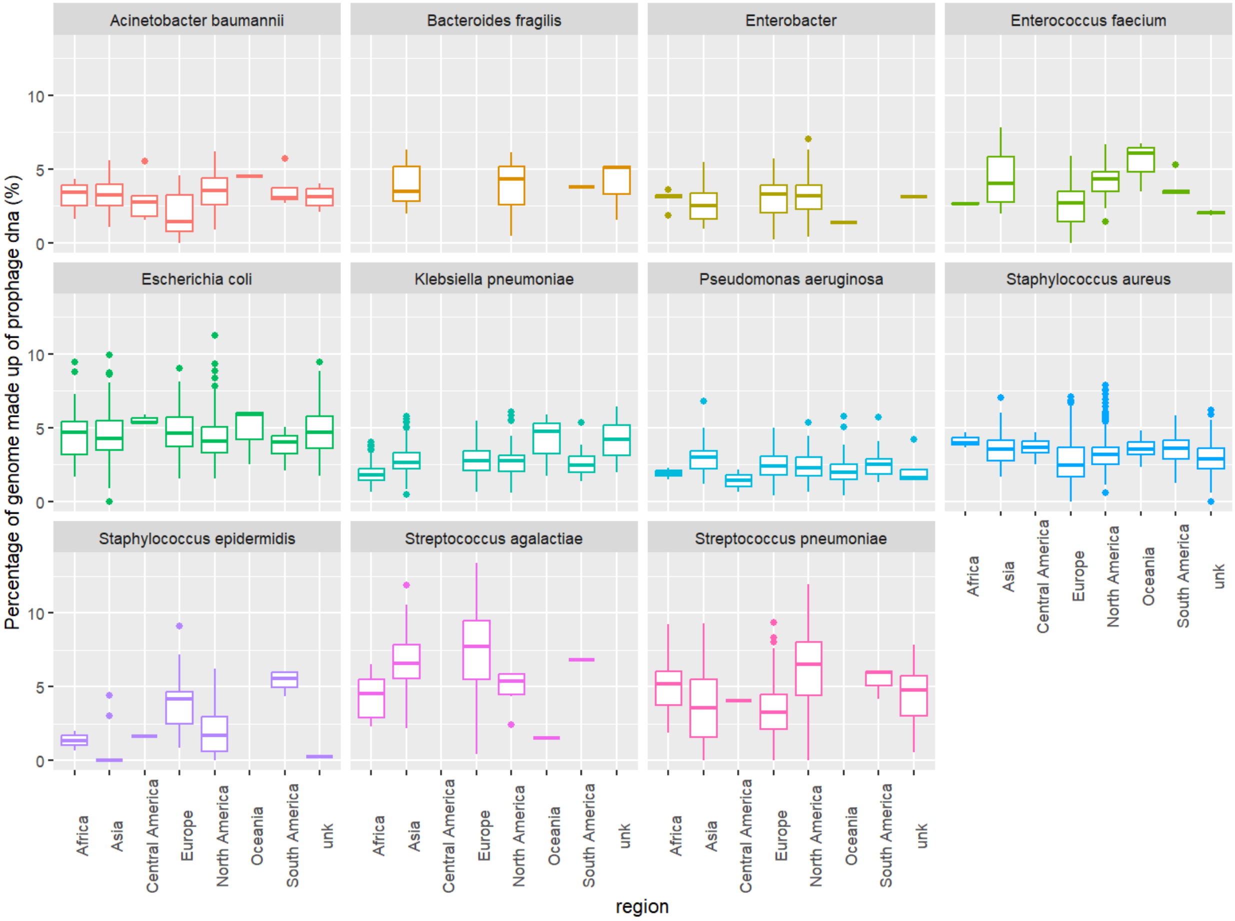

Figure 7: *Average amount of prophage DNA in different geographical regions for 11 different species of bacteria.*

Host health

Eight species in the dataset had samples from both symptomatic and asymptomatic humans. Six of the eight species showed significant differences in the amount of prophage DNA in asymptomatic vs. symptomatic samples, and of those six, five had more prophage DNA on average in genomes from symptomatic humans. Prophages are known to provide toxins and virulence genes to some bacterial species, this includes Shiga-like toxins in *E.coli*, and staph enterotoxins in *S. aureus* (Rodríguez-Rubio et al. 2021: Bielaszewska et al. 2012; Bokarewa, Jin, and Tarkowski 2006). There is evidence that *S. aureus* toxin genes have been passed onto other *Staphylococcus* species, such as S. epidermidis, through mobile genetic elements such as pathogenicity islands (Madhusoodanan et al., 2011), it is possible that phages may be contributing to the virulence of these other species. Chapter 2 of this thesis will examine the prescence of virulence genes in these prophages.However, healthy controls are relatively rare among samples deposited into GenBank, with the human microbiome project being the main source of samples from healthy people. Healthy controls are important for any clinical study and it is odd that so few are labelled as such. However, there are also prophages that reduce virulence such as prophage Φ13 found in *S. aureus* (Poupei et al., 2025). though, our results show that *S. aureus* sampled from symptomatic patients contain more prophage DNA, suggesting that perhaps the prophages that increase virulence may be more common, at least in *S aureus*, than those that reduce virulence.

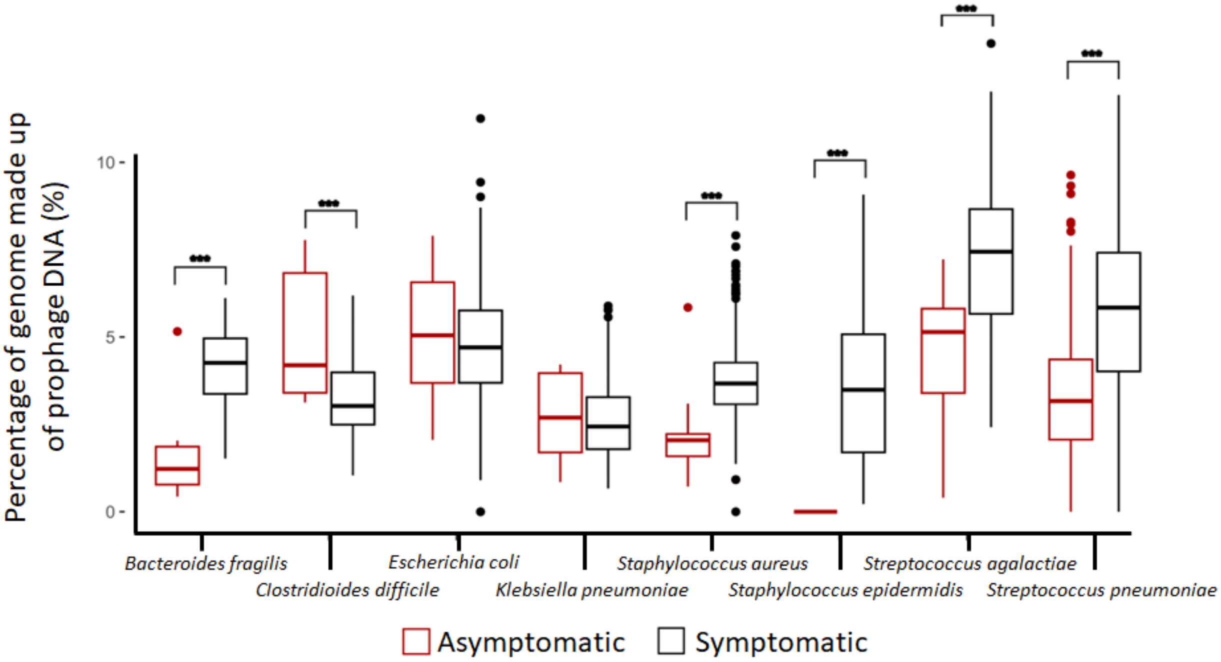

*Figure 8: Percentage of genome composed of prophage DNA across the eight species of bacteria that were found in both asymptomatic and symptomatic humans. Bacteria were either isolated from asymptomatic humans (red) or symptomatic humans (black). *** represents a p-value of >0.001.*

Species without prophages

Our dataset contained some genomes without any detected phage genes, though three species had none or almost no prophages at all. These were *Helicobacter pylori*, *Mycoplasma pneumoniae,* and *Chlamydia trachomatis*.

H. pylori is a species of gram-negative bacteria that infects the lining of the stomach and can cause abdominal pain and nausea. It is found across the world but most of the samples in the dataset were taken from Asia and Europe. They had one of the lowest average amounts of prophages with most samples having zero prophages. While there were only three North American samples, two of the three had prophages, with one sample having three, the most prophages in a *Helicobacte*r species, with only two of the three *Helicobacter cinaedi* samples having equal or more prophages. No matter where in the body helicobacter species are found that are lacking in prophages compared to most bacteria.

The stomach microbiome can be diverse when not infected by *H. pylori*. When *H. pylo*ri does infect the stomach, it quickly becomes the dominant species (Noto and Peek 2017). The Kill-the-Winner (KtW) hypothesis suggests that phages tend to choose lysis more often when their

host is the most dominant species, as the chance of the free virions finding a new host is high (Thingstad 2000). This would explain why genomes from the stomach are so low in prophage DNA, as most of the stomach samples were H. pylori genomes, which would have come from a lower-diversity microbiome. This highlights the need for more healthy samples
The two bacterial species with the least amount of prophages in the dataset were *Mycoplasma pneumoniae* and *Chlamydia trachomatis*, with no prophage DNA found in any of the samples in the dataset. Both of these species are pathogens, potentially leading to a similar low-diversity KtW scenario as *H. pylor*i.

Conclusion

Overall, while different bacterial species contain different average amounts of prophage DNA, these too are affected by body site, human health, and geographical region. Most body sites contained a range of species and many species were found across multiple body sites. This data provides some more insight into the patterns in the prophage DNA, while still remaining consistent with the rest of the data. The wide range of the most common bacteria shows that bacterial species are not overly biasing the data in most regions, and the species that can be found across symptomatic and asymptomatic humans suggest that it isn't entirely the difference in species that is driving the change in prophage abundance. However, body site was much less noticeable when examining specific species, with host health becoming the most significant factor.

These results also further highlight the problem of lacking metadata and rare samples. Even with thousands of genomes from across the world many regions of the world were undersampled, with many genomes containing no location metadata. While well-studied species such as *S. aureus* were very sommon in our dataset, most species were rarely sampled, showing that as large as the database is we have only sampled a small part of the world's microbial diversity, highlighting that we need to be able to get the most out of what we have, and well-annotated metadata is one way of maximising the potential of online databases.

Results Chapter 2: Prophages as a source of antimicrobial resistance in the human microbiome

Paper declaration

Laura K Inglis came up with the concept with the assistance of Robert A Edwards. They also analysed the data, produced the dataset used in this study, and wrote the manuscript.

Susanna R Grigson provided extra data in the form of prophage genomes that had been separated from the bacterial genomes. They also edited the document and provided advice on its contents.

Michael J Roach provided technical advice on the programs and coding languages used for data analysis. They also edited the final document and provided advice on its layout and contents.

Robert A Edwards assisted with the conceptualisation of the project. They also provided the initial data, assisted and provided advice regarding data analysis, edited the final document, and provided advice on its contents.

ChatGPT was used in the editing of some sections of this paper

Abstract

Prophages—viruses that integrate into bacterial genomes—are ubiquitous in the microbial realm. Prophages contribute significantly to horizontal gene transfer, including the potential spread of antimicrobial resistance (AMR) genes, because they can collect host genes. Understanding their role in the human microbiome is essential for fully understanding AMR dynamics and possible clinical implications.

We analysed almost 15,000 bacterial genomes for prophages and AMR genes. The bacteria were isolated from diverse human body sites and geographical regions, and their genomes were retrieved from GenBank.

AMR genes were detected in 6.6% of bacterial genomes, with a higher prevalence in people with symptomatic diseases. We found a wide variety of AMR genes combating multiple drug classes. We discovered AMR genes previously associated with plasmids, such as *blaOXA-23* in *Acinetobacter baumannii* prophages or genes found in prophages in species they had not been previously described in, such *as mefA-msrD* in *Gardnerella* prophages, suggesting prophage-mediated gene transfer of AMR genes. Prophages encoding AMR genes were found at varying frequencies across body sites and geographical regions, with Asia showing the highest diversity of AMR genes.

Importance

Antimicrobial resistance (AMR) is a growing threat to public health, and understanding how resistance genes spread between bacteria is essential for controlling their dissemination.

Bacteriophages, viruses that infect bacteria, have been recognised as potential vehicles for transferring these resistance genes, but their role in the human microbiome remains poorly understood. We examined nearly 15,000 bacterial genomes from various human body sites and regions worldwide to investigate how often prophages carry AMR genes in the human microbiome. Although AMR genes were uncommon in prophages, we identified diverse resistance genes across multiple bacterial species and drug classes, including some typically associated with plasmids. These findings reveal that prophages may contribute to the spread of resistance genes, highlighting an overlooked mechanism in the dynamics of AMR transmission. Ongoing monitoring of prophages is critical to fully understanding the pathways through which resistance genes move within microbial communities and impact human health.

Introduction

The virome is an essential microbiome component mainly consisting of bacteriophages (hereafter referred to as phages), which are viruses that infect bacteria. Phages have two main lifecycles: the lytic lifecycle, where virulent phages hijack the host's replication systems to create more virions, and the lysogenic life cycle, where temperate phages integrate into the bacterial host's genome. Temperate phages can integrate into the host's chromosome, where they can be replicated alongside the host. Temperate phages can also replicate extrachromosonally as phage plasmids, having characteristics of both plasmids and phages (E. Pfeifer and Rocha 2024; Eugen Pfeifer, Bonnin, and Rocha 2022; Cohen 1983; Cohen et al. 1996). Much like prophages, satellite phages, phages that don't encode all the necessary structural proteins and instead 'borrow' any required genes from other phages infecting the same host, typically integrate into a bacterial host until their required helper phage arrives (Dehò, Ghisotti, and Others 2006; deCarvalho et al. 2023). In addition, filamentous phages replicate and exit without lysing their hosts (Hay and Lithgow 2019).

Phages facilitate the evolution of their hosts by mediating the transfer of other mobile genetic elements such as transposons, plasmids, and genetic islands(Xia and Wolz 2014; Christie and Dokland 2012; Touchon, Moura de Sousa, and Rocha 2017). They can carry genes that probably originated in bacteria, known as auxiliary metabolic genes, and can provide these genes to their hosts in a process called lysogenic conversion. Errors in replication and phage excision can also move genes from the bacterial genome into the phage genome and vice versa, allowing phages to collect more genes to transfer to their next hosts (Touchon, Moura de Sousa, and Rocha 2017). Horizontal gene transfer connects the microbial world in a sort of "common gene pool" (Davis and Olsen 2010). The ability of phages to confer these auxiliary metabolic genes to their bacterial hosts and facilitate the transfer of other genetic elements that new bacteria can then take up allows genes to slowly hop around the microbiome, spreading both valuable and benign genes between species (Hendrix et al. 1999).

The term auxiliary metabolic genes is somewhat of a misnomer, though, as it is well known that prophages can contain genes other than metabolic genes, virulence genes being a well-known example. Several pathogens, such as *Vibrio cholerae* (Li et al. 2003), *Corynebacterium diphtheriae* (Muthuirulandi Sethuvel et al. 2019), and *Escherichia coli* (Rodríguez-Rubio et al.

2021), get many of their toxin-production genes from their prophages. There is also evidence to suggest that prophages regularly carry antimicrobial resistance genes.

Antimicrobial resistance (AMR) is a growing public health issue. Since the discovery of penicillin in 1928, antibiotics have become indispensable for treating life-threatening infections (Hutchings, Truman, and Wilkinson 2019). However, the short generation time of bacteria and the widespread use of antibiotics have led to a rapid arms race, causing the proliferation of AMR genes across various pathogens (Schmieder and Edwards 2012). Understanding how AMR genes are spread has been thoroughly researched since people became aware of the threat antibiotic resistance poses.

Since it is known that prophages can transfer bacterial genes between species, it is feasible that antimicrobial resistance (AMR) genes are transferred as well. AMR genes are common mobile genetic elements, but it was thought that the transduction events required to transfer AMR genes to and from phages were so uncommon as to be near impossible (Torres-Barceló 2018). It was not until recently that AMR genes were regularly being identified in phage genomes (Kondo, Kawano, and Sugai 2021; Eugen Pfeifer, Bonnin, and Rocha 2022; Moon et al. 2020; Brown-Jaque et al. 2018; Colomer-Lluch, Jofre, and Muniesa 2011). However, later studies suggest that overzealous interpretations of bioinformatics results and using lower identity thresholds may have resulted in overestimating the abundance of AMR genes in phage genomes (Enault et al. 2017). Here, we investigate the abundance of AMR genes in the prophage regions of almost 15,000 bacteria isolated from humans. With this, we aim to determine how common phage-encode AMR genes are across the human microbiome, and in doing so, we identified phage-transferred AMR genes that were previously thought to only be mobilised by plasmids.

Materials and Methods

Genome selection and filtering

We retrieved 949,935 bacterial genomes from GenBank on June 1, 2022. Genomes classified as metagenome-assembled or containing more than 50 contigs were excluded to ensure high-quality assemblies. This filtering set ensured an average fragment length exceeding 60 kbp, which is longer than the upper limit for many prophages (McKerral et al. (2023)). Duplicate accessions were removed, and the genomes were curated by isolation source and isolation region based on metadata from the Pathosystems Resource Integration Center (PATRIC). The curated accessions were then filtered to remove any genomes not isolated from humans. After filtering 14,987 genomes that could be associated with one of 31 different body site categories remained. 5,341 of these genomes could also be categorised by the health of the human host.

Prophage identification

We previously identified over 5 million high-quality prophages in these genomes using PhiSpy (Akhter, Aziz, and Edwards 2012; McKerral et al. 2023), one of the most accurate prophage prediction tools, with the lowest runtime currently available (Roach et al. 2022), described in Inglis et al. (2024).

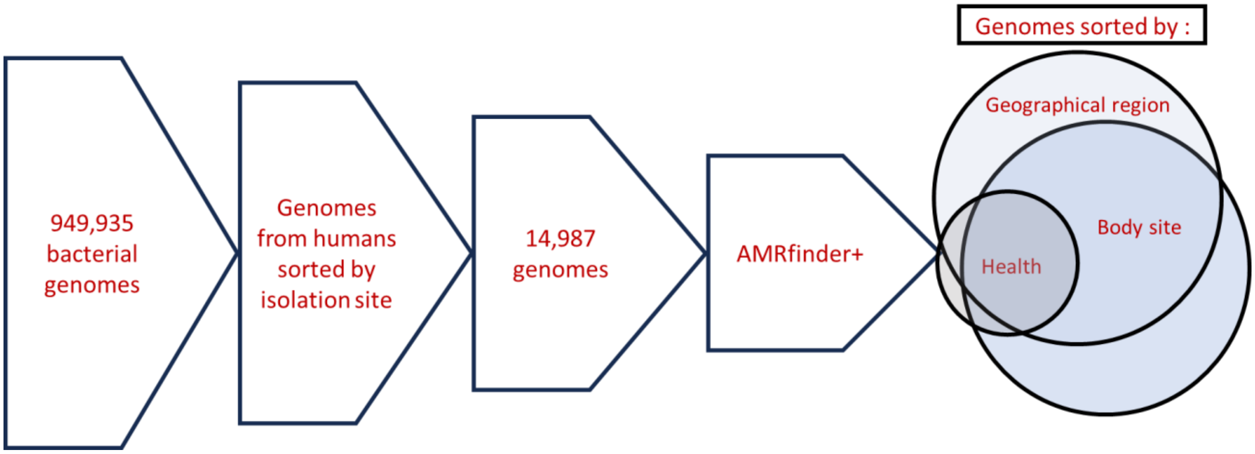

Figure 9: Flow chart of the methods

Gene detection

AMR, virulence, and stress genes were identified using AMRFinder+ (version 3.10.23, database version 2021-12-21.1) (Feldgarden et al. 2021). Both nucleotide and amino acid sequences were analysed to maximise detection accuracy. Stringent cutoffs for sequence identity (≥99%) and coverage (≥99%) were applied to minimise false positives. AMR genes detected within prophage regions were further validated by comparison against the Comprehensive Antibiotic Resistance Database (CARD) and corroborated through a literature search to avoid housekeeping genes.

Data analysis

The prevalence of AMR genes within prophage regions was compared across body sites, geographical areas, and host health status based on the PATRIC metadata to identify potential associations. Statistical analyses, including Kruskal-Wallis tests, were performed using SPSS to assess the significance of differences between AMR gene abundance across the aforementioned categories.

Results & Discussion

An overview of antimicrobial resistance genes in prophages

From the 14,987 bacterial genomes that were isolated from humans, 11,665 genomes came from 34 species that had more than 50 genomes in our dataset, while the remaining 3,322 genomes came from species with fewer than 50 genomes. 11,655 of the 14,987 genomes contained AMR genes, and 848 contained AMR genes in the predicted prophage regions. These 848 prophages contained a combined total of 1382 AMR genes, with an average of 1.6 genes per genome. The genomes contained a wide variety of AMR genes, and some genomes contained multiple genes. However, one AMR gene per genome was the most common result even when the genomes contained multiple phages, suggesting that not all are capable of producing a resistance phenotype. For virulence genes, 1,683 genomes contained virulence genes in their prophages, while stress genes were only found in 283 genomes.

Virulence and stress genes in prophages

Virulence genes are well-known auxiliary genes found in prophages and are the most common auxiliary genes found in our prophages. 14.4% of the genomes analysed carried a virulence gene in their prophage regions. 23.3% of nasal samples, 34.5% of skin samples, and 36.9% of urinary tract samples contained prophages with virulence genes. Most of the bacteria in these samples were *Staphylococcus aureu*s (65.3%) and *Escherichia coli* (33.7%). The most common genes were the increased serum survival protein *Iss*, nearly identical to the prophage-encoded protein *Bor.* The *Iss* protein is derived from the *Bor* protein, which is known to occur in *E. coli*'s phage λ (Johnson, Wannemuehler, and Nolan 2008). Three of the most common genes are a staphylococcal enterotoxin, a phage-encoded staphylokinase (Bokarewa, Jin, and Tarkowski 2006), and the complement inhibitor SCIN-A. The complete results list can be found in Supplementary Data 2e.

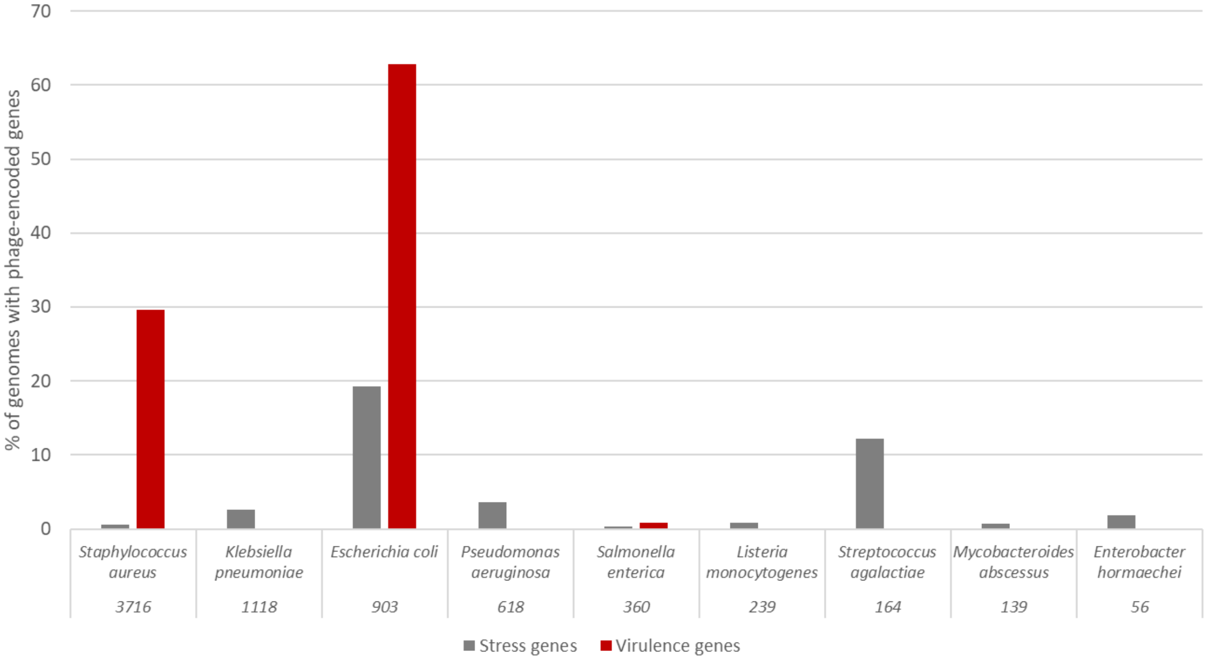

Figure 10: Percentage of genomes that carry either virulence or stress genes for the nine species represented by more than 50 genomes in our dataset that contained virulence or stress genes in prophage regions. The numbers beneath the species names are the number of genomes in our dataset.

As defined by AMRfinder+, stress genes were the least commonly detected among our results, with 2.4% of the analysed genomes containing prophage regions that carry these genes. Stress genes were grouped into heat, metal, acid, and biocide resistance genes, including efflux pumps, transport proteins, repressors, regulators, and reductases.

Stress genes were found in the predicted prophage regions of 14/31 body sites. Stool and urine samples had the most genomes with stress genes in their prophage-encoded regions, with 26.5% of bacteria whose prophages encode stress genes isolated from urine samples and

27.6% from stool. However, there were about half as many urine samples (n=1275) as stool samples (n=2967) in our dataset. While the bacterial genomes from all the body sites with more than 1000 bacterial isolates contained stress genes in their prophages, some body sites with few bacterial isolates, such as liver (n=9) and bile (n=27), also contained stress genes. The complete results list can be found in Supplementary Data 2e.

93.6% of the stress genes in our prophage regions were found in genomes of bacteria taken from symptomatic hosts. Out of the 283 genomes containing stress genes in prophage regions, only 8 came from bacteria from healthy or asymptomatic hosts, while 118 came from bacteria from symptomatic hosts, and the remaining 157 were from humans of unknown health status. Most prophages containing stress genes were found in *E. coli*, with almost 20% of *E. coli* genomes containing them. More common bacteria in our dataset, such as *S. aureus* and *Klebsiella pneumoniae*, had comparatively fewer prophages that contained stress genes, with 0.6% and 2.6% of the respective species containing stress genes in prophage regions. This suggests that E. coli phages commonly carry a variety of genes that may aid their host, and we do know that some of these genes are expressed by the E. coli host (Johnson, Wannemuehler, and Nolan 2008). Overall, the virulence genes found in the prophages are mostly what we expected to find.

AMR genes in prophages

The notion that phages can harbor and transmit antimicrobial resistance genes has long been debated. Originally, it was thought that phages rarely encoded AMR genes, but then, with the expansion of viromics, studies began to suggest that phages might often carry AMR genes (Modi et al. 2013). In 2017, Enault et. al. suggested that the thresholds used in these metagenomic analyses were not strict enough, showed that some of these ARGs predicted by bioinformatic methods did not confer resistance when transferred into bacteria, and that there are only two phages in the publicly available phage genomes in RefSeq that contain ARGs (Enault et al. 2017). Although they specifically analysed lytic phages, they briefly discussed how prophages are quite different, with a higher abundance of AMR genes than lytic phages (Enault et al. 2017), and that many prophages with known AMR genes have been shown to be incapable of lysis anymore.

To ensure that the AMR genes detected in this study were likely to confer a resistance phenotype, a literature search was conducted to find evidence of all the AMR genes found in predicted prophage regions (see Supplementary Data 2f). We found that two genes, *fosB2* and *tet(X1)*, from three genomes were truncated gene variants that no longer produced resistance (Yang et al. 2004; Wisdom et al. 1992). Gene *ant(6)-Ia*, with 19 occurrences in our prophages, which may have unnamed variants of the gene that do not confer resistance (Rodrigues Souza et al. 2020). Gene *crpP*, with four occurrences in our prophages, did not confer the resistance to ciprofloxacin as it was originally thought to, but may still provide low-level fluoroquinolone resistance (Zubyk and Wright 2021).

The most common resistance genes among our prophages were the aminoglycoside nucleotidyltransferase *ANT(9)-Ia*, which is known to mediate resistance to spectinomycin

(Sheng et al. 2023), the *ermB* and *tetM* genes. Although these genes are often found together (Roberts et al. 1999), we did not frequently find them together in our dataset.

Bodysite and AMR

Different areas of the body have different microbiomes and, consequently, different viromes. We split our bacteria by the area of the human body where they were isolated to determine whether the proportion of prophages with AMR genes varied across the different body sites.

We found AMR genes in predicted prophage regions in bacteria isolated from 21 out of the 31 different areas of the body. Of the body sites that didn't contain AMR genes, only one site, the stomach, had more than 100 genomes in the dataset. We previously showed that the stomach largely lacked prophages (Inglis, Roach, and Edwards 2024). In the body site with the most bacteria in our dataset, stool, over 11% of the genomes contained AMR genes on their prophages. This was slightly higher than the vaginal genomes, which while having fewer samples (Figure 11) has similar amounts of prophages (Inglis, Roach, and Edwards 2024). Stool and vaginal both had high amounts of prophage DNA per bacterial genome and high overall bacteria concentrations, while the skin also has a high bacterial concentration but has a lower amount of prophages (Chen Chen et al. 2017; Sender, Fuchs, and Milo 2016; Inglis, Roach, and Edwards 2024) and far fewer of those prophages contain AMR genes. This suggests that the prophages of stool and vaginal bacteria are more likely to contain prophages.

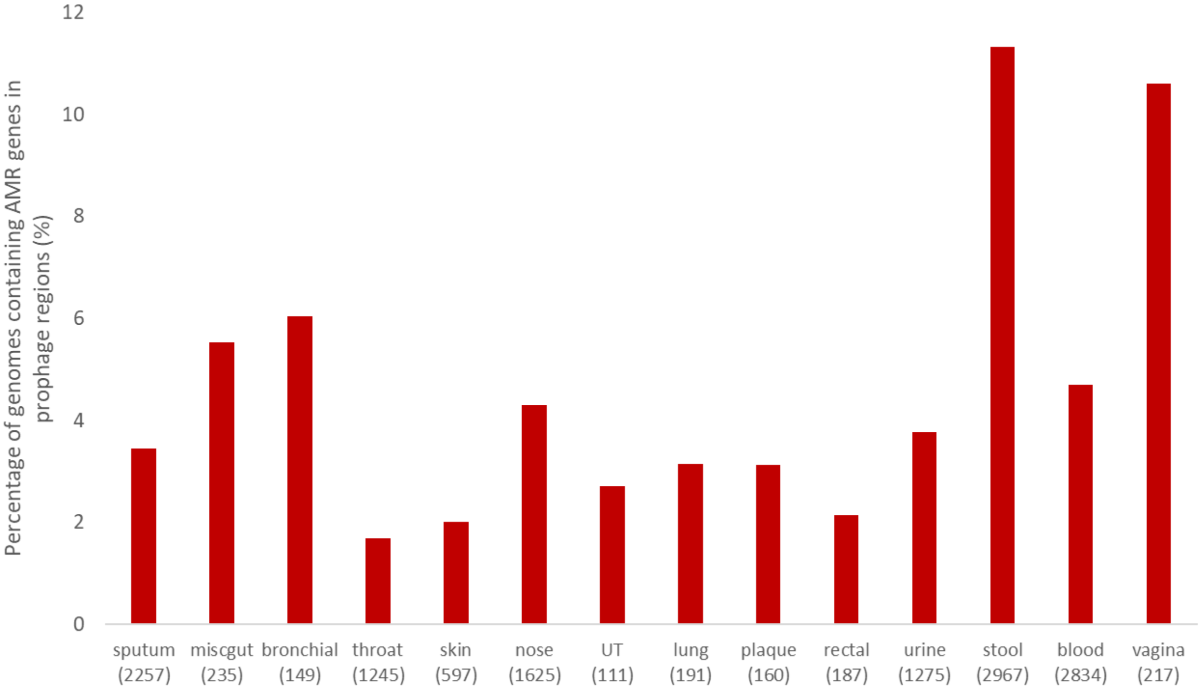

Figure 11: The percentage of genomes in our dataset that contained AMR genes in the predicted prophage regions. Bodysites are arranged left to right by the average amount of prophage DNA in the bacterial genome. The bracketed numbers are the number of bacterial genomes. UT and miscgut are abbreviations of urinary tract and miscellaneous gut.

The number of bacterial genomes that contained prophages with AMR genes generally correlated with the total number of genomes isolated from a body site, with a few notable exceptions, such as skin, tissue, and stool. Skin and tissue had many more bacterial isolates but fewer AMR genes in our dataset than body sites with similar numbers of genomes containing AMR genes, leading to a low proportion of genomes containing prophage-encoded AMR genes in these sites. Stool, the most frequently sampled body site, had 4.7% fewer genomes with AMR-containing prophages than the second most common body site, blood, but had 134% more genomes overall. A Chi-Square test of AMR gene presence across body site showed a significant difference (p <0.001). These results suggest that while a larger sample size does allow us to find more instances of prophage regions that contain AMR genes, some areas of the body appear to have higher rates of AMR genes being found in prophages than others.

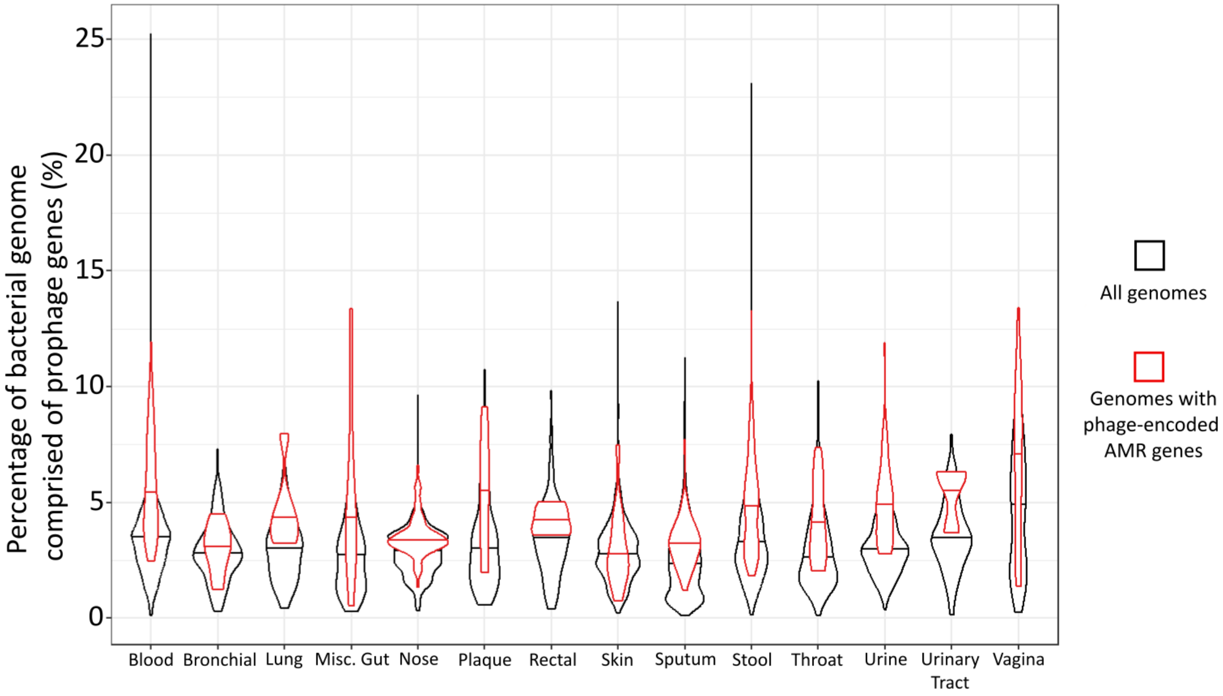

Figure 12: Violin plots comparing the percentage of the bacterial genome comprised of prophage DNA for all genomes in each body site (black) and genomes that contain AMR genes in their prophage regions (red). The horizontal line of the violin represents the median for each group.

Comparing the average proportion of prophage DNA per bacteria in each body site to the amount of prophage DNA per genome (Inglis, Roach, and Edwards 2024), we found that bacteria with AMR genes in prophage regions have, on average, higher amounts of prophage DNA in the genome than the average bacterial genome (Figure 12).
There is still variation in the amount of prophage DNA the bacterial genomes contain, and it is possible for bacteria to carry multiple prophages. Therefore, we hypothesise that the bacteria carrying the phages that encode AMR genes can also still contain the non-AMR-encoding phages.

Previously, we demonstrated that stool and vaginal samples, where over 10% of prophages harboured AMR genes, ranked among the five body sites with the highest prophage DNA content (Inglis, Roach, and Edwards 2024). While our results suggest that the likelihood of bacteria containing prophage-encoded AMR genes correlates with the amount of prophage DNA present, a Mann-Whitney U test showed there is no significant difference (p=0.303) between the presence of prophage-encoded AMR genes and total genome size.

Host Geography and AMR

The composition of the human microbiome and the prophages within them vary with geographical location (Suzuki and Worobey 2014; Yatsunenko et al. 2012; Gaulke and Sharpton 2018; Inglis, Roach, and Edwards 2024), and countries have varied rates of antibiotic consumption (Klein et al. 2018).

Most geographic regions had AMR genes in the prophage regions of 3-7% of their genomes. South America had the lowest proportion of genomes with AMR genes on their prophage regions at 2.3%. Africa and Asia, with 8.4% and 9.7% of the genomes, respectively, were the most likely to have AMR genes in their prophage regions

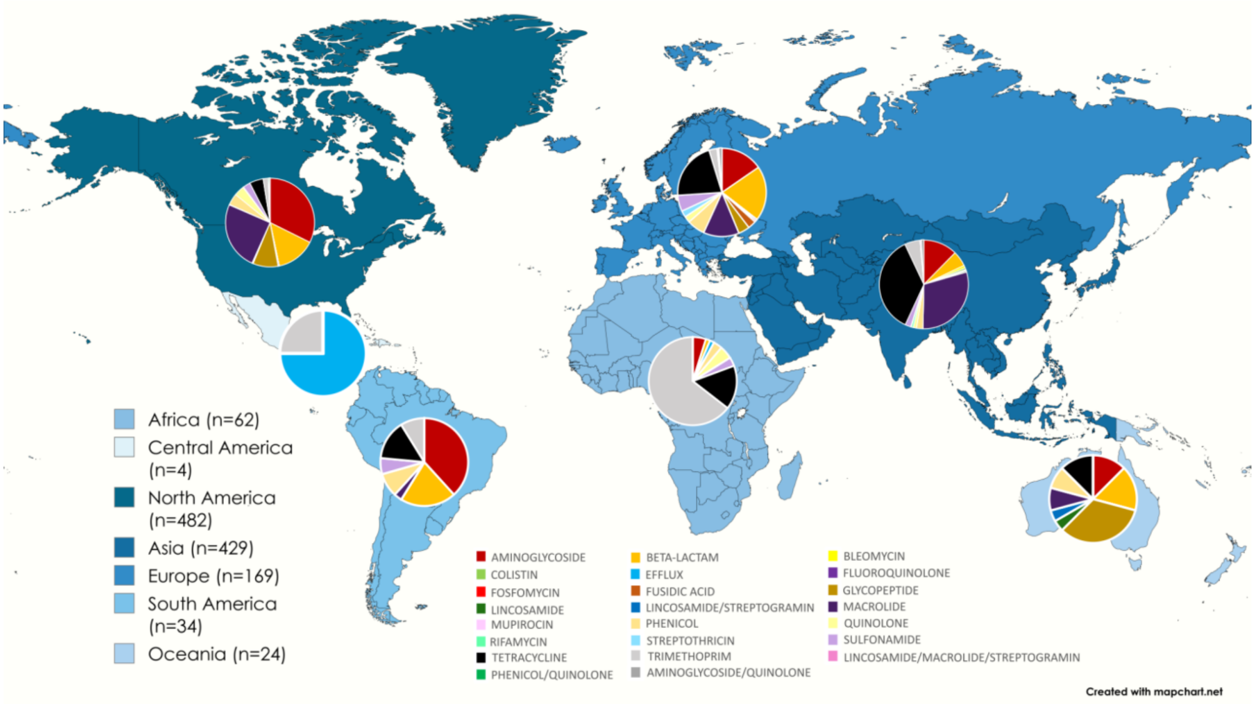

Figure 13: A map showing the AMR genes found in the prophage regions of genomes from each region split by drug class as defined by AMRfinder+, with n being the total number of genes found in each region.

Asia also has the most different kinds of AMR genes, with 73 unique genes across 17 classes, and North America has the second most, with 72 genes across 14 classes (Figure 13). The number of unique genes mostly aligns with the sample size, with regions with more genomes submitted having a wider variety of AMR genes represented. There were two exceptions to this trend, Africa and Asia. Africa had slightly more genomes than South America, 663 and 659 genomes, respectively, but South American prophages have almost double the rate of AMR genes than those from Africa. Asia has the most unique AMR genes but the third highest number of genomes and approximately half the number of genomes as from North America. None of the genes were found in prophages in every region.

A Kruskal-Wallis test showed some significant differences in the average number of AMR genes in prophage regions per genome. Asia, North America, and Africa were all significantly different from each other (p= <0.001-0.002). Samples from Asia were the most likely to contain AMR genes in phage regions, with 9.7% of genomes harbouring prophages that carried AMR genes. The least likely regions to find bacteria with prophage-encoded AMR genes are South America (2.3%) and Europe (2.7%).

All regions have a large variety of AMR classes, with Europe having the most classes represented (figure 5). Resistance to beta-lactams was the most common in all regions except for Africa, with aminoglycoside resistance being slightly more common, and Oceania, with efflux pumps being most common.

Host health and AMR

Overall, there was no significant difference in the number of prophage AMR genes per genome between healthy, asymptomatic, and symptomatic samples (Kruskal-Wallis p=0.409). However, more genomes were sampled from symptomatic humans than asymptomatic humans, and symptomatic samples were isolated from a wider variety of body sites and geographical regions. A wide variety of AMR genes was found in those samples. 6.6% of the 4,585 genomes from symptomatic hosts contained a predicted prophage with an AMR gene.

The genomes of the 159 bacterial isolates from healthy hosts contained fewer genes, 5.7% of which contained AMR genes in predicted prophage regions. The 609 genomes labeled as coming from asymptomatic individuals had a similar proportion of genomes with AMR genes (2.6%).

As bacterial genomes from healthy controls were much rarer in the GenBank database than genomes from symptomatic humans, the results we can draw from the comparisons are limited by low sample sizes. However, *Streptococcus pneumoniae* had more samples labeled as asymptomatic than symptomatic, and it showed a similar trend to other results, with bacteria isolated from asymptomatic hosts being less likely to have AMR genes in their prophage regions than bacteria isolated from symptomatic hosts. The difference is much more pronounced than most of the other species. However, bacteria from asymptomatic hosts are less likely to carry prophages with AMR genes than bacteria from symptomatic hosts, when looking at each species separately. I hypothesise that since antibiotics are given to treat illnesses and are thus more likely to be given to symptomatic humans, there is less selective pressure for antibiotic resistance in healthy populations. Therefore people who are ill should be expected to harbour more antibiotic resistant bacteria.

Out of the eight bacterial species that were sampled from both asymptomatic and symptomatic humans, seven species contained AMR genes in their phage regions (Figure 14), while one species, *Bacteroides fragilis*, contained no AMR genes on prophages at all. Two species, *Escherichia coli* and *Staphylococcus epidermidis*, contained no AMR genes in the asymptomatic samples. One species, *Clostridioides difficile,* formerly known as *Clostridium difficile*, has a higher percentage of prophage genomes from asymptomatic hosts having AMR genes than prophage genomes from symptomatic hosts. The remaining four have a lower rate of genomes with AMR genes in prophages in asymptomatic samples than in symptomatic samples. Since there is not enough metadata for many of these genomes to accurately determine whether these bacteria were the cause of these people's symptoms. However, these species are all common pathogens, and *S. aureus, K. pneumoniae*, and *E. coli* in particular, are common hospital-acquired infections (Boev and Kiss, 2017), and ESKAPEE pathogens which are of high concern for antibiotic resistance. If these species are not the original cause of the host's symptoms, a hospitalised patient would have a higher chance of coming into contact with resistant bacterial strains than the average person.

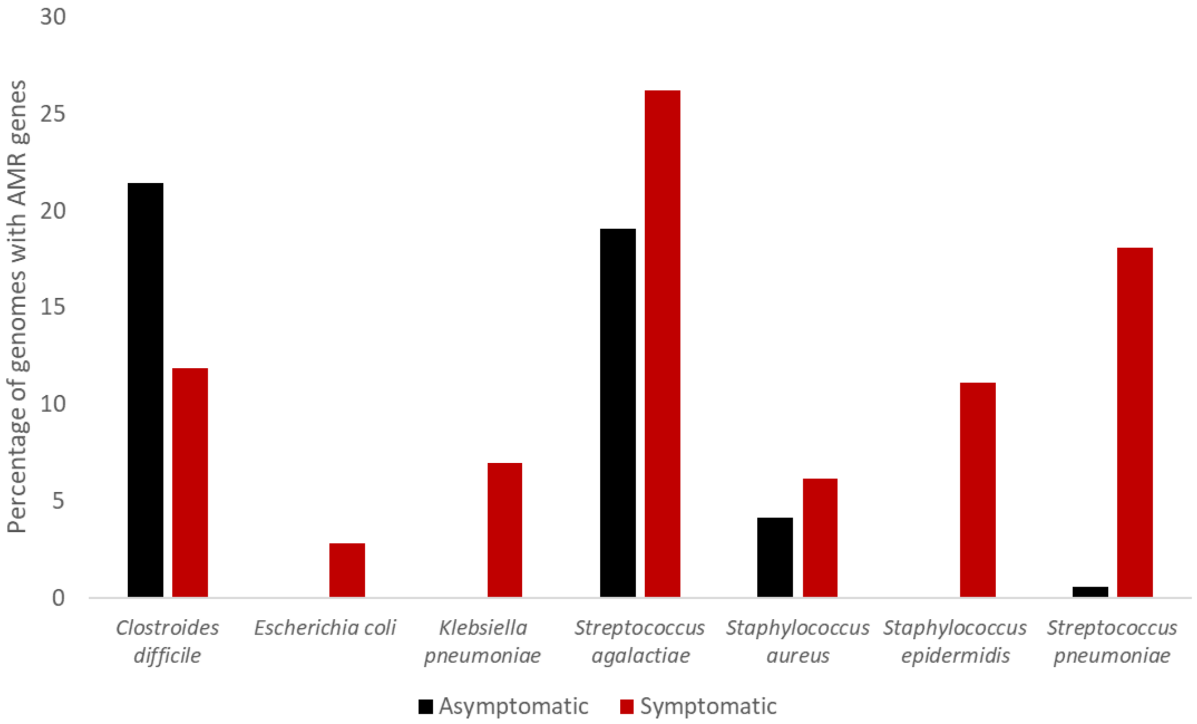

Figure 14: Percentage of bacterial genomes containing AMR genes in their prophage regions for seven species, split by the health of the human host (red: hosts that are symptomatic for illness, black: hosts that have no symptoms).

Known phage-encoded AMR genes

Our dataset contained a wide variety of AMR genes. These included genes previously known to be found in prophages, such as *MefA-MsrD,* and some that were only previously known to occur in other mobile genetic elements, such as *BlaOXA23* and *TMexCD-ToprJ.*

Fosfomycin resistance has been known to be encoded by multiple lytic phage isolates of γ phage (Schuch and Fischetti 2006), and it was thought that this *FosB* gene originally came from a prophage as the wild type Wβ phage does not carry the gene (Schuch and Fischetti 2006; Gillis and Mahillon 2014). While we did not find a prophage containing the *FosB* gene, we did

find prophages infecting the same species, *Bacillus anthracis*, that carried the spliced variant of the *fosB* gene, *fosB*2, even though *fosB2* cannot transform fibroblasts without the presence of a trans-activation domain as it has lost its activation domain (Wisdom et al. 1992).
The *TMexCD-ToprJ* gene clusters have been shown to confer resistance to carbapenems and tigecycline, a last-resort antibiotic for carbapenem-resistant bacteria (Zhu et al. 2023; Lv et al. 2020). This gene cluster has been identified in multiple Pseudomonas species, and we found it in *Pseudomonas aeruginosa*, both in the bacterial genomes and in a prophage region in genome GCA_008195485.1 collected in 2001.
The beta-lactamase *BlaOXA23* is a major source of carbapenem resistance for *Acinetobacter* species. Previously, *BlaOXA23* was found on conjugative plasmids (Zong et al. 2020), though few *Acinetobacter* species have conjugative plasmids, which was proposed to limit the transmission of this gene. The same penem resistance gene has also been shown to be found on transposons (Zong et al. 2020). However, we found three instances of the beta-lactamase in prophage regions of *Acinetobacter baumannii*. Phage-mediated transfer of beta-lactamases is potentially concerning as it suggests that there has been at least one instance of an AMR gene being transferred from a plasmid to a prophage. They were the only instances of AMR genes found in prophage regions of *A. baumannii*.
Phages are known to be able to transfer the macrolide efflux pump encoded by the *MefA-MsrD* gene pair to their *Streptococcus pneumoniae* hosts (Fox et al. 2021), however, we also found it in a *Gardnerella spp*. prophage.
ESKAPEE pathogens are a group of species, specifically known for their high rates of antibiotic resistance. They also face increased selection pressure due to the amount of antibiotics used to eliminate these multidrug-resistant infections. We found that while the AMR genes were very common among the ESKAPE species in our dataset, the rates of AMR genes occurring on their prophages were not above average. We hypothesise that prophages don't face as much selection pressure as their hosts to gather resistance genes. However, there were cases where either a gene was unique to prophages or a problematic AMR gene was shown to have moved to a prophage. For example, *Ant(6)-Ia* was found in many species, both in prophage regions and in the bacterial DNA, but in *Streptococcus pyogenes*, it was only found in the prophage regions
A concern in the fight against AMR is that genes can be transferred between bacteria by horizontal transmission, spreading resistance to previously susceptible bacteria. We found 34 bacterial species, each represented more than once in our dataset, that contained specific AMR genes *only* on prophage regions. However, homologous genes could also be found in both prophage and bacterial regions in the genomes of other species, suggesting that the prophages are mobilising these genes.

While most of our common species had prophages that contained AMR genes even if they were quite rare, there were species in our dataset that did not contain any prophages with AMR genes at all, suggesting that not all phages have equal access to the common gene pool and antimicrobial resistance genes.

Conclusion

After analysing the AMR genes in almost 15,000 bacterial genomes we found that AMR genes were relatively rare in human prophages, however, there was a wide variety of them. Different areas of the body and regions of the globe both had significantly different rates of AMR genes appearing on prophages and differing amounts of unique AMR genes occurring. Host health was only a significant factor when examining specific species. *S. pneumoniae* showed a significant difference between the number of asymptomatic and symptomatic genomes containing phage-encoded AMR genes, and three species, *E. coli, K. pneumoniae*, and *S. epidermidis*, contained no phage-encoded AMR genes in asymptomatic samples. Our results suggest that bacterial genomes with larger or multiple prophages were more likely to have a prophage that carries an AMR gene.

While AMR genes being found in prophage regions are currently rare, we identified at least one example of a gene that was known to be plasmid-bourne being found on a prophage, an instance of a gene that has moved between different prophages, of different life cycles, some genomes that only have AMR genes in their prophage regions, and genomes that appear to have acquired genes by prophages.

Acknowledgments

We would like to acknowledge the support provided by Flinders University for HPC research resources (Flinders University 2021).

We would also like to acknowledge the use of ChatGPT to assist with grammar and layout tips.

Funding

Awards from the NIH NIDDK RC2DK116713 and the Australian Research Council DP220102915 to R.A.E. supported this research. Flinders University Impact Seed Funding for Early Career Researchers supported M.J.R.

Results Chapter 3: Inferring Microbial Habitat from Function and Taxonomy: A Machine Learning Approach to Metagenome Classification

Paper declaration

Laura K Inglis came up with the concept with the assistance of Robert A Edwards. They chose the machine learning models to test, gathered the metagenomes, and created the training datasets. They also trained and tested each of the models and wrote the final document.

Susanna R Grigson provided advice on the implementation and validation of the machine learning algorithms. They also edited the final document and provided advice on its layout and contents.

Robert A Edwards assisted with the conceptualisation of the project. They also assisted and provided advice regarding data acquisition and analysis, and edited the final document.

ChatGPT was used in the editing of this paper

This paper has been submitted to Microbial Genomics and is awaiting review

Abstract

Metagenomic sequencing enables culture-independent profiling of microbial communities across diverse environments, yet publicly available metagenomic datasets are often limited by inconsistent or incomplete metadata. In particular, environmental annotations are frequently missing or ambiguous, hampering large-scale comparative studies. Here, we used machine learning to predict the environmental origin of metagenomes based on either taxonomic or functional profiles. Using 1,120 metagenomes with recorded isolation environments from the MGnify database, we developed a two-tiered environmental ontology comprising broad and narrow source categories. We trained and benchmarked 14 random forest classifiers using taxonomic features (strain to family) and functional features (subsystems), comparing performance across classification schemes. Our results show that taxonomic profiles, particularly at the genus level, yielded stronger predictive performance than functional profiles, although functional models exhibited greater generalisability to new datasets. Certain environments, such as soil and animal-associated samples, were accurately classified, while others like marine and built environment samples remained challenging. Feature importance analyses revealed that classification was driven by variation in core taxa and functions, rather than environment-specific taxonomic or functional markers. Our findings highlight the potential for automated environmental annotation of metagenomic datasets and underscore the need for improved metadata standards in public sequence repositories.

Impact statement

The rapidly expanding volume of publicly available metagenomic data offers immense opportunities for large-scale microbiome research. However, the inconsistent and incomplete environmental metadata associated with many archived samples presents a major barrier to meta-analysis. Our study addresses this gap by demonstrating that machine learning models can predict the environmental source of metagenomes using either taxonomic or functional data. We show that random forest classifiers trained on genus-level taxonomic profiles or high-level functional annotations can accurately assign samples to broad environmental categories. Importantly, these models remain effective even when applied to novel datasets, supporting their utility in metadata curation and environmental inference. Our work is valuable to microbiome researchers seeking to reuse public datasets and especially to those developing automated annotation and analysis pipelines.

Introduction

Microbial communities are ubiquitous across Earth's ecosystems, forming distinct and dynamic assemblages that play essential roles in environmental and host-associated processes. Historically, the study of these communities was constrained by culture-based techniques, which limited detection to taxa amenable to in vitro growth, excluding the vast majority of uncultivable organisms from analysis (Staley & Konopka 1985). The development of metagenomic sequencing has revolutionised microbiology by enabling culture-independent profiling of microbial diversity and function. Since the first metagenomic assemblies were reported in the early 2000s (Tyson et al. 2004; Breitbart et al. 2003), advances in sequencing technologies have expanded both the resolution and scale at which microbial ecosystems can be studied.

Two principal strategies dominate contemporary metagenomic studies: targeted amplicon sequencing and whole-genome shotgun (WGS) sequencing. Amplicon-based approaches focus on conserved marker genes, most commonly the bacterial 16S rRNA gene, to infer taxonomic composition. These methods are cost-effective and widely applied, particularly for surveys of bacterial communities (Johnson et al. 2019). However, their utility is reduced when characterising non-bacterial taxa. Although alternative markers, such as the internal transcribed spacer (ITS) region for fungi (De Filippis et al. 2017) or 18S rRNA for eukaryotes, are available, they offer lower resolution or limited taxonomic breadth. Moreover, there remains no universally conserved marker gene for viruses, further constraining amplicon-based assessments of viromes (Rohwer & Edwards 2002).

In contrast, WGS metagenomics sequences all the genomic material present in a sample, providing both taxonomic and functional insights, and enabling the detection of viruses and other microorganisms that are not accessible via targeted primers (Dinsdale et al. 2008). Although more resource-intensive, WGS metagenomics facilitates genome-resolved analyses, including the recovery of metagenome-assembled genomes (MAGs) and profiling of community functional potential (Papudeshi et al. 2017). As sequencing costs decline, the adoption of WGS sequencing has accelerated, resulting in the deposition of vast metagenomic datasets in public repositories, such as MGnify (Richardson et al. 2023) or the Sequence Read Archive (SRA) (Leinonen et al. 2010).

Despite the growth of these repositories, the utility of archived datasets is often hindered by inconsistent or incomplete metadata. Environmental annotations are frequently non-standardised or absent, impeding downstream reuse and large-scale meta-analyses. For instance, over 9% of WGS-labelled metagenomes in the SRA are tagged only as "metagenome" without further contextual information, severely limiting their value for ecological inference or comparative studies, and meaning that only a portion of the data within these databases can actually be used for future studies.

Machine learning offers a promising solution to infer missing environmental metadata based on features intrinsic to metagenomic datasets. Prior work has demonstrated that taxonomic profiles can be leveraged to predict the ecological origin of metagenomes (Burke et al. 2019). However, taxonomic profiles can vary substantially across similar environments due to ecological redundancy and niche-specific variation, which can potentially limit their generalizability. In contrast, the functional profile of a community, defined by the repertoire of encoded genes and pathways, may be more conserved across similar ecological contexts (Turnbaugh et al. 2009).

In this study, we evaluate the feasibility of using machine learning models trained on either taxonomic or functional profiles to classify WGS metagenomes based on their environmental sources. By comparing model performance across these two feature types, we aim to determine whether functional annotations provide a more robust signal for inferring ecological origin in the context of incomplete or missing metadata than the commonly used taxonomic profiles, and create a method by which untagged metagenomes can be used in future work. Our findings have implications for the reuse of public metagenomic data and the automated curation of environmental information in sequence repositories.

Methods

Dataset acquisition and preprocessing

All publicly available studies labelled as “*metagenomic*” within the MGnify database were queried via the MGnify API (version 1) (Richardson et al. 2023) in January 2024. This filtered for WGS metagenomes. MGnify was selected due to its relatively comprehensive metadata annotations compared to other repositories. From this dataset, metagenomes with clearly defined isolation sources were identified based on their associated metadata fields. These were manually curated and classified into discrete environmental source categories. To ensure a balanced and sufficiently powered dataset, only environmental categories represented by metagenomes from 40 or more unique studies were retained for model construction.

For each selected category, 50 metagenomes were randomly sampled to minimise overrepresentation biases, particularly from human-associated samples, which dominate metagenomic sequence repositories. The selected metagenomes were downloaded as FASTA-formatted files to the Flinders University High-Performance Computing cluster (Flinders University 2021).

Taxonomic profiling of each sample was performed using FOCUS (Silva et al. 2014), which provides read-based annotation from the phylum to the strain level. Functional annotations were inferred using SUPER-FOCUS v1.6 (Silva et al. 2016), which categorises reads into hierarchical subsystems of gene functions across three levels, culminating in predicted functional roles. To balance the number of features with the available samples, taxonomic data at the strain, species, genus, and family levels, as well as functional annotations from subsystem levels 1, 2, and 3, were selected for downstream analyses (Figure 15).

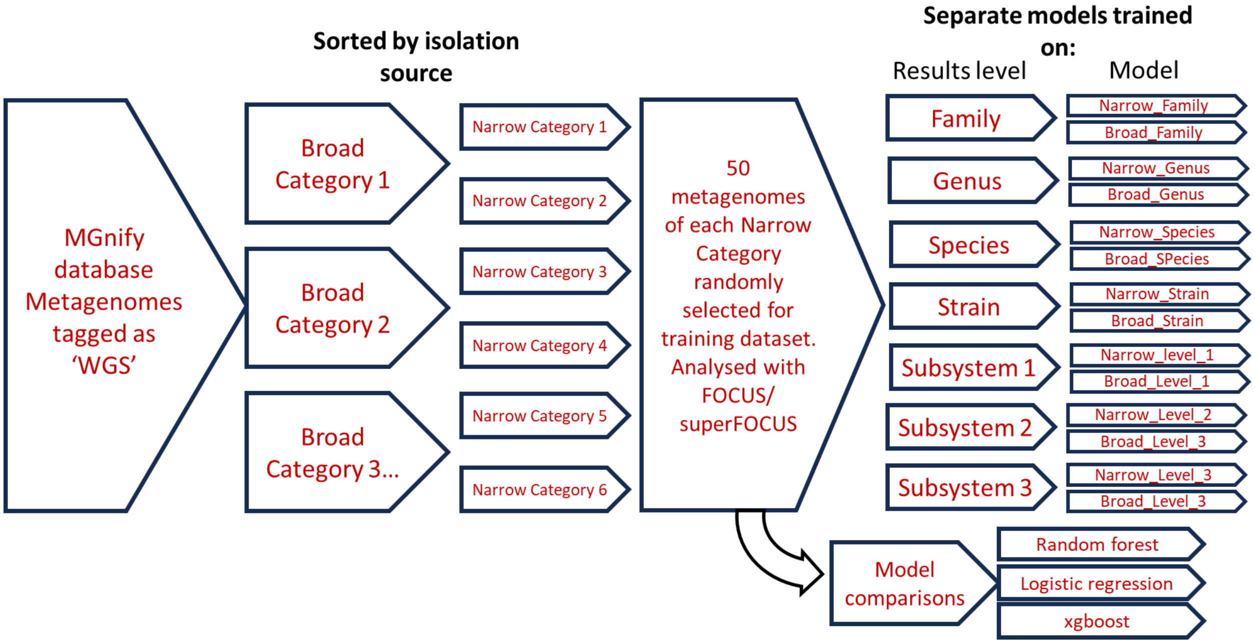

Figure 15: Graphical methods showing how the metagenomes were gathered, sorted and analysed to create training datasets for 14 random forest models

Machine learning model development

We implemented three classification algorithms: random forest (Breiman 2001), logistic regression (McFadden 1972), and XGBoost (Chen & Guestrin 2016). All models were developed using Python libraries within Google Colaboratory, employing the Scikit-learn (version 1.6.1) implementations of random forest and logistic regression, as well as the XGBoost Python package (version 2.1.4). These three algorithms were trained from scratch 10 times with reshuffled train/test data to compare how variable their results were.

For the random forest models, data were split into training (80%) and testing (20%) sets using the train_test_split method in scikit-learn with stratification to preserve class proportions. Extensive hyperparameter tuning was performed, including maximum tree depth, minimum sample size, number of trees, maximum leaf nodes, sample and feature subsampling strategies, and class balancing via the synthetic minority oversampling technique (SMOTE) (Chawla et al. 2011). Performance metrics across various parameter settings were evaluated (Supplementary Figures S2–S10), and the default settings were retained based on robust performance across all metrics.

Fourteen random forest models were trained, using seven FOCUS/SUPER-FOCUS-derived features (strain, species, genus, family, and subsystem levels 1–3) across two classification schemes: broad and narrow environmental categories. Broad categories included high-level source types (e.g. “*human*”), whereas narrow categories represented finer subdivisions (e.g. “*gut*”, “*skin*”, “*oral cavity*”).

Model performance was assessed using the area under the receiver operating characteristic curve (ROC-AUC), accuracy, precision, recall, and F_1 score using Scikit-learn. Confusion matrices were generated to identify common misclassifications. Feature importance metrics were computed to determine the relative contribution of input features to classification outcomes.

Precision was defined as the proportion of true positives among all predicted positives, while recall reflected the proportion of true positives among all actual positives. The F_1 score, the harmonic mean of precision and recall, was used as a summary measure of model accuracy. All evaluation metrics were calculated on the test set.

Results and discussion

Dataset overview and ontology structure

A total of 1,120 metagenomes were incorporated into the training dataset, curated from the MGnify repository and annotated using a simplified ontology derived from the Environment Ontology (ENVO) (Buttigieg et al. 2016). The constructed ontology comprised two hierarchical levels: broad environmental categories (e.g. "*freshwater*", "*human gastrointestinal*") and, where metadata permitted, narrower subclassifications (e.g. "*freshwater lake*", "*freshwater sediment*").

The dataset included between 37 and 50 metagenomes per narrow class and between 41 and 195 per broad category (Figure 16). Of the 12 broad categories included ("*animal*", "*built environment*”, "*freshwater*", "*marine*", "*human gastrointestinal*", "*human respiratory*", "*human oral*", "*human skin*", "*human other*", "*plant*", "*soil*", and "*wastewater*"), seven were further subdivided into 24 narrow classes, each represented by at least 35 metagenomes (Figure 2).

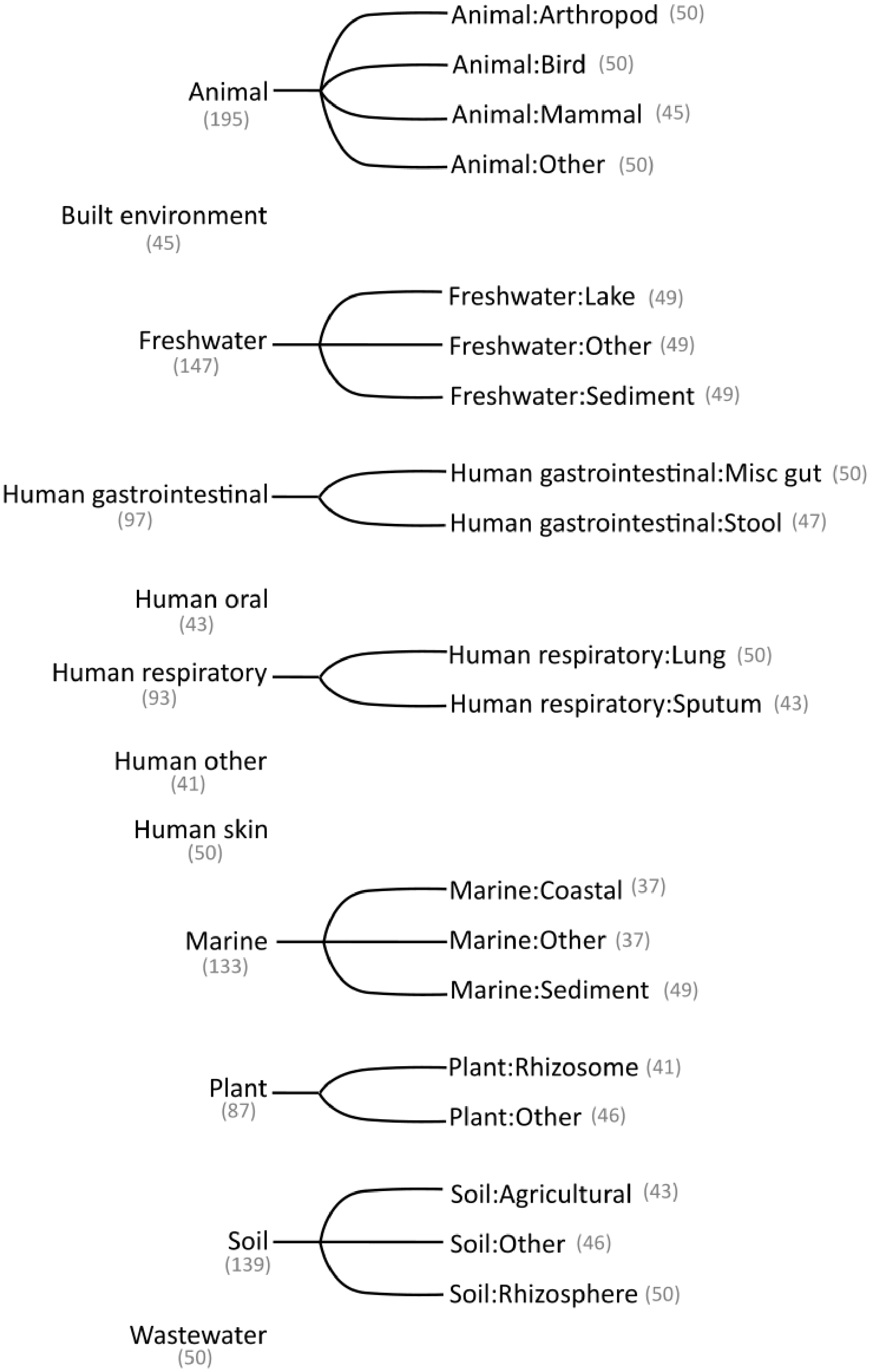
Animal (195)
Animal:Arthropod (50)
Animal:Bird (50)
Animal:Mammal (45)
Animal:Other (50)
Built environment (45)
Freshwater (147)
Freshwater:Lake (49)
Freshwater:Other (49)
Freshwater:Sediment (49)
Human gastrointestinal (97)
Human gastrointestinal:Misc gut (50)
Human gastrointestinal:Stool (47)
Human oral (43)
Human respiratory (93)
Human respiratory:Lung (50)
Human respiratory:Sputum (43)
Human other (41)
Human skin (50)
Marine (133)
Marine:Coastal (37)
Marine:Other (37)
Marine:Sediment (49)
Plant (87)
Plant:Rhizosome (41)
Plant:Other (46)
Soil (139)
Soil:Agricultural (43)
Soil:Other (46)
Soil:Rhizosphere (50)
Wastewater (50)

Figure 16: Flowchart of the categories used. N is shown in brackets. The lines illustrate how each narrow category is related to the broader classes.

Taxonomic and functional clustering of environmental categories

To assess whether the constructed ontology captured meaningful microbiome structure, we performed non-metric multidimensional scaling (NMDS) on both taxonomic (genus-level) and functional (subsystem level 2) profiles. These analyses revealed environment-specific clustering, with more apparent separation observed in taxonomic profiles than in functional ones (Figures 3 and 4). For instance, the broad category "*plant*" exhibited heterogeneous taxonomic profiles, whereas its narrow subclass, "*plant:rhizosphere,*" showed tighter clustering, suggesting that increased resolution improves ecological specificity.

Several factors not currently captured in metadata, such as seasonal dynamics (Gilbert et al. 2012), environmental connectivity (Crump et al. 2007), and microgeochemical gradients (Edwards et al. 2006; Alsop et al. 2014), likely contribute to the observed within-group variability. Moreover, a lack of clustering at broader scales may hinder model generalisation, while over-clustering around narrow microenvironments may risk overfitting.

Suppose microbial communities are structured predominantly by unmeasured microenvironmental variables such as local geochemistry, host physiology, or temporal dynamics. In that case, the training dataset may reflect these idiosyncratic conditions rather than the broader environmental categories used for classification. This overrepresentation of microenvironment-specific signatures risks model overfitting, thereby compromising its ability to generalise to unseen data. Conversely, if microbial communities lack consistent structure at the chosen environmental resolution, the resulting absence of informative patterns may reduce model discriminability, leading to lower predictive accuracy.

Taxonomic resolution had a marked influence on the observed clustering patterns. At the family level—the broadest taxonomic rank considered in this study—clustering was minimal (Figure S14), likely due to the taxonomic breadth encompassed at this level. Such coarse classifications may obscure meaningful ecological or biogeographic variation, reinforcing the notion that “everything is everywhere” when generalisation exceeds the ecological signal. In contrast, genus-level profiles displayed the most significant degree of environment-specific clustering, suggesting that this taxonomic resolution provides an optimal balance between the ecological specificity and cross-sample comparability.

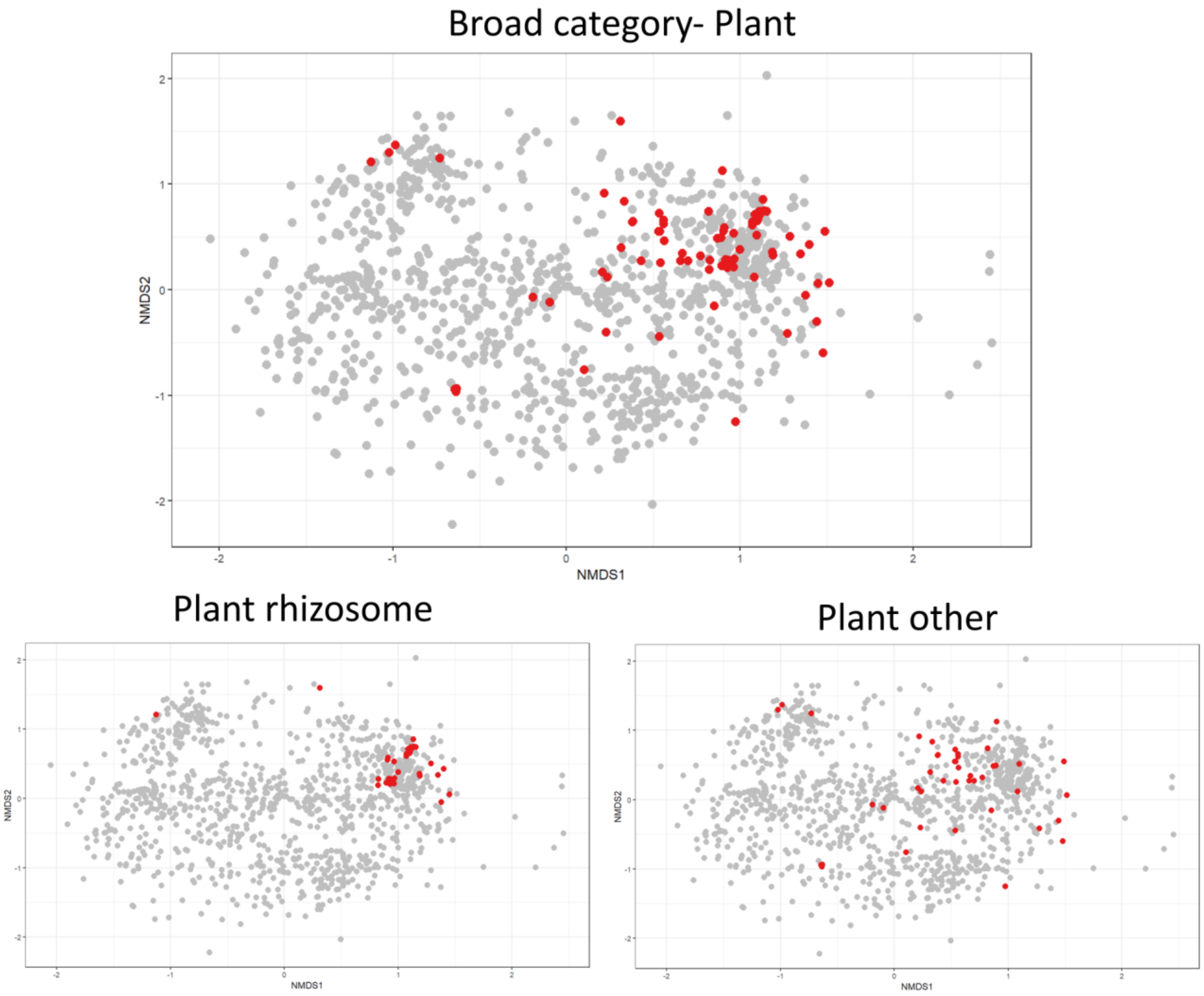

Figure 17: NMDS plots of the genus-level FOCUS results. The category listed is shown in red, while all other metagenomes are shown in grey.

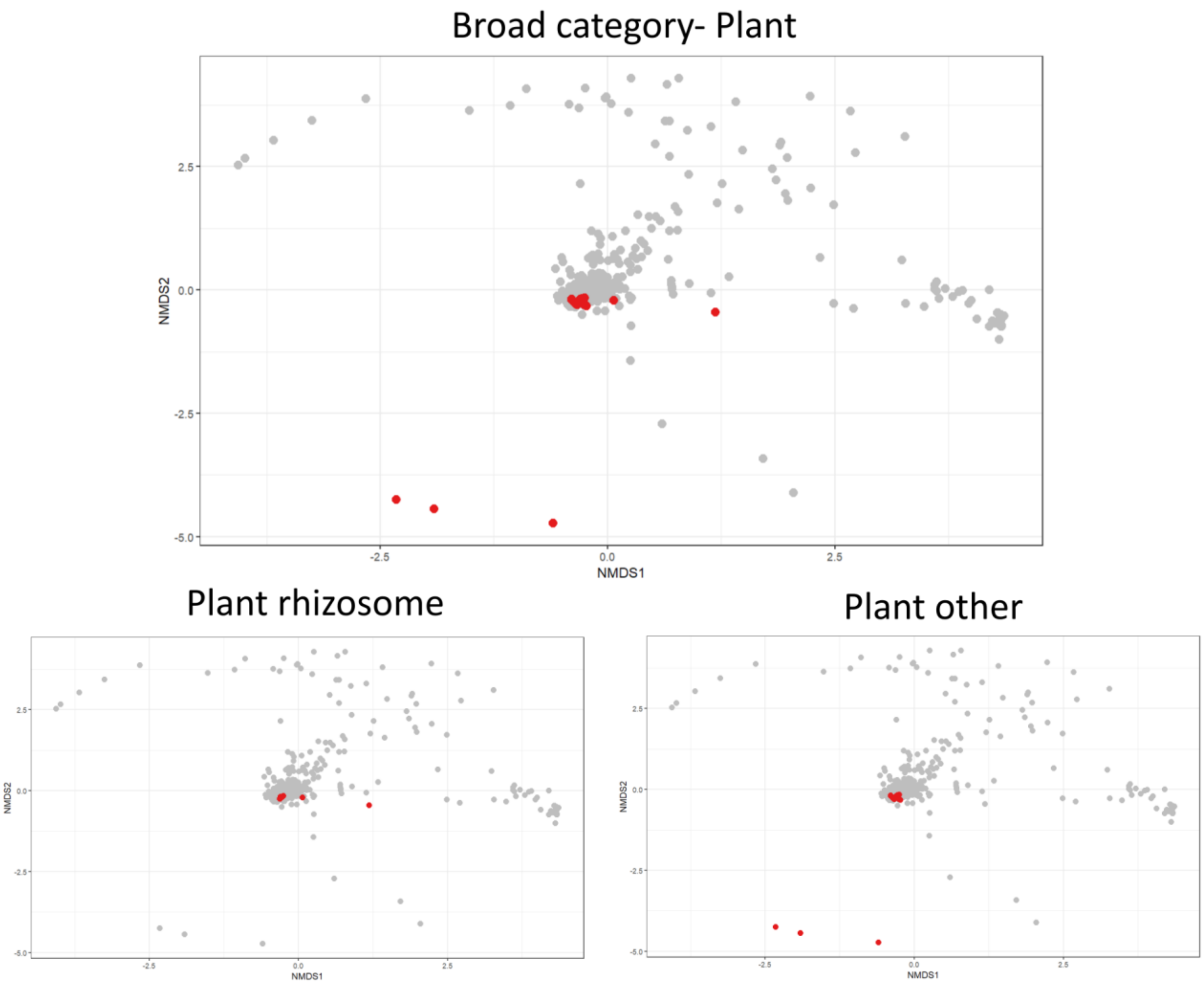

Figure 18: NMDS plots of the subsystem level 2 SUPERFOCUS results. The category listed is shown in red, while all other metagenomes are shown in grey.

Functional profiles are more conserved than the taxonomic profiles (Turnbaugh et al. 2009). Core genes involved in fundamental cellular processes, such as replication and structural maintenance, are ubiquitous across environments. In contrast, niche-specific genes, including those involved in virulence, will be more informative for our models. For example, virulence factors are enriched in host-associated environments where they play key roles in host-microbe interactions (Dinsdale et al. 2008). Although functional genes are typically less sensitive to environmental fluctuations, specific components of the functional profile, such as mobile genetic elements and phage-associated genes, exhibit variability associated with geographic isolation (Haggerty & Dinsdale 2017; Dinsdale et al. 2008).

Overall, the clustering patterns we observed (Figures S11-S17) indicate that our environmental categories are sufficiently distinct to support machine learning classification. However, model performance will likely vary depending on the taxonomic or functional levels used, and the machines will be able to learn some environments more easily than others.

Model benchmarking and algorithm comparison

We compared three classification algorithms (random forest (Breiman 2001), logistic regression (McFadden 1972), and XGBoost (Chen & Guestrin 2016)) across both taxonomic and functional datasets.

Random forest is an ensemble learning method based on decision trees, where input features are used to split the data into increasingly homogeneous subsets iteratively. Each decision tree within the ensemble contributes a single classification vote, and the final prediction is determined by majority voting across all trees. This approach is computationally efficient, resistant to overfitting, particularly in settings with moderate dataset sizes, like the one we employ here, and offers interpretable insights into feature importance and decision pathways. Unlike linear models, random forest allows the identification of hierarchical, non-linear relationships between features and class labels.

XGBoost, by contrast, implements gradient-boosted decision trees where each tree contributes a weighted score, and subsequent trees are iteratively trained to correct residual prediction errors of the ensemble. This model is highly optimised for performance but less interpretable than a random forest. Furthermore, it offers fewer tuneable hyperparameters accessible to end-users compared to the extensive configurability of random forests.

Logistic regression, on the other hand, is not a decision tree based method, and instead provides the odds of a particular outcome (Sperandei, 2014). This method can be faster than large forests. However, with logistic regression models you cannot see exactly what the model is learning. With decision trees you can see exactly what causes each branch point. This lack of transparency can allow the model to learn from unintended data, if proper care is not taken.

Hyperparameter tuning for the machine learning models was conducted to optimise model performance, with full parameter specifications and benchmarking results provided in the Supplementary Data.

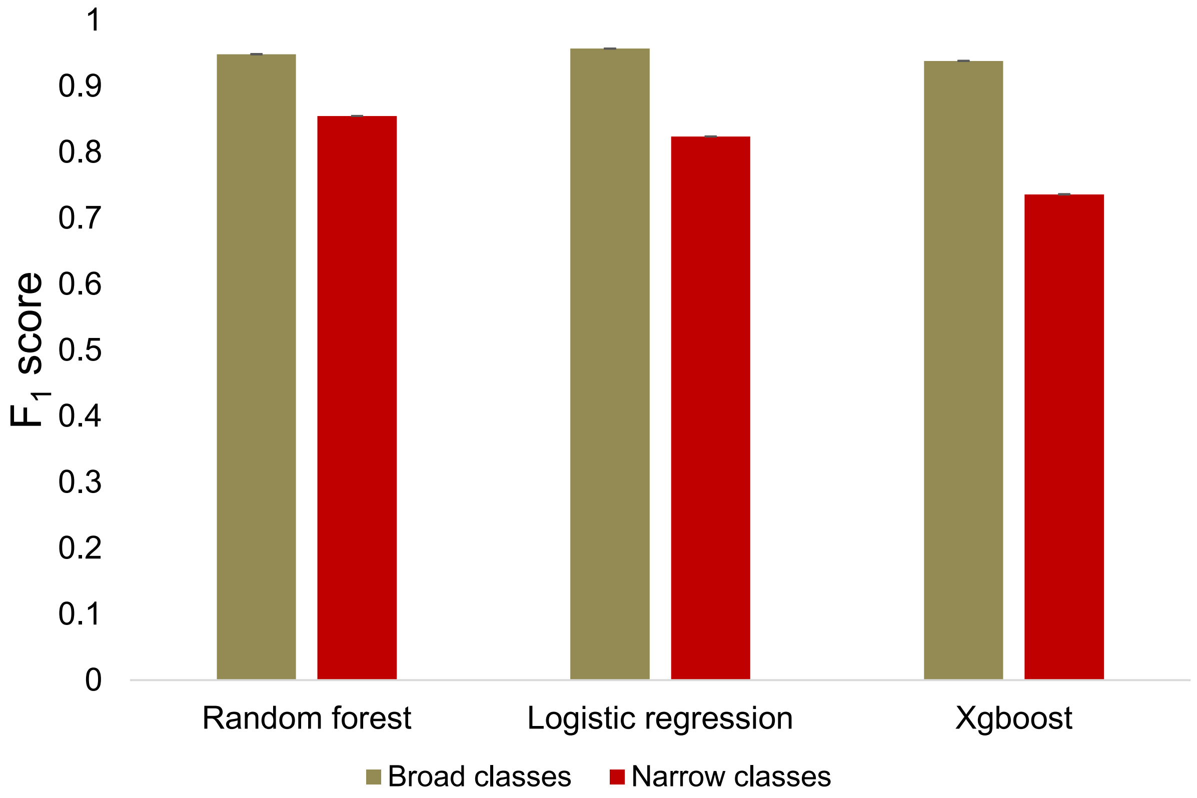

Figure 19: F_1 scores for three different models: random forest, logistic regression, and XGboost for both the broad and narrow class sets using all features.

Performance was assessed using F_1 scores for models trained on broad and narrow environmental categories (Figure 19). Random forest outperformed or matched the other algorithms in both contexts, particularly for narrow categories, and offered additional advantages, including model interpretability and computational efficiency.

Given its superior performance and transparency, random forest was selected for downstream analyses. Fourteen models were trained using either taxonomic (strain, species, genus, family) or functional (subsystem levels 1, 2, 3) data and evaluated on broad and narrow classification tasks.

Classification performance and feature comparison

We compared the taxonomic and functional profiles generated by FOCUS and superFOCUS, as they are generated independently for each dataset.

Receiver operating characteristic area under the curve (ROC-AUC) analyses demonstrated strong performance across all models (Figure 20). Taxonomic models marginally outperformed functional models, and broad classification tasks achieved slightly higher scores than narrow ones.

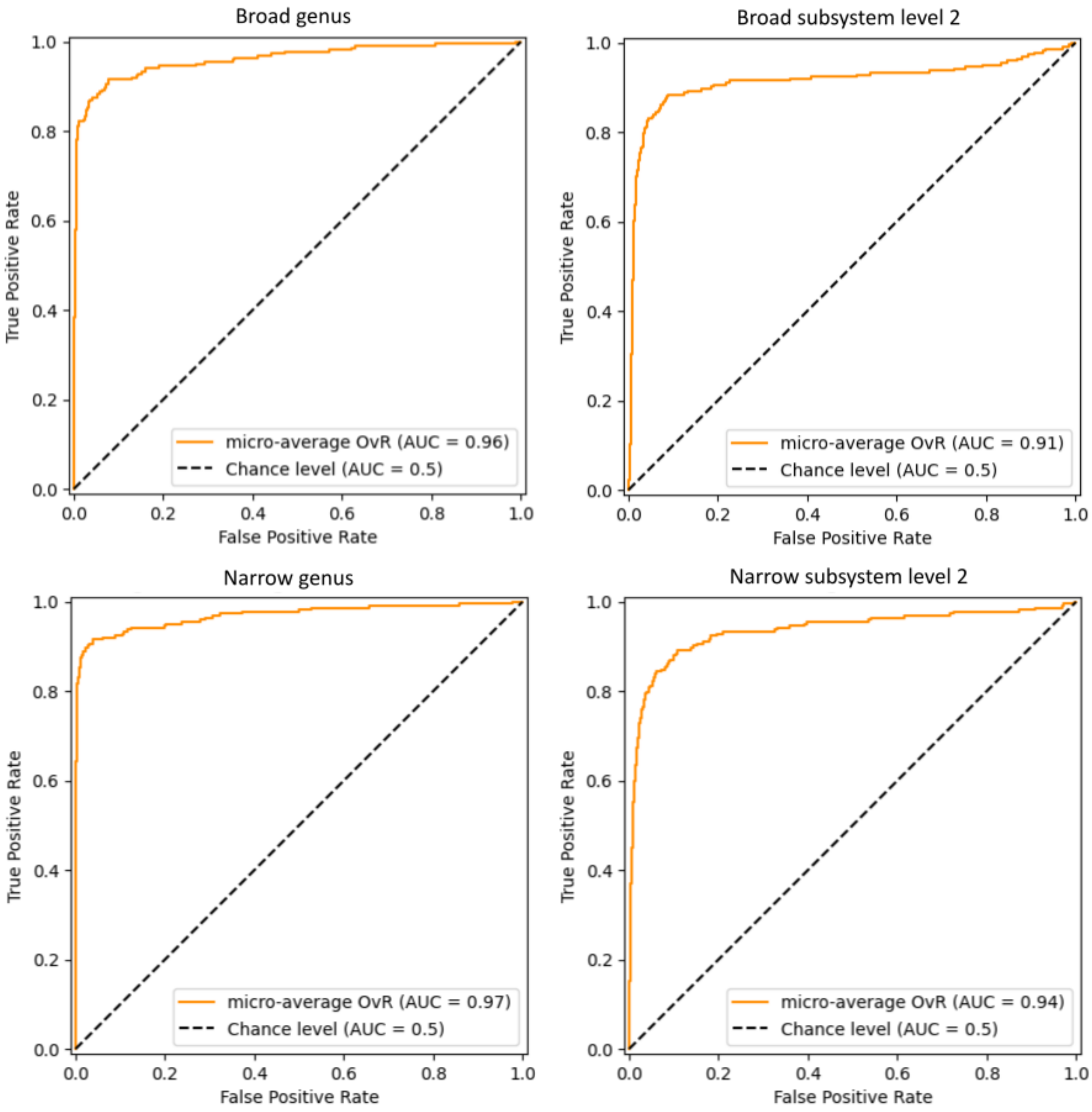

Figure 20: ROC-AUC curve showing the slight differences in the AUC values of broad class sets vs narrow class sets, and taxonomic profile vs functional profile. The orange line represents the average combined results of the OvR (one vs the rest) for each class (12 board classes, 24 narrow classes). The dashed line (Chance level) represents the results expected for random predictions.

Evaluation of model metrics revealed similar trends: all models achieved high accuracy, but F_1 scores were consistently lower for narrow categories (Figure 21). Specific categories were more readily classified than others. For instance, the "*marine*" class exhibited lower performance across models, whereas "*soil*" and "*animal*" classes performed well at specific taxonomic and functional levels (Figures 22 and 23). Our comparison suggests that the narrow subclasses are too similar to each other, while the broad-level classification captures most of the variation in functional and taxonomic profiles between environments.

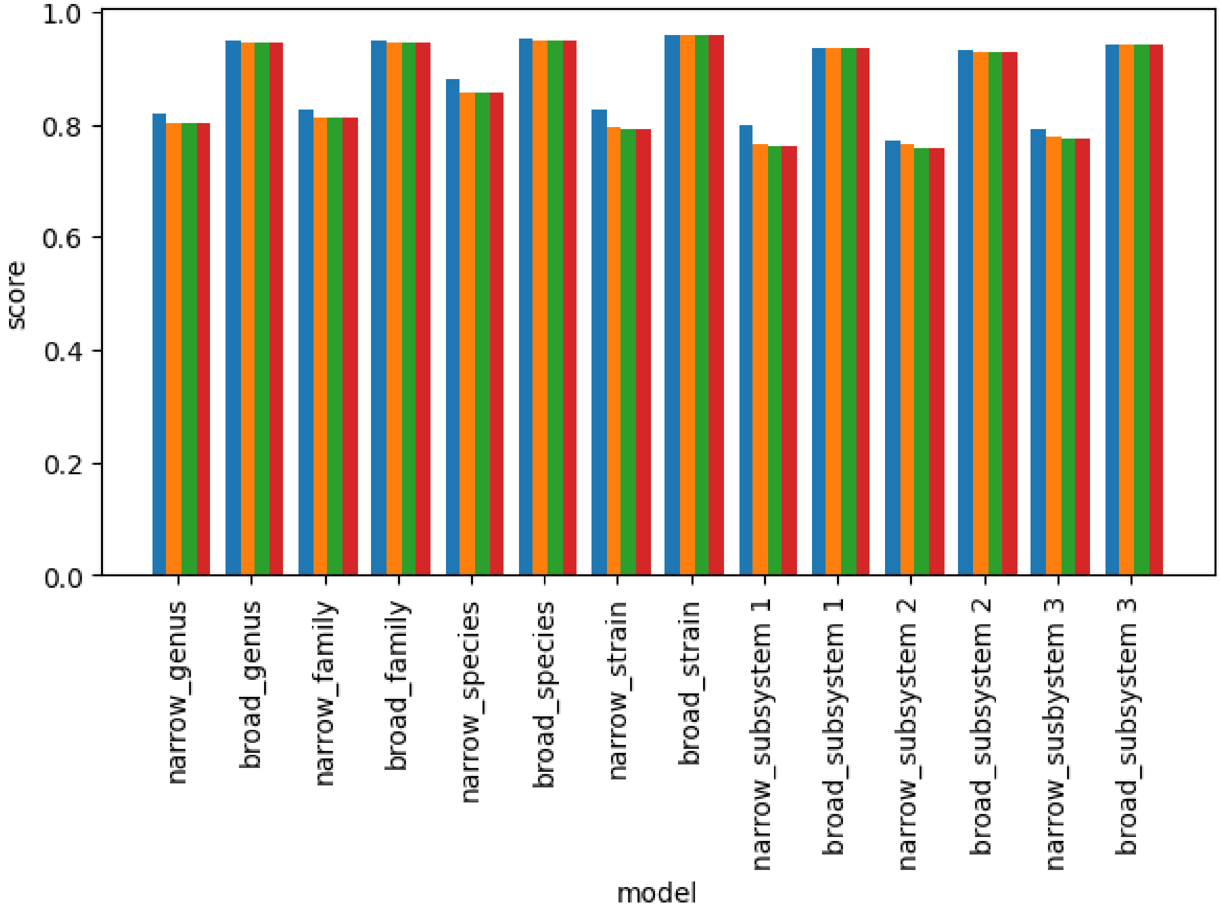

Figure 21: The precision (blue), recall (orange), accuracy (green), and F_1 (red) scores for each model.

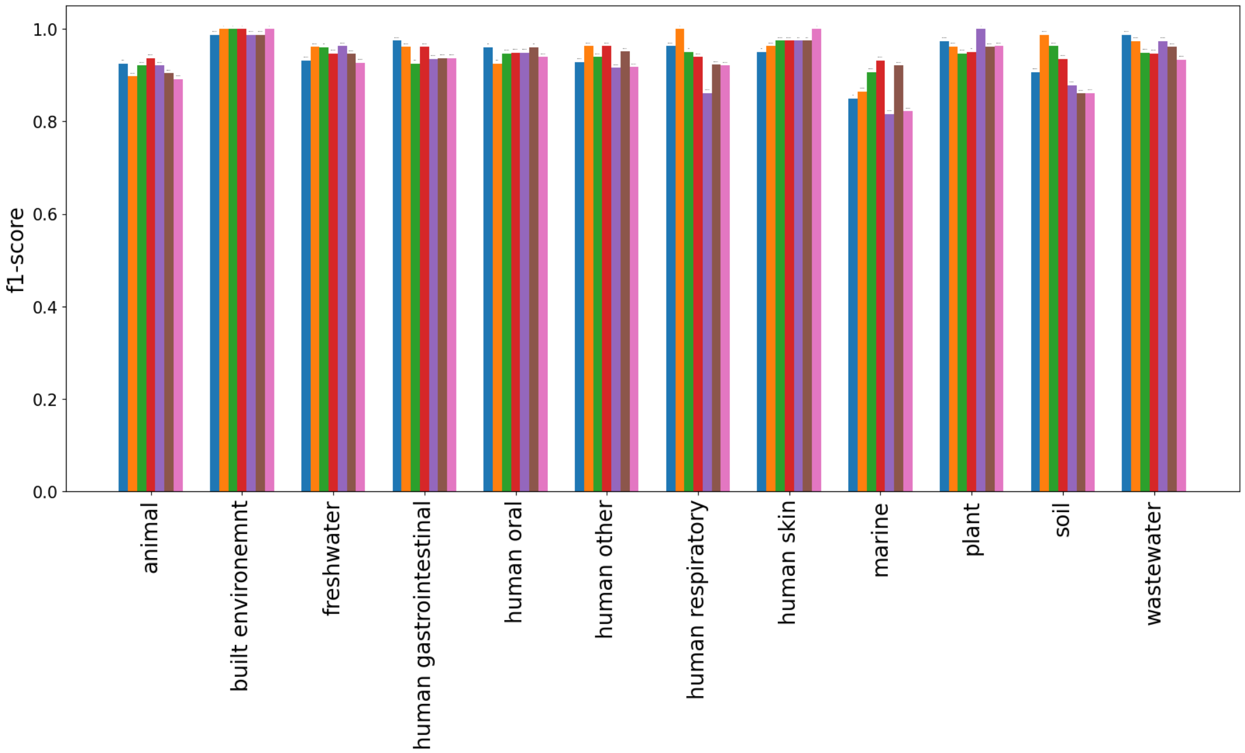

Figure 22: F$_1$ score per category for each broad class model. From left to right: species (blue), genus (orange), family (green), strain (red), subsystem 1 (purple), subsystem 2 (brown), subsystem 3 (pink).

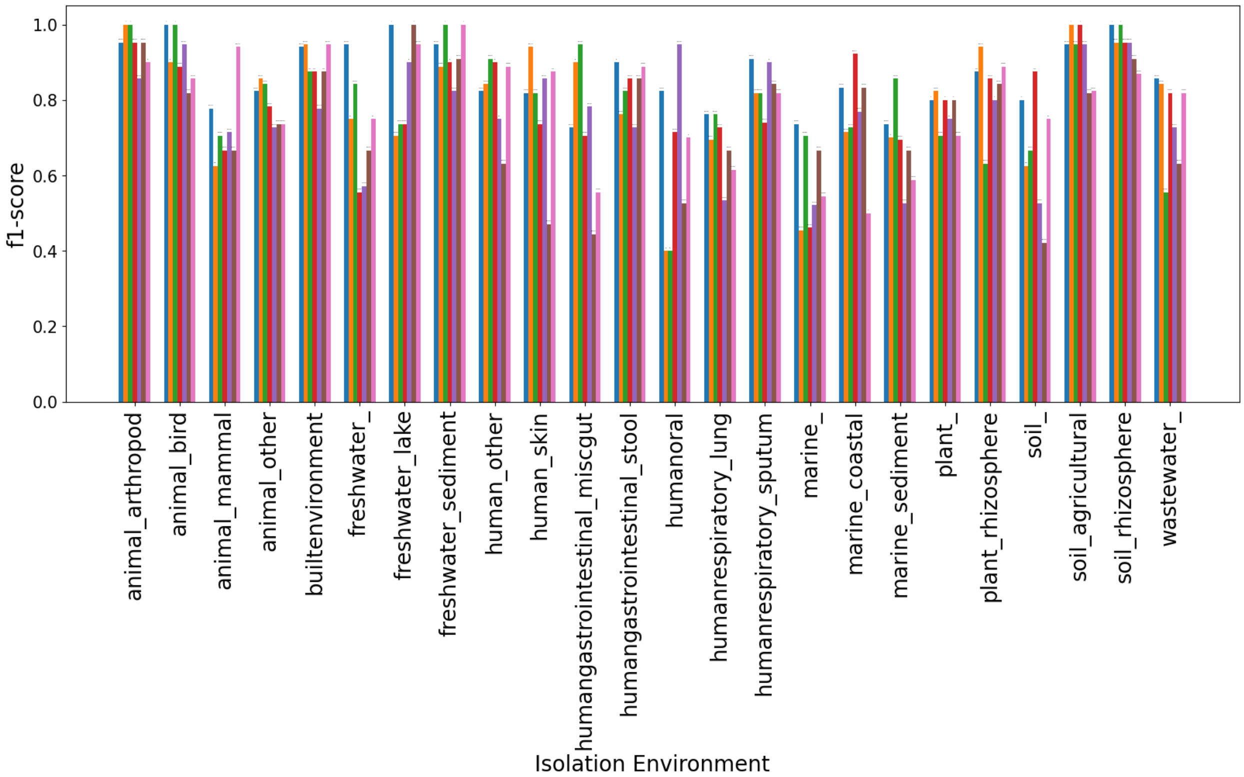

Figure 23: F₁ score per category for each narrow class model. From left to right– species (blue), genus (orange), family (green), strain (red), subsystem 1 (purple), subsystem 2 (brown), subsystem 3 (pink).

Clustering patterns observed in the NMDS analyses (Figures S11-S17) further supported differential classification performance across narrow environmental subclasses. Distinct compositional separation was evident for categories such as "*Animal:Bird*" and "*Animal:Arthropod*." At the same time, "*Marine:Other*" and "*Marine:Coastal*" displayed considerable overlap, suggesting a continuum of microbial community structure in coastal and transitional marine environments.

F_1 score distributions among narrow classes were highly variable. Some subclasses achieved higher classification performance than their corresponding broad categories, indicating that fine-scale environmental distinctions can improve model discrimination when underlying microbial differences are sufficiently pronounced. In contrast, other Animal-associated subclasses were classified with lower accuracy than the aggregate "*Animal*" category, which likely reflects the shared microbiota between mammals and human-associated samples, given their frequent cohabitation and overlapping ecological niches (e.g., pets, livestock). The catch-all "*Animal:Other*" category was taxonomically diverse. It included infrequently sampled groups such as marine vertebrates, which plausibly share microbial features with their surrounding aquatic habitats, further complicating accurate classification (Austin 2006).

Categories labelled as “other” or “miscellaneous” consistently exhibited the poorest performance across models. These groups contain heterogeneous and ambiguous metadata, as they encompass samples that do not fit clearly into defined environmental classes. To investigate whether these categories overlapped with more well-defined classes, additional models were trained, excluding the “other” categories, and subsequently used to predict their environmental origin. Although this exclusion slightly improved model performance on the training set, yielding accuracy comparable to that of broad-class models, prediction confidence and accuracy for the excluded “other” samples remained low. In many cases, predictions failed to map even to the correct broad class, indicating that these samples likely represent ecologically distinct or compositionally intermediate microbiomes. These findings suggest that although removing ambiguous categories may streamline training, the residual complexity of such samples remains a barrier to accurate environmental classification.

Misclassifications

While global performance metrics, such as F_1, provide a quantitative overview of model accuracy, they do not reveal specific patterns of misclassification. To explore these, we examined confusion matrices for each model (Figure 24). As anticipated, perfect classification was not observed, reflecting the inherent complexity of microbiome data and avoiding concerns of overfitting. Random misclassifications were present; more notably, consistent misclassification patterns emerged, particularly among narrow environmental categories.

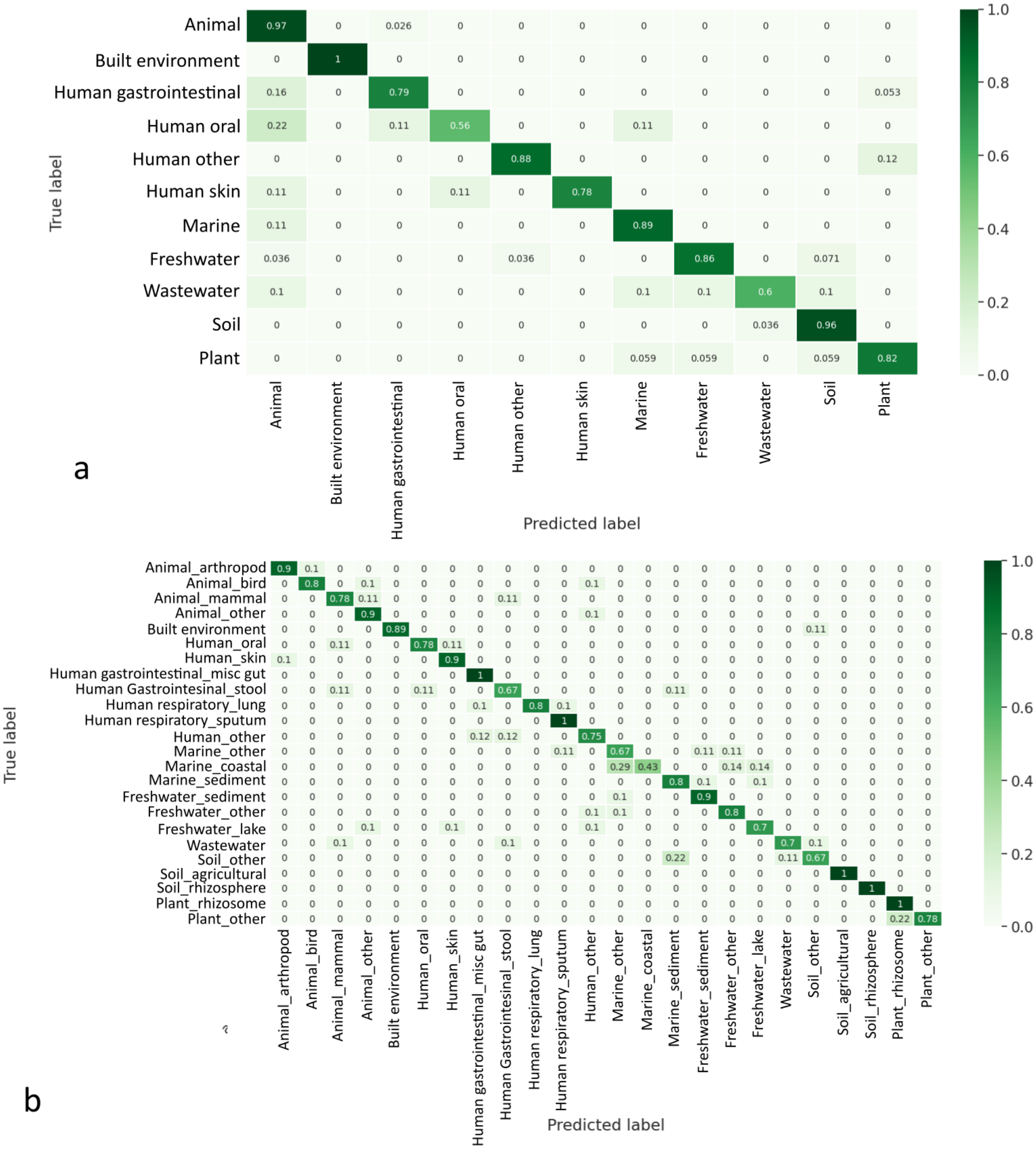

Figure 24: confusion matrices for the a) broad genus and b) narrow genus model. The Y-axis shows the actual label, while the X-axis shows the predicted label. The colour of the heatmap indicates how often the model predicted a specific class for each actual class. Matrices for other models are available in the supplementary data (Figures S19-S24).

These patterns were most apparent within subclasses nested under the same broad environmental category, suggesting that misclassification often occurred between ecologically similar or physically connected environments. These misclassifications are consistent with the known continuum of microbial community composition across environmental gradients, the lack of strict ecological boundaries in many real-world settings, and reflect biological overlap (Alsop et al. 2014).

However, analysis of these misclassifications provided further insight into the model's behaviour. For instance, samples labelled as “*marine:other*” were frequently misclassified as originating from freshwater environments. Similarly, “*marine:coastal*” samples also showed occasional overlap with freshwater predictions. These patterns likely reflect the inclusion of estuarine or brackish water samples within the marine classes, reflective of environments that share microbiological characteristics with freshwater systems due to lower salinity and transitional geochemical conditions.

Conversely, other closely related environments were well resolved. The subclasses “*soil:agricultural*,” “*soil:rhizosphere*,” and “*plant:rhizosphere*” demonstrated clear separation and high classification accuracy. These results suggest that, where metadata accurately reflects distinct ecological contexts, models can successfully capture fine-scale microbiome structure and assign environmental origin with high precision.

Feature importance and model interpretation

To understand the contribution of individual taxa and functions to model performance, we assessed feature importance across all random forest models. Feature importance provides a quantitative measure of each feature’s influence on model predictions (Musolf et al. 2022). It is particularly relevant in complex microbiome datasets, which often include a large number of features, especially at fine taxonomic resolutions such as species and strain levels.

Two commonly used metrics were employed: Gini importance and permutation importance. Gini importance is computed as the average reduction in node impurity (Gini index) attributable to each feature across all decision trees in the ensemble. However, this method is biased towards features that occur more frequently in the dataset (Nicodemus 2011). In contrast, permutation importance evaluates the decrease in model performance resulting from the random permutation of a given feature, offering an unbiased, albeit more computationally intensive, alternative (Musolf et al. 2022).

Feature importance patterns varied depending on the taxonomic level and the level of environmental granularity. At the family level, where the total number of features was comparatively limited, the most informative features were largely consistent between broad and narrow classification tasks. In contrast, species- and strain-level models exhibited substantial divergence, with minimal overlap among the top-ranked features between classification schemes. This variability likely reflects the higher dimensionality and environmental specificity of features at finer taxonomic scales.

In taxonomic models, members of the genera *Anaplasma*, *Mycoplasma*, and *Coprococcus* emerged as consistently important features, particularly at the genus and species levels. *Coprococcus* and *Anaplasma centrale* were among the top contributors in species-level models.

These genera are frequently detected in human-associated microbiomes, with *Mycoplasma* comprising several pathogenic species, and *Coprococcus* recognised as a commensal butyrate producer that may be important in human health (Liddicoat et al. 2024). Notably, all three taxa also occur in diverse animal hosts and environmental samples, underscoring their ecological ubiquity and potential utility in classification models (Breitbart et al. 2003; Johnson et al. 2019).

For models trained on functional profiles, genes involved in protein biosynthesis and virulence functions were among the most influential features. Interestingly, there was greater divergence between Gini and permutation-based rankings for functional models compared to taxonomic models. Gini importance consistently ranked phage-associated genes as top contributors, while these same features were deprioritised by permutation analysis, likely due to their relatively high frequency across samples.

Overall, the most informative features across all models were those that exhibited intermediate prevalence and moderate ecological breadth. Rather than relying on rare, environment-specific markers, models preferentially exploited quantitative variation in the relative abundance of core taxa and functions that span multiple habitats. Our work suggests that subtle shifts in common microbiome constituents offer stronger predictive signals than uniquely associated features in sparse datasets.

Model performance on independent metagenomic data

To assess model generalisability, we evaluated classification performance on an independent validation set comprising metagenomes from all broad environmental categories except "*Wastewater*", due to the lack of new wastewater samples in the MGnify database. To provide a robust test of model performance beyond the original training and testing partitions, we separately downloaded this dataset from the MGnify database, and it uses previously unseen samples.

Confusion matrix analysis (Figure 25) revealed that misclassifications were infrequent and dispersed across classes. As we saw previously, the errors often aligned with known ecological similarities. The fine-scale taxonomic levels, such as species and particularly strain, exhibit substantial intra-environmental variability and are frequently geographically localised or influenced by host-specific factors. For example, variation in human gut microbiomes can be driven by diet, age, and geographic origin (Turnbaugh et al. 2009; Lassalle et al. 2018). These sources of heterogeneity can obscure ecological signals in unseen datasets. In contrast, functional gene profiles are generally more conserved across samples and environments (Rodriguez-Brito et al. 2010). Different microbial taxa may encode overlapping functional repertoires to fulfil similar ecological roles, resulting in greater stability of functional classifications. Functional redundancy likely explains why those models outperformed their taxonomic counterparts when applied to novel data.

Overall, classification accuracy on the independent dataset was marginally lower than on the original test set (Figures 24-25). Nevertheless, categories that had performed well during model training (e.g. "*Soil*" and "*Animal*") maintained relatively high predictive accuracy. Exceptions included the "*Built environment*" and "*Human:Other*" classes, both of which showed substantial reductions in performance. The “*Built environment*” category is inherently heterogeneous, encompassing a range of urban and indoor environments with diverse microbial exposures. Furthermore, it was sparsely represented in the MGnify database, and associated metadata were frequently ambiguous or insufficiently specific. These factors likely contributed to reduced model performance for this group and underscore the importance of metadata quality in environmental classification tasks.

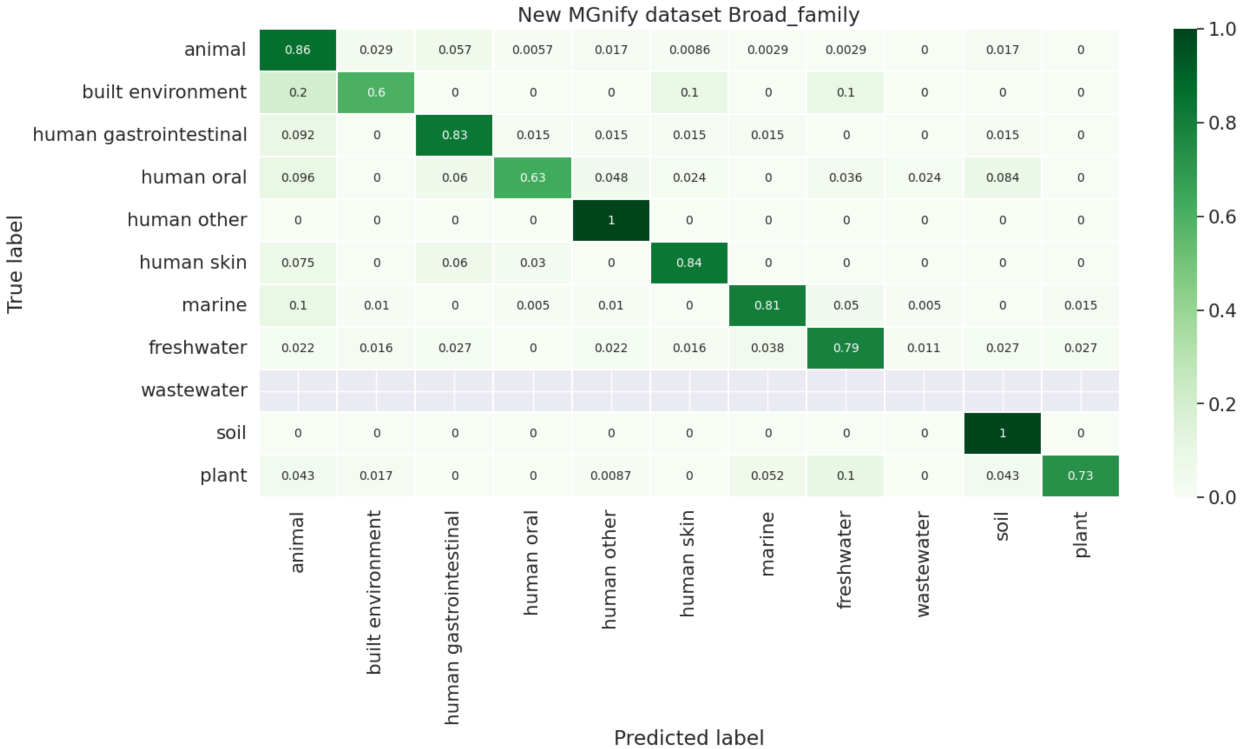

Figure 25: confusion matrix for the broad_family model tested on the new test dataset.

Prediction confidence and ambiguity in closely related environments

To further assess model reliability and internal decision-making, we applied the prediction probability function of the random forest classifier to an independent validation set. This function estimates prediction confidence by reporting the proportion of decision trees that vote for each class label. In datasets with a high number of features and only moderate ecological separation between categories, variation in tree-level predictions is expected. Indeed, we observed frequent prediction ambiguity in cases where samples were misclassified between ecologically or phylogenetically related groups. For example, samples annotated as *“Animal:Mammal”* and *“Human gastrointestinal”* often received nearly equivalent support from the classifier, suggesting a high degree of biological similarity between these categories.

Since the mammal categories consisted mostly of stool samples, it is perhaps unsurprising that it is confused most often from human gut/stool samples. This overlap is likely due to a shared evolutionary history. Host phylogeny is a well-established determinant of gut microbiome composition. Humans, as mammals, exhibit microbiome profiles that are more similar to those of other mammals than to those of non-mammalian animals. As such, microbial communities derived from human and non-human mammalian guts may be difficult to distinguish based solely on taxonomic or functional features. Consistent with this, the "*Human gastrointestinal:Stool*", *"Human gastrointestinal:Misc gut"* and "*Animal:Mammal*" categories demonstrated some of the lowest classification accuracies in the validation dataset, and their prediction probability scores were often similar, regardless of the actual label. These findings highlight the biological limitations of environmental classification, particularly when host-related categories exhibit extensive overlap in their microbiomes. Adding more varied mammal metagenomes to the dataset in the future may help the model to separate these three groups, as other areas of the body may not be as similar as the gut.

While the overall classification performance of models trained on broad environmental categories was generally robust, notable exceptions were observed for the "*Marine*" class when using functional data. Functional models exhibited reduced performance on marine samples, even when the associated prediction confidence exceeded 0.5. This suggests that the classifier reached a high internal consensus despite generating incorrect predictions. Among the functional models, only those trained on subsystem level 1, which represents the broadest tier of functional annotation, were able to classify marine samples with moderate accuracy. In contrast, marine samples remained challenging to classify when using subsystem levels 2 and 3, likely due to functional redundancy and limited specificity of gene annotations in aquatic environments. Other broad environmental classes were predicted accurately across models, regardless of whether taxonomic or functional data were used.

Overall, this was the first study to use machine learning to categorise metagenomes by isolation source using their functional profiles. We showed that while the taxonomic profile produced a stronger predictive performance, but that the functional profile was more generalisable to new datasets. Since the functional profile is more robust to new data, it is the preferred method. However, care should be taken if the functional profile of the metagenome of interest contains many unknown. This algorithm will allow other to utilise metagenomes that would otherwise be unusable for environmental comparisons due to their lacking metadata, potentially opening up vast reservoirs of data stored in online databases.

Acknowledgments

We would like to acknowledge the support provided by Flinders University for HPC research resources(Flinders University 2021).

Funding

Awards from the NIH NIDDK RC2DK116713 and the Australian Research Council DP220102915 and DP250103825 to R.A.E. supported this research.

Ethical statement

Ethical approval was not required because our study did not involve the collection of new biological samples or the use of human or animal subjects. All data used were from publicly available metagenomic sequences obtained from the MGnify database.

Discussion chapter

The human microbiome is a complex ecosystem, with the microbes thought to outnumber human cells by a factor of ten (Turnbaugh et al. 2007). There are hundreds of species in the human microbiome, but many are exclusive to different areas of the body. For example, *Enterobacte*r species are usually found in the gut, and *Helicobacter pylori* is usually confined to the stomach.

The composition of the human microbiome also varies with factors such as diet, lifestyle, medication, and geography. It also changes with host health. Pathogenic species outcompeting the ‘normal’ microbiome are associated with disease, and even species that would otherwise be harmless in their regular environment can become pathogenic when they gain access to a different area of the body.

Phages are the smallest and most numerous members of the microbiome. They consist largely of lytic and temperate phages. Lytic phages act like a typical virus. They infect their host, then hijack the host’s replication machinery to create more virions before lysing the cell and releasing the new virions to infect more cells. Temperate phages, on the other hand, don't always do this. These phages have the option of integrating into their host’s genome after infection. While integrated into their host's genome, they are replicated as their host is replicated, and as long as the phage remains intact, it can switch to the lytic lifecycle when conditions change.

There are a few hypotheses that attempt to explain what makes a temperate phage choose the lytic or lysogenic strategy. Piggyback-the-Winner, Kill-the-Winner, and Piggyback-the-loser all explain what factors influence the phage's decision, but due to looking at very different environments, they produce slightly different conclusions. However, all can be observed in nature.

My question then became, how do the lysogenic phages vary by body site? Do some areas of the body favour the lytic lifecycle while others favour lysogeny? As different body sites have different environmental conditions, there should be variation in the phages. To answer these questions, I predicted the phage regions of over 14,000 bacterial genomes that were isolated from humans.

I found that some regions did appear to favour lysogeny, while some regions were more skewed towards lysis. In particular, the stomach had almost no phages at all. The average amount of prophages in different areas of the human body largely follows the Piggyback-the-Winner or Piggyback-the-Loser/Kill-the-Winner hypotheses. The Piggyback-the-Winner hypothesis suggests that phages in high-nutrient and high-diversity environments, such as the human gut, will be more likely to remain lysogenic (B. Knowles et al. 2016). Kill-the-Winner suggests that phages in a high-nutrient, low-diversity environment will switch to the lytic lifestyle (Thingstad 2000). Finally, the Piggyback-the-Loser/refugium hypothesis suggests that phages in a low-nutrient environment will remain lysogenic (B. Knowles et al. 2016). The different areas of the human body and host health are split across these nutrient and diversity groups The stomach was very different to the other areas of the body, and possibly due to its uniquely harsh and acidic conditions, has one of the lowest bacterial concentrations of the human body at 10^{3-4} cells/ml (Sender, Fuchs, and Milo 2016). It also had by far the least amount of prophage DNA per genome. It is a microbiome that is often dominated by a single species, particularly in the case of *H. pylori* infection. It also receives a steady supply of nutrients, so it is unsurprising that it follows the Kill-the-Winner scenario. The vagina, on the other end of the spectrum, was the body site with the most average prophages DNA. It is also an environment that is relatively stable and has a much higher bacterial concentration of 10^{10-11} cells/ml, but is still often dominated by a single genus (Chen Chen et al. 2017; Saraf et al. 2021).

I found that some species also had little to no prophages. One of these species, *H. pylori*, was found in the stomach, which I hypothesise follows KtW dynamics, as, while it is a harsh environment, once infected by *H. pylori,* it is dominated by a single species. The other two species that contained no prophages were *M. plasmodiae* and *C. trachomatis.* The *C. trachomatis* samples came mostly from vaginal samples which had an overall high amount of prophages, likely because it is a stable environment that is generally dominated by *Lactobacillus* species (Saraf et al. 2021).

There were significant differences between genomes from healthy humans and genomes from symptomatic humans, but whether bacteria from the dysbiotic microbiomes were more likely to have more prophage DNA varied by body site. This leads to the question of whether specific species change their phages in response to a change in host health. Some bacteria can be opportunistic pathogens and are found in samples from both symptomatic and asymptomatic hosts in the dataset. Looking specifically at those genomes, I find that most of them gain prophage DNA when they are in symptomatic hosts. It is already known that *S. aureus, E. coli, V. cholerae* and others contain prophages that allow them to produce toxins (Rodríguez-Rubio et al. 2021; Waldor and Mekalanos 1996; Cao et al. 2012), so it is perhaps not surprising that five out of the eight species in the dataset that were found in both symptomatic and asymptomatic hosts had significantly more prophage DNA when sampled from symptomatic humans.

The variety of prophage DNA in the data could also include different types of prophages. There are a few different kinds of prophages that could be populating the genomes in the dataset. There are the typical phage, phage-plasmids, and satellite phages. Phage-plasmids are a sort of combination of plasmid and phage. They exist in the host cell as plasmids instead of integrating into the bacterial chromosome and carry both plasmid genes and phage genes and it is estimated that 5% of phages and 7% of plasmids are these phage-plasmids (Pfeifer and Rocha 2024; Pfeifer et al. 2021). Satellite phages are small phages that often don't encode structural genes and rely on their ‘helper’ phage infecting the same bacteria so they can borrow their genes (deCarvalho et al. 2023; Dehò, Ghisotti, and Others 2006). Most satellite phages are lysogenic, though there are some lytic ones (deCarvalho et al. 2023). A cursory glance at the prophage regions predicted by PhiSpy suggests that it picks up all three kinds of prophages, as well as potentially incomplete phages..

Overall, the human microbiome is incredibly complex and is shaped by a myriad of environmental factors, and the lysogenic phageome is no exception. But the lysogenic phageome is of particular note because when temperate phages infect their hosts, they integrate into the bacterial DNA and become prophages, with the host/phage combination being referred to as a lysogen. Errors in replication can cause genes to be swapped onto the phages, allowing them to spread genes further. One of the most important groups of genes in terms of the human microbiome and health is antimicrobial resistance (AMR) genes. These genes have provided one of the most difficult challenges to human health since the discovery of antibiotics.

While integrated, the bacteria have access to the phage’s genes, which can give the bacteria genes they would otherwise not have access to. One well-known example of this gene sharing is the Shiga toxin genes in Escherichia coli. These genes are provided by lambdoid bacteriophages (Rodríguez-Rubio et al. 2021). Another benefit they provide to their hosts is superinfection exclusion, where the host is protected from infection by further phages (Mavrich and Hatfull 2019). This can potentially cause problems for techniques like phage therapy which rely on infecting bacteria with phages. Having an overview of how common phages are in different situations can help inform researchers on whether they need to consider the effects of prophages.

It is also important to know what genes prophages are providing to their hosts, as they can have detrimental effects and even be spread further through the microbiome. PhiSpy predicted many different kinds of phages, and they all have the potential to carry AMR genes, with at least a couple of potential instances of each in the dataset. PhiSpy also has the potential to predict non-functional incomplete phages as well, and these phages can also carry AMR genes. This should not affect the ability of the bacteria to utilise these genes and still suggests that they may have arrived by phage, but these incomplete prophages do not have the capability to transfer the AMR genes further.

While more recent papers are suggesting that AMR genes in phages are very common (Kondo, Kawano, and Sugai 2021), my results suggest that that isn’t quite the case. AMR genes in phages were not rare in the dataset, but they were uncommon, though a wide variety of genes were found, suggesting that the transfer of AMR genes into phages happens quite often. The discovery of genes that were found the bacterial DNA of some species but the prophage regions of others was interesting as it shows AMR genes being acquired by prophages and moving them into new species, providing evidence for the theory of the common gene pool. If AMR genes have found their way into the common gene pool then that is yet another way which these genes will slowly spread.The main limitations of my work were that I used a collection of human-isolated bacterial genomes as a proxy for the human microbiome and that the metadata of samples deposited in online databases was lacking and required a lot of manual curation to parse useful information.. While the number of genomes and different species I had access to was large, it was biased towards medical samples and pathogenic bacteria, as that is what human studies tend to focus on. In particular, *Staphylococcus aureus* and *Escherichia coli* made up the majority of genomes in the database.

The databases largely lacked samples from a healthy human microbiome. However, it can also be said that we lack a hard definition of ‘healthy’. Currently, ‘healthy controls’ are often merely subjects that do not have the disease being studied, and they may have other ailments, as there are very few perfectly healthy people (Marchesini, Marchignoli, and Petta 2017). The screening process for healthy controls can also be less comprehensive than screening for patients (Pavletic 2020). This makes the healthy human microbiome difficult to quantify.

There are a few potential solutions to the metadata issue but all are flawed in various ways. When submitting data to online databases there are fields for metadata, though these are numerous and optional. Increasing the rate of people using these fields could be done by either simplifying them by reducing the number of fields, or making some field mandatory. These both have their own issues however. Making uploading more difficult could turn people off from uploading, and merely simplifying the process will not have as strong of an uptake. Scientists working in or with industry can face restrictions on the use of their data which will always remain a factor so long as industry has a place in science. This also leads into the question of ethics in data collection and storage. Human samples face strict ethics guidelines regarding identifying information. Having too much metadata can make the person behind the sample identifiable, particularly. Even data like location and disease state can be sufficiently identifiable in some situations.

One solution that avoids the issue of needing more metadata could be to create a database for curated genomes with complete metadata, but this solution falls on the shoulders of a few people instead of the whole, and doesn’t solve the problem of genomes lacking metadata, just makes it easier to find the ones that are properly labelled.

Another potential solution is to use metagenomes instead. A metagenome is the genetic material from an entire environment. Metagenomes are created by extracting all the genetic material from a sample and ‘shotgun’ sequencing it, then piecing the separate genomes back together from the sequenced fragments. This approach allows us to get DNA from the microbial dark matter, and entirely new candidate phyla of bacteria have been discovered this way.

However, this solution presents its own problems, and there are fewer metagenomes deposited online than genomes, which compounds many of the issues they have. Metagenomes require significantly more time and storage space to download due to their large size. Many of the metagenomes in the SRA were also of low quality and were almost entirely cut when using the standard settings of Trimmomatic.

But the biggest problem with metagenomes from databases was still the lack of metadata.

The fact that there are fewer metagenomes deposited online than genomes makes the problem of poor metadata worse. Instead of ~20,000 samples, my attempt to download all the metagenomes from MGnify yielded 856 studies that could be identified as human, though these studies often contained multiple metagenomes.

To try and alleviate that problem for future studies, I attempted to automate the categorisation of metagenomic sequences using machine learning. I attempted to sort metagenomes by their isolation environment and used 7 different taxonomic or functional profiles and two levels of environmental granularity. I found that both taxonomic and functional profiles could be used to train the algorithms with a similar level of accuracy. However, the broader environmental classes produced a more accurate model than their narrow subclasses, even though some of the narrow classes were more taxonomically/functionally conserved. This could be due to the larger average sample sizes of the broad classes, as the sample sizes for many environments were severely limited by the previously mentioned issues of few metagenomes with descriptive enough metadata. This research, will be useful for curating metagenomes to make the data stored in databases more accessible for future research.

Significance and future directions

This thesis lays the groundwork for many future studies into the human phageome by providing an overview of how common integrated phages are in different areas of the human body and different regions of the world. We explored this further by examining what AMR genes they could be providing to their hosts, but prophages can provide other benefits to their hosts, such as protection from infection by other phages. It would be useful to determine how common superinfection exclusion is, particularly in pathogenic bacteria.

While there was a wealth of bacterial genomes found in GenBank, and I found that the amount of prophage DNA in the human microbiome varied significantly by body site, host health and other factors, there are still significant gaps in the dataset we used. Due to lacking metadata I was unable to specify what illnesses the human hosts were suffering from. The data set also lacked genomes from a wide variety of healthy people. This is particularly concerning as a healthy baseline is important for many studies. It would be useful to collect more data from healthy individuals to compile a more complete picture of the healthy human phageome. Collecting genomes from species that are not associated with the illness that symptomatic humans have would be useful to examine how the changes in the microbiome and the human host body affect the commensals. However, most genomes in GenBank are from studies investigating pathogenic bacteria that are most often the cause of the patient's ailment. To get enough samples to examine these questions would require collecting and sequencing our own samples.

The presence of AMR genes in prophages is significant as antimicrobial resistance is currently one of the greatest threats to human health, and knowing how it is spread between species is crucial to limit its proliferation. The next step for this work would be to determine whether the phages conferred a resistance phenotype to their hosts and whether the phages that did were able to return to a free viral state and infect new bacteria. If both these points are confirmed for any phages the next step would be to determine their host range and where in the world they are found.

Training an algorithm to sort metagenomes with incomplete metadata will allow many more sequences to be available for future studies, including being able to look at a more complete human microbiome through metagenomics. However, the algorithm could still be further refined by adding more training data. This would allow for more training data per isolation soucre to account for more finer-scale variations, and including more isolation source categories. Because our training dataset was limited to isolation source categories that contained more than 50 studies in MGnify, some environments were not as detailed as others or left out entirely, limiting the algorithms usefulness in those areas. Since some categories are very broad, having a more specific classification system would be more useful for a wider variety of studies. At this point the model does not have training data for many parts of the human body, or any foods. However, this requires more metagenomes that we can accurately classify into these narrower categories. Having more data may also make the model more resilient against factors other than environment that may be affecting the results, such as different techniques used by different labs.

Conclusions

The aims of this thesis were threefold: One, to investigate the potential for using machine learning to automatically sort environmental metagenomes by their isolation source, and compare the potential for the functional profile as the features for this training over the taxonomic profile, which had been explored previously. Two, to examine the prophages of the human microbiome and show how the lysogenic component of the human virome varies alongside the rest of the microbiome. Three- to explore what genes prophages are providing to their microbial hosts and determine how common it is for prophages to encode for antibiotic resistance genes.

To achieve these aims I gathered genomes and metagenomes from GenBank and MGnify and manually curated tens of thousands of genomes, analysed them with PhiSpy and AMRfinder+, ran statistical analyses, trained and tested several machine learning models on multiple sets of features, and pulled meaning out of terabytes of data, creating four papers, two of which have been published and two which have been submitted.

I found that the functional profile was even more informative than the taxonomic profile for training a machine learning algorithm, and I found that most of the features that were important to the algorithms were phage genes, which were found in varying concentrations in every environment, suggesting that variations in the amount of phage in a microbiome is an important indicator of environment.

In regards to the prophages in the human microbiome, I showed just how much they varied across the human body, with different areas of the body, even if they were physically linked, showing different lytic/lysogenic dynamics, and some regions and bacterial species have almost no detectable phages at all. The average amount of prophage DNA in bacteria was also affected by other factors such as the host of the health, even in species that are found commensally, and by geographical region, also even in species found across regions, showing that the lysogenic phageome is affected by many more environmental factors than just the isolation site.

Since commencing this project, the pendulum has swung back toward antimicrobial resistance genes being commonly found in phages, with papers published in the last couple of years yet again claiming that they are common. However, I found that AMR genes were quite rare in our dataset, with less than 7% of genomes in our dataset containing prophages that carried AMR genes. However, these genes were quite varied, encoding varying levels of resistance to 19 different antibiotic groups, and AMR genes in phages appeared to be selected for, especially in Asia and Oceania. Prophages containing AMR genes were most common in bacteria isolated from stool samples and least common in bacteria isolated from throat samples. I also found cases where specific AMR genes were encoded on bacteria genomes in some species but only found in the prophage regions of another species, suggesting that phages are carrying genes to species that did not contain them before, proving that they are yet another method that antibiotic resistance is spreading throughout the environment, and in the case of our study, within the human microbiome.

There are still several gaps in our knowledge of the microbiome and the human virome in particular, but my work has succeeded in filling a couple of those gaps and providing more to the foundation on which future work can be built.

Bibliography

Abellan-Schneyder, I., M. S. Matchado, and S. Reitmeier. 2021. “Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. mSphere 6: e01202-20.” https://scholar.google.com/citations?user=Q0imgU4AAAAJ&hl=en&oi=sra.

Akhter, Sajia, Ramy K. Aziz, and Robert A. Edwards. 2012. “PhiSpy: A Novel Algorithm for Finding Prophages in Bacterial Genomes That Combines Similarity- and Composition-Based Strategies.” *Nucleic Acids Research* 40 (16): e126.

Aziz, Ramy K., Bhakti Dwivedi, Sajia Akhter, Mya Breitbart, and Robert A. Edwards. 2015. “Multidimensional Metrics for Estimating Phage Abundance, Distribution, Gene Density, and Sequence Coverage in Metagenomes.” *Frontiers in Microbiology* 6 (May):381.

Bielaszewska, Martina, Evgeny A. Idelevich, Wenlan Zhang, Andreas Bauwens, Frieder Schaumburg, Alexander Mellmann, Georg Peters, and Helge Karch. 2012. “Effects of Antibiotics on Shiga Toxin 2 Production and Bacteriophage Induction by Epidemic Escherichia Coli O104:H4 Strain.” *Antimicrobial Agents and Chemotherapy* 56 (6): 3277–82.

Billaud, Maud, Quentin Lamy-Besnier, Julien Lossouarn, Elisabeth Moncaut, Moira B. Dion, Sylvain Moineau, Fatoumata Traoré, et al. 2021. “Analysis of Viromes and Microbiomes from Pig Fecal Samples Reveals That Phages and Prophages Rarely Carry Antibiotic Resistance Genes.” *ISME Communications* 1 (1): 55.

Blumberg, Kai L., Alise J. Ponsero, Matthew Bomhoff, Elisha M. Wood-Charlson, Edward F. DeLong, and Bonnie L. Hurwitz. 2021. “Ontology-Enriched Specifications Enabling Findable, Accessible, Interoperable, and Reusable Marine Metagenomic Datasets in Cyberinfrastructure Systems.” *Frontiers in Microbiology* 12 (December):765268.

Boev, Christine, Elizabeth Kiss. 2017. “Hospital-Acquired Infections: Current Trends and Prevention.” *Critical Care Nursing Clinics of North America* 29(1):51-65.

Bokarewa, M. I., T. Jin, and A. Tarkowski. 2006. “Staphylococcus Aureus: Staphylokinase.” *The International Journal of Biochemistry & Cell Biology* 38 (4): 504–9.

Braga, Lucas P. P., Aymé Spor, Witold Kot, Marie-Christine Breuil, Lars H. Hansen, João C. Setubal, and Laurent Philippot. 2020. “Impact of Phages on Soil Bacterial Communities and Nitrogen Availability under Different Assembly Scenarios.” *Microbiome* 8 (1): 52.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.

Breitbart, Mya, Ian Hewson, Ben Felts, Joseph M. Mahaffy, James Nulton, Peter Salamon, and Forest Rohwer. 2003. “Metagenomic Analyses of an Uncultured Viral Community from Human Feces.” *Journal of Bacteriology* 185 (20): 6220–23.

Breitbart, Mya, and Forest Rohwer. 2005a. “Here a Virus, There a Virus, Everywhere the Same Virus?” *Trends in Microbiology* 13 (6): 278–84.

———. 2005b. “Method for Discovering Novel DNA Viruses in Blood Using Viral Particle Selection and Shotgun Sequencing.” *BioTechniques* 39 (5): 729–36.

Brown-Jaque, Maryury, William Calero-Cáceres, Paula Espinal, Judith Rodríguez-Navarro, Elisenda Miró, Juan José González-López, Thais Cornejo, Juan Carlos Hurtado, Ferran Navarro, and Maite Muniesa. 2018. “Antibiotic Resistance Genes in Phage Particles Isolated from Human Faeces and Induced from Clinical Bacterial Isolates.” *International Journal of Antimicrobial Agents* 51 (3): 434–42.

Brown, J. R., T. Bharucha, and J. Breuer. 2018. “Encephalitis Diagnosis Using Metagenomics: Application of next Generation Sequencing for Undiagnosed Cases.” *The Journal of Infection*. https://www.sciencedirect.com/science/article/pii/S0163445317304139.

Brum, Jennifer R., Bonnie L. Hurwitz, Oscar Schofield, Hugh W. Ducklow, and Matthew B. Sullivan. 2016. “Seasonal Time Bombs: Dominant Temperate Viruses Affect Southern Ocean Microbial Dynamics.” *The ISME Journal* 10 (2): 437–49.

Brüssow, Harald. 2023. “The Human Microbiome Project at Ten Years - Some Critical Comments and Reflections on ‘Our Third Genome’, the Human Virome.” *Microbiome Research Reports* 2 (1): 7.

Burke, Jillian, Katelyn McNair, Adrian Cantu, Melissa Giluso, and Robert A. Edwards. 2019. “PARTIE-HAT: High Thoughput Automatic Tagging of Bacterial Genomes and Metagenomes.” *In Review*.

Buttigieg, Pier Luigi, Norman Morrison, Barry Smith, Christopher J. Mungall, Suzanna E. Lewis, and ENVO Consortium. 2013. “The Environment Ontology: Contextualising Biological and Biomedical Entities.” *Journal of Biomedical Semantics* 4 (1): 43.

Calle, Fernando de la. 2017. “Marine Microbiome as Source of Natural Products.” *Microbial Biotechnology* 10 (6): 1293–96.

Camarillo-Guerrero, Luis F., Alexandre Almeida, Guillermo Rangel-Pineros, Robert D. Finn, and Trevor D. Lawley. 2021. “Massive Expansion of Human Gut Bacteriophage Diversity.” *Cell* 184 (4): 1098–1109.e9.

Cao, Rong, Nikoleta Zeaki, Nina Wallin-Carlquist, Panagiotis N. Skandamis, Jenny Schelin, and Peter Rådström. 2012. “Elevated Enterotoxin A Expression and Formation in Staphylococcus Aureus and Its Association with Prophage Induction.” *Applied and Environmental Microbiology* 78 (14): 4942–48.

Castillo, Daniel, Kathryn Kauffman, Fatima Hussain, Panos Kalatzis, Nanna Rørbo, Martin F. Polz, and Mathias Middelboe. 2018. “Widespread Distribution of Prophage-Encoded Virulence Factors in Marine Vibrio Communities.” *Scientific Reports* 8 (1): 9973.

Cazares, Daniel, Adrian Cazares, Wendy Figueroa, Gabriel Guarneros, Robert A. Edwards, and Pablo Vinuesa. 2021. “A Novel Group of Promiscuous Podophages Infecting Diverse Gammaproteobacteria from River Communities Exhibits Dynamic Intergenus Host Adaptation.” *mSystems* 6 (1). https://doi.org/10.1128/mSystems.00773-20.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2011. “SMOTE: Synthetic Minority over-Sampling Technique.” *arXiv [cs.AI]*. arXiv. https://doi.org/10.1613/jair.953.

Chen, Chao. 2004. “Using Random Forest to Learn Imbalanced Data.” https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf.

Chen, Chen, Xiaolei Song, Weixia Wei, Huanzi Zhong, Juanjuan Dai, Zhou Lan, Fei Li, et al. 2017. “The Microbiota Continuum along the Female Reproductive Tract and Its Relation to Uterine-Related Diseases.” *Nature Communications* 8 (1): 875.

Cheng, H. H., P. J. Muhlrad, M. A. Hoyt, and H. Echols. 1988. “Cleavage of the cII Protein of Phage Lambda by Purified HflA Protease: Control of the Switch between Lysis and Lysogeny.” *Proceedings of the National Academy of Sciences of the United States of America* 85 (21): 7882–86.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785.

Chibani-Chennoufi, Sandra, Anne Bruttin, Marie-Lise Dillmann, and Harald Brüssow. 2004. “Phage-Host Interaction: An Ecological Perspective.” *Journal of Bacteriology* 186 (12): 3677–86.

Chiu, Charles Y., and Steven A. Miller. 2019. “Clinical Metagenomics.” *Nature Reviews. Genetics* 20 (6): 341–55.

Cho, Ilseung, and Martin J. Blaser. 2012. “The Human Microbiome: At the Interface of Health and Disease.” *Nature Reviews. Genetics* 13 (4): 260–70.

Christie, Gail E., and Terje Dokland. 2012. “Pirates of the Caudovirales.” *Virology* 434 (2): 210–21.

Clokie, Martha Rj, Andrew D. Millard, Andrey V. Letarov, and Shaun Heaphy. 2011. “Phages in Nature.” *Bacteriophage* 1 (1): 31–45.

Cohen, G. 1983. “Electron Microscopy Study of Early Lytic Replication Forms of Bacteriophage P1 DNA.” *Virology* 131 (1): 159–70.

Cohen, G., E. Or, W. Minas, and N. L. Sternberg. 1996. “The Bacteriophage P1 Lytic Replicon: Directionality of Replication and Cis-Acting Elements.” *Gene* 175 (1-2): 151–55.

Colomer-Lluch, Marta, Juan Jofre, and Maite Muniesa. 2011. “Antibiotic Resistance Genes in the Bacteriophage DNA Fraction of Environmental Samples.” *PloS One* 6 (3): e17549.

Crits-Christoph, A., R. S. Kantor, M. R. Olm, and O. N. Whitney. 2021. “Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants. MBio 12, e02703–e02720.”

Crump, Ron C., Heather E. Adams, John E. Hobbie, and George W. Kling. 2007. “Biogeography of Bacterioplankton in Lakes and Streams of an Arctic Tundra Catchment.” *Ecology* 88 (6): 1365–78.

Cuthbertson, Leah, Alan W. Walker, Anna E. Oliver, Geraint B. Rogers, Damian W. Rivett, Thomas H. Hampton, Alix Ashare, et al. 2020. “Lung Function and Microbiota Diversity in Cystic Fibrosis.” *Microbiome* 8 (1): 45.

Davis, James J., and Gary J. Olsen. 2010. “Modal Codon Usage: Assessing the Typical Codon Usage of a Genome.” *Molecular Biology and Evolution* 27 (4): 800–810.

deCarvalho, Tagide, Elia Mascolo, Steven M. Caruso, Júlia López-Pérez, Kathleen Weston-Hafer, Christopher Shaffer, and Ivan Erill. 2023. “Simultaneous Entry as an Adaptation to Virulence in a Novel Satellite-Helper System Infecting Streptomyces Species.” *The ISME Journal* 17 (12): 2381–88.

De Filippis, Francesca, Manolo Laiola, Giuseppe Blaiotta, and Danilo Ercolini. 2017. “Different Amplicon Targets for Sequencing-Based Studies of Fungal Diversity.” *Applied and Environmental Microbiology* 83 (17). https://doi.org/10.1128/AEM.00905-17.

Dehò, Gianni, Daniela Ghisotti, and Others. 2006. “The Satellite Phage P4.” *The Bacteriophages* 2:391.

DeLong, Edward F., Christina M. Preston, Tracy Mincer, Virginia Rich, Steven J. Hallam, Niels-Ulrik Frigaard, Asuncion Martinez, et al. 2006. “Community Genomics among Stratified Microbial Assemblages in the Ocean’s Interior.” *Science (New York, N.Y.)* 311 (5760): 496–503.

Dinsdale, Elizabeth A., Robert A. Edwards, Dana Hall, Florent Angly, Mya Breitbart, Jennifer M. Brulc, Mike Furlan, et al. 2008. “Functional Metagenomic Profiling of Nine Biomes.” *Nature* 452 (7187): 629–32.

Dutilh, Bas E., Noriko Cassman, Katelyn McNair, Savannah E. Sanchez, Genivaldo G. Z. Silva, Lance Boling, Jeremy J. Barr, et al. 2014. “A Highly Abundant Bacteriophage Discovered in the Unknown Sequences of Human Faecal Metagenomes.” *Nature Communications* 5 (1): 4498.

Edwards, Robert A., Beltran Rodriguez-Brito, Linda Wegley, Matthew Haynes, Mya Breitbart, Dean M. Peterson, Martin O. Saar, Scott Alexander, E. Calvin Alexander Jr, and Forest Rohwer. 2006. “Using Pyrosequencing to Shed Light on Deep Mine Microbial Ecology.” *BMC Genomics* 7 (1): 57.

Edwards, Robert A., Alejandro A. Vega, Holly M. Norman, Maria Ohaeri, Kyle Levi, Elizabeth A. Dinsdale, Ondrej Cinek, et al. 2019. “Global Phylogeography and Ancient Evolution of the Widespread Human Gut Virus crAssphage.” *Nature Microbiology* 4 (10): 1727–36.

Enault, François, Arnaud Briet, Léa Bouteille, Simon Roux, Matthew B. Sullivan, and Marie-Agnès Petit. 2017. “Phages Rarely Encode Antibiotic Resistance Genes: A Cautionary Tale for Virome Analyses.” *The ISME Journal* 11 (1): 237–47.

Erez, Zohar, Ida Steinberger-Levy, Maya Shamir, Shany Doron, Avigail Stokar-Avihail, Yoav Peleg, Sarah Melamed, et al. 2017. “Communication between Viruses Guides Lysis-Lysogeny Decisions.” *Nature* 541 (7638): 488–93.

Falony, Gwen, Marie Joossens, Sara Vieira-Silva, Jun Wang, Youssef Darzi, Karoline Faust, Alexander Kurilshikov, et al. 2016. “Population-Level Analysis of Gut Microbiome Variation.” *Science (New York, N.Y.)* 352 (6285): 560–64.

Feldgarden, Michael, Vyacheslav Brover, Narjol Gonzalez-Escalona, Jonathan G. Frye, Julie Haendiges, Daniel H. Haft, Maria Hoffmann, et al. 2021. “AMRFinderPlus and the Reference Gene Catalog Facilitate Examination of the Genomic Links among Antimicrobial Resistance, Stress Response, and Virulence.” *Scientific Reports* 11 (1): 12728.

Fernández, A., S. Huang, S. Seston, J. Xing, R. Hickey, C. Criddle, and J. Tiedje. 1999. “How Stable Is Stable? Function versus Community Composition.” *Applied and Environmental Microbiology* 65 (8): 3697–3704.

Flinders University. 2021. “Deep Thought (HPC).” Flinders University. https://doi.org/10.25957/FLINDERS.HPC.DEEPTHOUGHT.

Fox, Valeria, Francesco Santoro, Gianni Pozzi, and Francesco Iannelli. 2021. “Predicted Transmembrane Proteins with Homology to Mef(A) Are Not Responsible for Complementing mef(A) Deletion in the mef(A)-msr(D) Macrolide Efflux System in Streptococcus Pneumoniae.” *BMC Research Notes* 14 (1): 432.

Gaulke, Christopher A., and Thomas J. Sharpton. 2018. “The Influence of Ethnicity and Geography on Human Gut Microbiome Composition.” *Nature Medicine* 24 (10): 1495–96.

Gilbert, Jack A., Joshua A. Steele, J. Gregory Caporaso, Lars Steinbrück, Jens Reeder, Ben Temperton, Susan Huse, et al. 2012. “Defining Seasonal Marine Microbial Community Dynamics.” *The ISME Journal* 6 (2): 298–308.

Gillis, Annika, and Jacques Mahillon. 2014. “Phages Preying on Bacillus Anthracis, Bacillus Cereus, and Bacillus Thuringiensis: Past, Present and Future.” *Viruses* 6 (7): 2623–72.

Gómez, Pedro, Jonathan Bennie, Kevin J. Gaston, and Angus Buckling. 2015. “The Impact of Resource Availability on Bacterial Resistance to Phages in Soil.” *PloS One* 10 (4): e0123752.

Guarner, Francisco, and Juan-R Malagelada. 2003. “Gut Flora in Health and Disease.” *Lancet* 361 (9356): 512–19.

Gupta, Shashank, Martin S. Mortensen, Susanne Schjørring, Urvish Trivedi, Gisle Vestergaard, Jakob Stokholm, Hans Bisgaard, Karen A. Krogfelt, and Søren J. Sørensen. 2019. “Amplicon Sequencing Provides More Accurate Microbiome Information in Healthy Children Compared to Culturing.” *Communications Biology* 2 (1): 291.

Haggerty, John Matthew, and Elizabeth Ann Dinsdale. 2017. “Distinct Biogeographical Patterns of Marine Bacterial Taxonomy and Functional Genes.” *Global Ecology and Biogeography: A Journal of Macroecology* 26 (2): 177–90.

Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman. 1998. “Molecular Biological Access to the Chemistry of Unknown Soil Microbes: A New Frontier for Natural Products.” *Chemistry & Biology* 5 (10): R245–49.

Hay, Iain D., and Trevor Lithgow. 2019. “Filamentous Phages: Masters of a Microbial Sharing Economy.” *EMBO Reports* 20 (6): e47427.

Hematian, Ali, Nourkhoda Sadeghifard, Reza Mohebi, Morovat Taherikalani, Abbas Nasrolahi, Mansour Amraei, and Sobhan Ghafourian. 2016. “Traditional and Modern Cell Culture in Virus Diagnosis.” *Osong Public Health and Research Perspectives* 7 (2): 77–82.

Hendriksen, Rene S., Patrick Munk, Patrick Njage, Bram van Bunnik, Luke McNally, Oksana Lukjancenko, Timo Röder, et al. 2019. “Global Monitoring of Antimicrobial Resistance Based on Metagenomics Analyses of Urban Sewage.” *Nature Communications* 10 (1): 1124.

Hendrix, R. W., M. C. Smith, R. N. Burns, M. E. Ford, and G. F. Hatfull. 1999. “Evolutionary Relationships among Diverse Bacteriophages and Prophages: All the World’s a Phage.” *Proceedings of the National Academy of Sciences of the United States of America* 96 (5): 2192–97.

Hover, B. M., S. H. Kim, M. Katz, Z. Charlop-Powers, J. G. Owen, M. A. Ternei, and Brady S. 2018. “Culture-Independent Discovery of the Malacidins as Calcium-Dependent Antibiotics with Activity against Multidrug-Resistant Gram-Positive Pathogens.” *Nature Microbiology* 3 (4): 415–22.

Howard-Varona, Cristina, Katherine R. Hargreaves, Stephen T. Abedon, and Matthew B. Sullivan. 2017. “Lysogeny in Nature: Mechanisms, Impact and Ecology of Temperate Phages.” *The ISME Journal* 11 (7): 1511–20.

Hutchings, Matthew I., Andrew W. Truman, and Barrie Wilkinson. 2019. “Antibiotics: Past, Present and Future.” *Current Opinion in Microbiology* 51 (October):72–80.

Inglis, Laura K., Michael J. Roach, and Robert A. Edwards. 2024. “Prophages: An Integral but Understudied Component of the Human Microbiome.” *Microbial Genomics* 10 (1). https://doi.org/10.1099/mgen.0.001166.

Johnson, Timothy J., Yvonne M. Wannemuehler, and Lisa K. Nolan. 2008. “Evolution of the Iss Gene in Escherichia Coli.” *Applied and Environmental Microbiology* 74 (8): 2360–69.

Johnson, Jethro S., Daniel J. Spakowicz, Bo-Young Hong, Lauren M. Peterson, Patrick Demkowicz, Lei Chen, Shana R. Leopold, Blake M. Hanson, Hanako O. Agresta, Mark Gerstein, Erica Sodergren, and George M. Weinstock. 2019. “Evaluation of 16s gene sequencing for species and strain-level microbiome analysis.” *Nature Communications* 10:5029

Kim, Min-Soo, and Jin-Woo Bae. 2018. “Lysogeny Is Prevalent and Widely Distributed in the Murine Gut Microbiota.” *The ISME Journal* 12 (4): 1127–41.

Klein, Eili Y., Thomas P. Van Boeckel, Elena M. Martinez, Suraj Pant, Sumanth Gandra, Simon A. Levin, Herman Goossens, and Ramanan Laxminarayan. 2018. “Global Increase and Geographic Convergence in Antibiotic Consumption between 2000 and 2015.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (15): E3463–70.

Knowles, Ben, Barbara Bailey, Lance Boling, Mya Breitbart, Ana Cobián-Güemes, Javier Del Campo, Rob Edwards, et al. 2017. “Variability and Host Density Independence in Inductions-Based Estimates of Environmental Lysogeny.” *Nature Microbiology* 2 (7): 17064.

Knowles, B., C. B. Silveira, B. A. Bailey, K. Barott, V. A. Cantu, A. G. Cobián-Güemes, F. H. Coutinho, et al. 2016. “Lytic to Temperate Switching of Viral Communities.” *Nature* 531 (7595): 466–70.

Kogan, Michael I., Yulia L. Naboka, Khalid S. Ibishev, Irina A. Gudima, and Kurt G. Naber. 2015. “Human Urine Is Not Sterile - Shift of Paradigm.” *Urologia Internationalis* 94 (4): 445–52.

Kondo, Kohei, Mitsuoki Kawano, and Motoyuki Sugai. 2021. “Distribution of Antimicrobial Resistance and Virulence Genes within the Prophage-Associated Regions in Nosocomial Pathogens.” *mSphere* 6 (4): e0045221.

Kumpitsch, Christina, Kaisa Koskinen, Veronika Schöpf, and Christine Moissl-Eichinger. 2019. “The Microbiome of the Upper Respiratory Tract in Health and Disease.” *BMC Biology* 17 (1): 87.

Kushida, Tatsuya, Kouji Kozaki, Yuka Tateisi, Katsutaro Watanabe, T. Masuda, Katsuji Matsumura, Takahiro Kawamura, and T. Takagi. 2017. “Efficient Construction of a New Ontology for Life Sciences by Sub-Classifying Related Terms in the Japan Science, Technology Agency Thesaurus.” *International Conference on Biomedical Ontology*. https://ceur-ws.org/Vol-2137/paper_20.pdf.

Landgraff, Chrystal, Lu Ya Ruth Wang, Cody Buchanan, Matthew Wells, Justin Schonfeld, Kyrylo Bessonov, Jennifer Ali, Erin Robert, and Celine Nadon. 2021. “Metagenomic Sequencing of Municipal Wastewater Provides a near-Complete SARS-CoV-2 Genome Sequence Identified as the B.1.1.7 Variant of Concern from a Canadian Municipality Concurrent with an Outbreak.” *bioRxiv*. medRxiv. https://doi.org/10.1101/2021.03.11.21253409.

Lassalle, Florent, Matteo Spagnoletti, Matteo Fumagalli, Liam Shaw, Mark Dyble, Catherine Walker, Mark G. Thomas, Andrea Bamberg Migliano, and Francois Balloux. 2018. “Oral Microbiomes from Hunter-Gatherers and Traditional Farmers Reveal Shifts in Commensal Balance and Pathogen Load Linked to Diet.” *Molecular Ecology* 27 (1): 182–95.

Lederberg, E. M., and J. Lederberg. 1953. “Genetic Studies of Lysogenicity in Escherichia Coli.” *Genetics* 38 (1): 51–64.

Leinonen, R., H. Sugawara, and Martin Shumway. 2010. “The Sequence Read Archive.” *Nucleic Acids Research* 39 (suppl_1): D19–21.

Liddicoat, Craig, Robert A. Edwards, Michael Roach, Jake M. Robinson, Kiri Joy Wallace, Andrew D. Barnes, Joel Brame, et al. 2024. “Bioenergetic Mapping of ‘Healthy Microbiomes’ via Compound Processing Potential Imprinted in Gut and Soil Metagenomes.” *The Science of the Total Environment* 940 (173543): 173543.

Li, Manrong, Mamuka Kotetishvili, Yuansha Chen, and Shanmuga Sozhamannan. 2003. “Comparative Genomic Analyses of the Vibrio Pathogenicity Island and Cholera Toxin Prophage Regions in Nonepidemic Serogroup Strains of Vibrio Cholerae.” *Applied and Environmental Microbiology* 69 (3): 1728–38.

Lim, Yan Wei, Jose S. Evangelista 3rd, Robert Schmieder, Barbara Bailey, Matthew Haynes, Mike Furlan, Heather Maughan, Robert Edwards, Forest Rohwer, and Douglas Conrad. 2014. “Clinical Insights from Metagenomic Analysis of Sputum Samples from Patients with Cystic Fibrosis.” *Journal of Clinical Microbiology* 52 (2): 425–37.

Lloyd-Price, Jason, Galeb Abu-Ali, and Curtis Huttenhower. 2016. “The Healthy Human Microbiome.” *Genome Medicine* 8 (1): 51.

Lok, C. 2015. “Mining the Microbial Dark Matter.” *Nature*, June 16, 2015.

Lv, Luchao, Miao Wan, Chengzhen Wang, Xun Gao, Qiwen Yang, Sally R. Partridge, Yang Wang, et al. 2020. “Emergence of a Plasmid-Encoded Resistance-Nodulation-Division Efflux Pump Conferring Resistance to Multiple Drugs, Including Tigecycline, in Klebsiella Pneumoniae.” *mBio* 11 (2). https://doi.org/10.1128/mBio.02930-19.

Lwoff, André. 1953. “LYSOGENY1.” *Bacteriological Reviews* 17 (4): 269–337.

Madhusoodanan, Jyoti, Keun Seok Seo, Brian Remortel, Joo Youn Park, Sun Young Hwang, Lawrence K. Fox, Yong Ho Park, Claudia F. Deobald, Dan Wang, Song Liu, Sean C. Daugherty, Ann Lindley Gill, Gregory A. Bohach, Steven R Gill. 2011. “An Enterotoxin-Bearing Pathogenicity Island in *Staphylococcus epidermidis.*” *Journal of Bacteriology* 193(8):1854-1862

Mahoney, Aaron K., Chuntao Yin, and Scot H. Hulbert. 2017. “Community Structure, Species Variation, and Potential Functions of Rhizosphere-Associated Bacteria of Different Winter Wheat (Triticum Aestivum) Cultivars.” *Frontiers in Plant Science* 8 (February):132.

Man, Wing Ho, Wouter A. A. de Steenhuijsen Piters, and Debby Bogaert. 2017. “The Microbiota of the Respiratory Tract: Gatekeeper to Respiratory Health.” *Nature Reviews. Microbiology* 15 (5): 259–70.

Marchesini, Giulio, Francesca Marchignoli, and Salvatore Petta. 2017. “Evidence-Based Medicine and the Problem of Healthy Volunteers.” *Annals of Hepatology* 16 (6): 832–34.

Martín, Virginia, Antonio Maldonado-Barragán, Laura Moles, Mercedes Rodriguez-Baños, Rosa Del Campo, Leonides Fernández, Juan M. Rodríguez, and Esther Jiménez. 2012. “Sharing of Bacterial Strains between Breast Milk and Infant Feces.” *Journal of Human Lactation: Official Journal of International Lactation Consultant Association* 28 (1): 36–44.

Mavrich, Travis N., and Graham F. Hatfull. 2019. “Evolution of Superinfection Immunity in Cluster A Mycobacteriophages.” *mBio* 10 (3). https://doi.org/10.1128/mBio.00971-19.

McDaniel, Lauren, Mya Breitbart, Jennifer Mobberley, Amy Long, Matthew Haynes, Forest Rohwer, and John H. Paul. 2008. “Metagenomic Analysis of Lysogeny in Tampa Bay: Implications for Prophage Gene Expression.” *PloS One* 3 (9): e3263.

McDaniel, Lauren D., Karyna Rosario, Mya Breitbart, and John H. Paul. 2014. “Comparative Metagenomics: Natural Populations of Induced Prophages Demonstrate Highly Unique, Lower Diversity Viral Sequences: Metagenomic Comparison of Induced and Ambient Viruses.” *Environmental Microbiology* 16 (2): 570–85.

McFadden, D. 1972. “Conditional Logit Analysis of Qualitative Choice Behavior.” *Frontiers in Econometrics*. https://escholarship.org/content/qt61s3q2xr/qt61s3q2xr.pdf.

McKerral, Jody C., Bhavya Papudeshi, Laura K. Inglis, Michael J. Roach, Przemyslaw Decewicz, Katelyn McNair, Antoni Luque, Elizabeth A. Dinsdale, and Robert A. Edwards. 2023. “The Promise and Pitfalls of Prophages.” *bioRxiv : The Preprint Server for Biology*, April. https://doi.org/10.1101/2023.04.20.537752.

McNair, K., P. Decewicz, S. Akhter, R. K. Aziz, and S. Daniel. 2019. “PhiSpy.”

Mirsepasi-Lauridsen, Hengameh Chloé, Katleen Vrankx, Jørgen Engberg, Alice Friis-Møller, Jørn Brynskov, Inge Nordgaard-Lassen, Andreas Munk Petersen, and Karen Angeliki Krogfelt. 2018. “Disease-Specific Enteric Microbiome Dysbiosis in Inflammatory Bowel Disease.” *Frontiers in Medicine* 5 (November):304.

Mitchell, Alex L., Alexandre Almeida, Martin Beracochea, Miguel Boland, Josephine Burgin, Guy Cochrane, Michael R. Crusoe, et al. 2020. “MGnify: The Microbiome Analysis Resource in 2020.” *Nucleic Acids Research* 48 (D1): D570–78.

Modi, Sheetal R., Henry H. Lee, Catherine S. Spina, and James J. Collins. 2013. “Antibiotic Treatment Expands the Resistance Reservoir and Ecological Network of the Phage Metagenome.” *Nature* 499 (7457): 219–22.

Moon, Kira, Jeong Ho Jeon, Ilnam Kang, Kwang Seung Park, Kihyun Lee, Chang-Jun Cha, Sang Hee Lee, and Jang-Cheon Cho. 2020. “Freshwater Viral Metagenome Reveals Novel and Functional Phage-Borne Antibiotic Resistance Genes.” *Microbiome* 8 (1): 75.

Morris, Peter, Laura J. Marinelli, Deborah Jacobs-Sera, Roger W. Hendrix, and Graham F. Hatfull. 2008. “Genomic Characterization of Mycobacteriophage Giles: Evidence for Phage Acquisition of Host DNA by Illegitimate Recombination.” *Journal of Bacteriology* 190 (6):

2172–82.

Motlagh, Amir Mohaghegh, Ananda S. Bhattacharjee, Felipe H. Coutinho, Bas E. Dutilh, Sherwood R. Casjens, and Ramesh K. Goel. 2017. “Insights of Phage-Host Interaction in Hypersaline Ecosystem through Metagenomics Analyses.” *Frontiers in Microbiology* 8 (March):352.

Musolf, Anthony M., Emily R. Holzinger, James D. Malley, and Joan E. Bailey-Wilson. 2022. “What Makes a Good Prediction? Feature Importance and Beginning to Open the Black Box of Machine Learning in Genetics.” *Human Genetics* 141 (9): 1515–28.

Muthuirulandi Sethuvel, Dhiviya Prabaa, Nithya Subramanian, Agila Kumari Pragasam, Francis Yesurajan Inbanathan, Prashant Gupta, Jaichand Johnson, Naresh Chand Sharma, et al. 2019. “Insights to the Diphtheria Toxin Encoding Prophages amongst Clinical Isolates of Corynebacterium Diphtheriae from India.” *Indian Journal of Medical Microbiology* 37 (3): 423–25.

Nardone, Gerardo, and Debora Compare. 2015. “The Human Gastric Microbiota: Is It Time to Rethink the Pathogenesis of Stomach Diseases?” *United European Gastroenterology Journal* 3 (3): 255–60.

Nicodemus, K. K. 2011. “Letter to the Editor: On the Stability and Ranking of Predictors from Random Forest Variable Importance Measures.” *Briefings in Bioinformatics* 12 (4). https://doi.org/10.1093/bib/bbr016.

Noto, Jennifer M., and Richard M. Peek Jr. 2017. “The Gastric Microbiome, Its Interaction with Helicobacter Pylori, and Its Potential Role in the Progression to Stomach Cancer.” *PLoS Pathogens* 13 (10): e1006573.

O’Dwyer, David N., Robert P. Dickson, and Bethany B. Moore. 2016. “The Lung Microbiome, Immunity, and the Pathogenesis of Chronic Lung Disease.” *The Journal of Immunology* 196 (12): 4839–47.

Oh, Jee-Hwan, Xiaoxi B. Lin, Shenwei Zhang, Stephanie L. Tollenaar, Mustafa Özçam, Case Dunphy, Jens Walter, and Jan-Peter van Pijkeren. 2019. “Prophages in Lactobacillus Reuteri Are Associated with Fitness Trade-Offs but Can Increase Competitiveness in the Gut Ecosystem.” *Applied and Environmental Microbiology* 86 (1). https://doi.org/10.1128/AEM.01922-19.

Païssé, Sandrine, Carine Valle, Florence Servant, Michael Courtney, Rémy Burcelin, Jacques Amar, and Benjamin Lelouvier. 2016. “Comprehensive Description of Blood Microbiome from Healthy Donors Assessed by 16S Targeted Metagenomic Sequencing.” *Transfusion* 56 (5): 1138–47.

Papudeshi, Bhavya, J. Matthew Haggerty, Michael Doane, Megan M. Morris, Kevin Walsh, Douglas T. Beattie, Dnyanada Pande, et al. 2017. “Optimizing and Evaluating the Reconstruction of Metagenome-Assembled Microbial Genomes.” *BMC Genomics* 18 (1): 915.

Parmar, Krupa, Nishant Dafale, Rajesh Pal, Hitesh Tikariha, and Hemant Purohit. 2018. “An Insight into Phage Diversity at Environmental Habitats Using Comparative Metagenomics Approach.” *Current Microbiology* 75 (2): 132–41.

Paterson, James S., Renee J. Smith, Jody C. McKerral, Lisa M. Dann, Elise Launer, Peter Goonan, Tavis Kleinig, Jed A. Fuhrman, and James G. Mitchell. 2019. “A Hydrocarbon-Contaminated Aquifer Reveals a Piggyback-the-Persistent Viral Strategy.” *FEMS*

Microbiology Ecology 95 (8): fiz116.

Pavletic, Adriana J. 2020. “Why Knowing Healthy Controls Matters.” *International Journal of Clinical Practice* 74 (1): e13424.

Pfeifer, E., and Eduardo P. C. Rocha. 2024. “Phage-Plasmids Promote Recombination and Emergence of Phages and Plasmids.” *Nature Communications* 15 (February). https://doi.org/10.1038/s41467-024-45757-3.

Pfeifer, Eugen, Rémy A. Bonnin, and Eduardo P. C. Rocha. 2022. “Phage-Plasmids Spread Antibiotic Resistance Genes through Infection and Lysogenic Conversion.” *mBio* 13 (5): e0185122.

Pfeifer, Eugen, Jorge A. Moura de Sousa, Marie Touchon, and Eduardo P. C. Rocha. 2021. “Bacteria Have Numerous Distinctive Groups of Phage-Plasmids with Conserved Phage and Variable Plasmid Gene Repertoires.” *Nucleic Acids Research* 49 (5): 2655–73.

Poupei, Olivier, Gérald Kenanian, Lhousseine Tounqui, Charlotte Abrial, Tarek Msadek, Sarah Dubrac. 2025. “Timely excision of prophage Φ13 is essential for the *Staphylococcus aureus* infectious process.” *Infection and Immunity* 0:e00314-25

Prestat, E., M. M. David, J. Hultman, N. Taş, R. Lamendella, J. Dvornik, and Jansson J. 2014. “FOAM (functional Ontology Assignments for Metagenomes): A Hidden Markov Model (HMM) Database with Environmental Focus.” *Nucleic Acids Research* 42 (19): e145–e145.

Ptashne, M. 2004. “A Genetic Switch, Phage Lambda Revisited.” *(No Title)*, April. https://cir.nii.ac.jp/crid/1370285712577206913.

Rezaei Javan, Reza, Elisa Ramos-Sevillano, Asma Akter, Jeremy Brown, and Angela B. Brueggemann. 2019. “Prophages and Satellite Prophages Are Widespread in Streptococcus and May Play a Role in Pneumococcal Pathogenesis.” *Nature Communications* 10 (1): 4852.

Richardson, Lorna, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L. Bileschi, Tony Burdett, Josephine Burgin, et al. 2023. “MGnify: The Microbiome Sequence Data Analysis Resource in 2023.” *Nucleic Acids Research* 51 (D1): D753–59.

Riedl, Ruth A., Colin M. L. Burnett, Nicole A. Pearson, John J. Reho, Mohamad Mokadem, Robert A. Edwards, Tammy L. Kindel, John R. Kirby, and Justin L. Grobe. 2021. “Gut Microbiota Represent a Major Thermogenic Biomass.” *Function (Oxford, England)* 2 (3): zqab019.

Roach, Michael J., Katelyn McNair, Sarah K. Giles, Laura Inglis, Evan Pargin, Simon Roux, Przemysław Decewicz, and Robert A. Edwards. 2021. “Philympics 2021: Prophage Predictions Perplex Programs.” *bioRxiv*. bioRxiv. https://doi.org/10.1101/2021.06.03.446868.

Roach, Michael J., Katelyn McNair, Maciej Michalczyk, Sarah K. Giles, Laura K. Inglis, Evan Pargin, Jakub Barylski, Simon Roux, Przemysław Decewicz, and Robert A. Edwards. 2022. “Philympics 2021: Prophage Predictions Perplex Programs.” *F1000Research* 10 (758): 758.

Roberts, M. C., J. Sutcliffe, P. Courvalin, L. B. Jensen, J. Rood, and H. Seppala. 1999. “Nomenclature for Macrolide and Macrolide-Lincosamide-Streptogramin B Resistance Determinants.” *Antimicrobial Agents and Chemotherapy* 43 (12): 2823–30.

Rodrigues Souza, Stephanie Silva, Adriana Rocha Faria, Andréa Andrade Rangel Freitas, Paul J. Planet, Vânia Lúcia Carreira Merquior, and Lúcia Martins Teixeira. 2020. “Occurrence and Associated Characteristics of a mutated*ant(6’)-Ia*gene among*Enterococcus*

*Faecium*strains Expressing Phenotypic Susceptibility to High Levels of Streptomycin." *bioRxiv*. bioRxiv. https://doi.org/10.1101/2020.12.28.424548.

Rodriguez-Brito, Beltran, Linlin Li, Linda Wegley, Mike Furlan, Florent Angly, Mya Breitbart, John Buchanan, et al. 2010. "Viral and Microbial Community Dynamics in Four Aquatic Environments." *The ISME Journal* 4 (6): 739–51.

Rodríguez-Rubio, Lorena, Nadja Haarmann, Maike Schwidder, Maite Muniesa, and Herbert Schmidt. 2021. "Bacteriophages of Shiga Toxin-Producing Escherichia Coli and Their Contribution to Pathogenicity." *Pathogens* 10 (4). https://doi.org/10.3390/pathogens10040404.

Rohwer, Forest, and Rob Edwards. 2002. "The Phage Proteomic Tree: A Genome-Based Taxonomy for Phage." *Journal of Bacteriology* 184 (16): 4529–35.

Romero-Calle, Danitza, Raquel Guimarães Benevides, Aristóteles Góes-Neto, and Craig Billington. 2019. "Bacteriophages as Alternatives to Antibiotics in Clinical Care." *Antibiotics (Basel, Switzerland)* 8 (3): 138.

Rothman, Jason A., Theresa B. Loveless, Madison L. Griffith, Joshua A. Steele, John F. Griffith, and Katrine L. Whiteson. 2020. "Metagenomics of Wastewater Influent from Southern California Wastewater Treatment Facilities in the Era of COVID-19." *Microbiology Resource Announcements* 9 (41). https://doi.org/10.1128/mra.00907-20.

Saraf, Viqar Sayeed, Sohail Aslam Sheikh, Aftab Ahmad, Patrick M. Gillevet, Habib Bokhari, and Sundus Javed. 2021. "Vaginal Microbiome: Normalcy vs Dysbiosis." *Archives of Microbiology* 203 (7): 3793–3802.

Schmieder, Robert, and Robert Edwards. 2012. "Insights into Antibiotic Resistance through Metagenomic Approaches." *Future Microbiology* 7 (1): 73–89.

Schuch, Raymond, and Vincent A. Fischetti. 2006. "Detailed Genomic Analysis of the Wbeta and Gamma Phages Infecting Bacillus Anthracis: Implications for Evolution of Environmental Fitness and Antibiotic Resistance." *Journal of Bacteriology* 188 (8): 3037–51.

Sender, Ron, Shai Fuchs, and Ron Milo. 2016. "Revised Estimates for the Number of Human and Bacteria Cells in the Body." *PLoS Biology* 14 (8): e1002533.

Sheng, Xiusheng, Wei Lu, Aifang Li, Junwan Lu, Chunhan Song, Jiefeng Xu, Youming Dong, et al. 2023. "ANT(9)-Ic, a Novel Chromosomally Encoded Aminoglycoside Nucleotidyltransferase from Brucella Intermedia." *Microbiology Spectrum* 11 (3): e0062023.

Shkoporov, Andrey N., Adam G. Clooney, Thomas D. S. Sutton, Feargal J. Ryan, Karen M. Daly, James A. Nolan, Siobhan A. McDonnell, et al. 2019. "The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific." *Cell Host & Microbe* 26 (4): 527–41.e5.

Shkoporov, A. N., and C. Hill. 2019. "Bacteriophages of the Human Gut: The 'known Unknown' of the Microbiome." *Cell Host & Microbe* 25 (2): 195–209.

Silva, Genivaldo Gueiros Z., Daniel A. Cuevas, Bas E. Dutilh, and Robert A. Edwards. 2014. "FOCUS: An Alignment-Free Model to Identify Organisms in Metagenomes Using Non-Negative Least Squares." *PeerJ* 2 (June):e425.

Silva, Genivaldo Gueiros Z., Kevin T. Green, Bas E. Dutilh, and Robert A. Edwards. 2016. "SUPER-FOCUS: A Tool for Agile Functional Analysis of Shotgun Metagenomic Data." *Bioinformatics* 32 (3): 354–61.

Silveira, Cynthia B., Antoni Luque, and Forest Rohwer. 2021. "The Landscape of Lysogeny across Microbial Community Density, Diversity and Energetics." *Environmental Microbiology* 23 (8): 4098–4111.

Silveira, Cynthia B., and Forest L. Rohwer. 2016. "Piggyback-the-Winner in Host-Associated Microbial Communities." *Npj Biofilms and Microbiomes* 2 (1): 16010.

Sperandei, Sandro. 2014. "understanding logistic regression analysis." *Biochemia Medica* 15;24 (1):12-18.

Staley, J. T., and A. Konopka. 1985. "Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats." *Annual Review of Microbiology* 39 (1): 321–46.

Steenhuijsen Piters, Wouter A. A. de, Elisabeth G. W. Huijskens, Anne L. Wyllie, Giske Biesbroek, Menno R. van den Bergh, Reinier H. Veenhoven, Xinhui Wang, et al. 2016. "Dysbiosis of Upper Respiratory Tract Microbiota in Elderly Pneumonia Patients." *The ISME Journal* 10 (1): 97–108.

Stokar-Avihail, Avigail, Nitzan Tal, Zohar Erez, Anna Lopatina, and Rotem Sorek. 2019. "Widespread Utilization of Peptide Communication in Phages Infecting Soil and Pathogenic Bacteria." *Cell Host & Microbe* 25 (5): 746–55.e5.

Suttle, Curtis A. 2007. "Marine Viruses--Major Players in the Global Ecosystem." *Nature Reviews. Microbiology* 5 (10): 801–12.

Suzuki, Taichi A., and Michael Worobey. 2014. "Geographical Variation of Human Gut Microbial Composition." *Biology Letters* 10 (2): 20131037.

Sweere, Johanna M., Jonas D. Van Belleghem, Heather Ishak, Michelle S. Bach, Medeea Popescu, Vivekananda Sunkari, Gernot Kaber, et al. 2019. "Bacteriophage Trigger Antiviral Immunity and Prevent Clearance of Bacterial Infection." *Science (New York, N.Y.)* 363 (6434): eaat9691.

Tagliaferri, Thaysa Leite, Mathias Jansen, and Hans-Peter Horz. 2019. "Fighting Pathogenic Bacteria on Two Fronts: Phages and Antibiotics as Combined Strategy." *Frontiers in Cellular and Infection Microbiology* 9 (February):22.

Thingstad, T. Frede. 2000. "Elements of a Theory for the Mechanisms Controlling Abundance, Diversity, and Biogeochemical Role of Lytic Bacterial Viruses in Aquatic Systems." *Limnology and Oceanography* 45 (6): 1320–28.

Thompson, Cristiane C., Livia Vidal, Vinicius Salazar, Jean Swings, and Fabiano L. Thompson. 2021. "Microbial Genomic Taxonomy." In *Trends in the Systematics of Bacteria and Fungi*, 168–78. UK: CABI.

Torres-Barceló, Clara. 2018. "The Disparate Effects of Bacteriophages on Antibiotic-Resistant Bacteria." *Emerging Microbes & Infections* 7 (1): 168.

Torres, Pedro J., Robert A. Edwards, and Katelyn A. McNair. 2017. "PARTIE: A Partition Engine to Separate Metagenomic and Amplicon Projects in the Sequence Read Archive." *Bioinformatics (Oxford, England)* 33 (15): 2389–91.

Tortorella, Emiliana, Pietro Tedesco, Fortunato Palma Esposito, Grant Garren January, Renato Fani, Marcel Jaspars, and Donatella de Pascale. 2018. "Antibiotics from Deep-Sea Microorganisms: Current Discoveries and Perspectives." *Marine Drugs* 16 (10): 355.

Touchon, Marie, Jorge A. Moura de Sousa, and Eduardo Pc Rocha. 2017. "Embracing the Enemy: The Diversification of Microbial Gene Repertoires by Phage-Mediated Horizontal Gene Transfer." *Current Opinion in Microbiology* 38 (August):66–73.

Turnbaugh, Peter J., Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, et al. 2009. “A Core Gut Microbiome in Obese and Lean Twins.” *Nature* 457 (7228): 480–84.

Turnbaugh, Peter J., Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon. 2007. “The Human Microbiome Project.” *Nature* 449 (7164): 804–10.

Tyson, Gene W., Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar, and Jillian F. Banfield. 2004. “Community Structure and Metabolism through Reconstruction of Microbial Genomes from the Environment.” *Nature* 428 (6978): 37–43.

University, Flinders. 2021. “Deep Thought (HPC).”

Vandenkoornhuyse, Philippe, Achim Quaiser, Marie Duhamel, Amandine Le Van, and Alexis Dufresne. 2015. “The Importance of the Microbiome of the Plant Holobiont.” *The New Phytologist* 206 (4): 1196–1206.

Vries, Jutte J. C. de, Julianne R. Brown, Natacha Couto, Martin Beer, Philippe Le Mercier, Igor Sidorov, Anna Papa, et al. 2021. “Recommendations for the Introduction of Metagenomic next-Generation Sequencing in Clinical Virology, Part II: Bioinformatic Analysis and Reporting.” *Journal of Clinical Virology: The Official Publication of the Pan American Society for Clinical Virology* 138 (104812): 104812.

Vuong, Cuong, and Michael Otto. 2002. “Staphylococcus Epidermidis Infections.” *Microbes and Infection* 4 (4): 481–89.

Waldor, M. K., and J. J. Mekalanos. 1996. “Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin.” *Science (New York, N.Y.)* 272 (5270): 1910–14.

Walls, Ramona L., John Deck, Robert Guralnick, Steve Baskauf, Reed Beaman, Stanley Blum, Shawn Bowers, et al. 2014. “Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies.” *PloS One* 9 (3): e89606.

Wattam, Alice R., James J. Davis, Rida Assaf, Sébastien Boisvert, Thomas Brettin, Christopher Bun, Neal Conrad, et al. 2017. “Improvements to PATRIC, the All-Bacterial Bioinformatics Database and Analysis Resource Center.” *Nucleic Acids Research* 45 (D1): D535–42.

Wilson, Michael R., Samia N. Naccache, Erik Samayoa, Mark Biagtan, Hiba Bashir, Guixia Yu, Shahriar M. Salamat, et al. 2014. “Actionable Diagnosis of Neuroleptospirosis by next-Generation Sequencing.” *The New England Journal of Medicine* 370 (25): 2408–17.

Winter, Amy K., and Sonia T. Hegde. 2020. “The Important Role of Serology for COVID-19 Control.” *The Lancet Infectious Diseases* 20 (7): 758–59.

Wintersdorff, Christian J. H. von, John Penders, Julius M. van Niekerk, Nathan D. Mills, Snehali Majumder, Lieke B. van Alphen, Paul H. M. Savelkoul, and Petra F. G. Wolffs. 2016. “Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer.” *Frontiers in Microbiology* 7 (February):173.

Wisdom, R., J. Yen, D. Rashid, and I. M. Verma. 1992. “Transformation by FosB Requires a Trans-Activation Domain Missing in FosB2 That Can Be Substituted by Heterologous Activation Domains.” *Genes & Development* 6 (4): 667–75.

Xia, Guoqing, and Christiane Wolz. 2014. “Phages of Staphylococcus Aureus and Their Impact on Host Evolution.” *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 21 (January):593–601.

Yagi, Kazuma, Gary B. Huffnagle, Nicholas W. Lukacs, and Nobuhiro Asai. 2021. “The Lung Microbiome during Health and Disease.” *International Journal of Molecular Sciences* 22 (19): 10872.

Yang, Wangrong, Ian F. Moore, Kalinka P. Koteva, David C. Bareich, Donald W. Hughes, and Gerard D. Wright. 2004. “TetX Is a Flavin-Dependent Monooxygenase Conferring Resistance to Tetracycline Antibiotics.” *The Journal of Biological Chemistry* 279 (50): 52346–52.

Yatera, Kazuhiro, Shingo Noguchi, and Hiroshi Mukae. 2018. “The Microbiome in the Lower Respiratory Tract.” *Respiratory Investigation* 56 (6): 432–39.

Yatsunenko, Tanya, Federico E. Rey, Mark J. Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, et al. 2012. “Human Gut Microbiome Viewed across Age and Geography.” *Nature* 486 (7402): 222–27.

Yi, Hana, Dongeun Yong, Kyungwon Lee, Yong-Joon Cho, and Jongsik Chun. 2014. “Profiling Bacterial Community in Upper Respiratory Tracts.” *BMC Infectious Diseases* 14 (1): 583.

Youngblut, Nicholas D., Georg H. Reischer, William Walters, Nathalie Schuster, Chris Walzer, Gabrielle Stalder, Ruth E. Ley, and Andreas H. Farnleitner. 2019. “Host Diet and Evolutionary History Explain Different Aspects of Gut Microbiome Diversity among Vertebrate Clades.” *Nature Communications* 10 (1): 2200.

Zaura, Egija, Bart J. F. Keijser, Susan M. Huse, and Wim Crielaard. 2009. “Defining the Healthy ‘Core Microbiome’ of Oral Microbial Communities.” *BMC Microbiology* 9 (December):259.

Zhu, Jie, Jingnan Lv, Zhichen Zhu, Tao Wang, Xiaofang Xie, Haifang Zhang, Liang Chen, and Hong Du. 2023. “Identification of TMexCD-TOprJ-Producing Carbapenem-Resistant Gram-Negative Bacteria from Hospital Sewage.” *Drug Resistance Updates: Reviews and Commentaries in Antimicrobial and Anticancer Chemotherapy* 70 (100989): 100989.

Zinder, N. D. 1958. “Lysogenization and Superinfection Immunity in Salmonella.” *Virology* 5 (2): 291–326.

Zong, Gongli, Chuanqing Zhong, Jiafang Fu, Yu Zhang, Peipei Zhang, Wenchi Zhang, Yan Xu, Guangxiang Cao, and Rongzhen Zhang. 2020. “The Carbapenem Resistance Gene blaOXA-23 Is Disseminated by a Conjugative Plasmid Containing the Novel Transposon Tn6681 in Acinetobacter Johnsonii M19.” *Antimicrobial Resistance and Infection Control* 9 (1): 182.

Zubyk, Haley L., and Gerard D. Wright. 2021. “CrpP Is Not a Fluoroquinolone-Inactivating Enzyme.” *Antimicrobial Agents and Chemotherapy* 65 (8): e0077321.

COVID-19 Sewage Surveillance Program - COVID-19 (Coronavirus). 2022. COVID-19 Sewage Surveillance Program - COVID-19 (Coronavirus). [ONLINE] Available at: https://www.health.nsw.gov.au/Infectious/covid-19/Pages/sewage-surveillance.aspx.

Inglis L, Edwards R. Prophages in Humans. 2023. Flinders University.

Public health alert - New venues of concern - News. 2022b. Public health alert - New venues of concern - News. [ONLINE] Available at: https://www.health.nsw.gov.au/news/Pages/20210813_03.aspx..

Welcome to the NCBO BioPortal | NCBO BioPortal. 2022. Welcome to the NCBO BioPortal | NCBO BioPortal. [ONLINE] Available at: https://bioportal.bioontology.org.

www.ncbi.nlm.nih.gov. 2022. No page title. [ONLINE] Available at: https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/

Appendices

Supplementary Data 1: Dataset overview

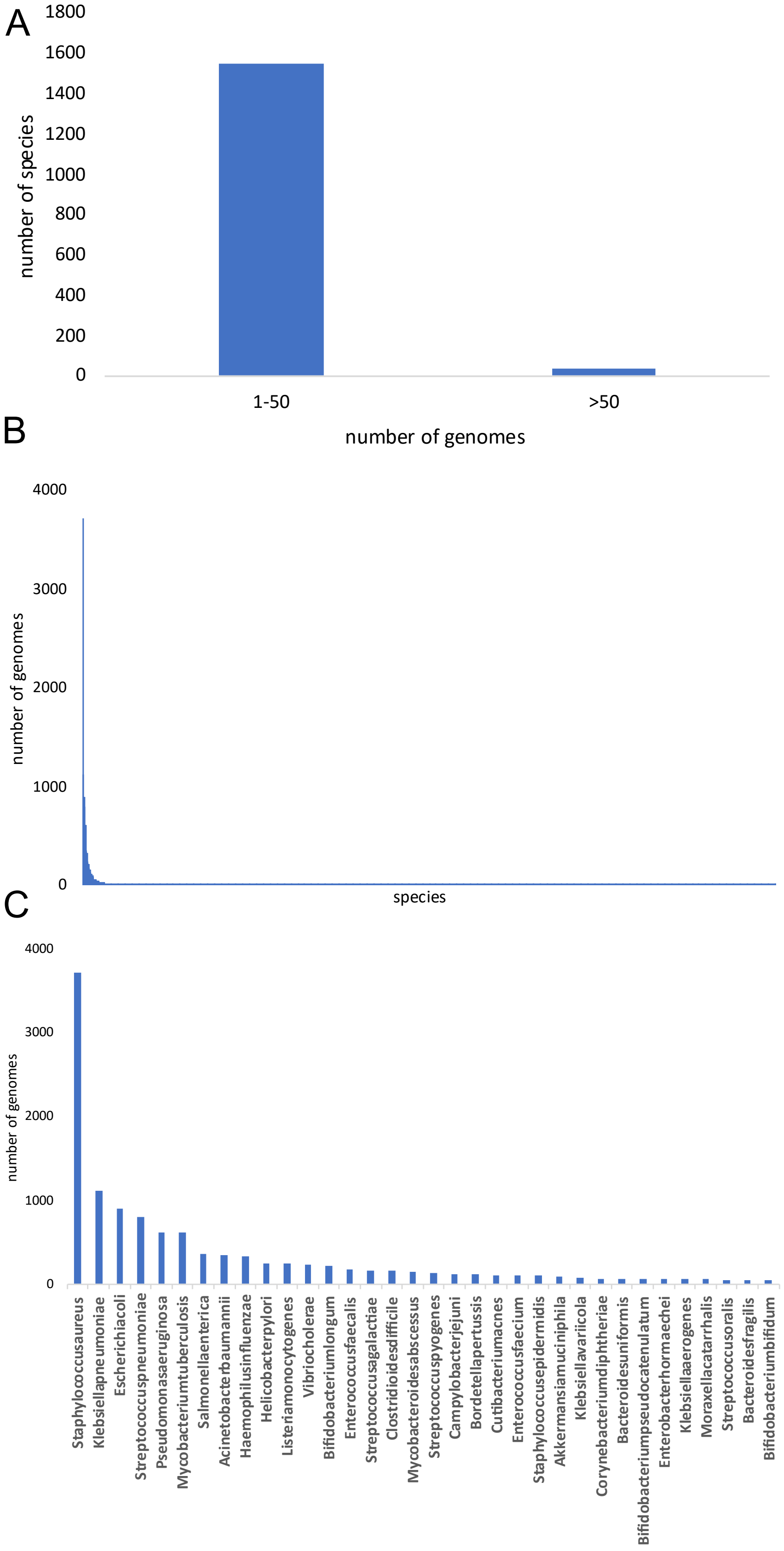
A
1800
1600
1400
1200
1000
800
600
400
200
0
number of species
1-50
>50
number of genomes
B
4000
3000
2000
1000
0
number of genomes
species
C
4000
3000
2000
1000
0
number of genomes
Staphylococcusaureus
Klebsiellapneumoniae
Escherichiacoli
Streptococcuspneumoniae
Pseudomonasaeruginosa
Mycobacteriumtuberculosis
Salmonellaenterica
Acinetobacterbaumannii
Haemophilusinfluenzae
Helicobacterpylori
Listeriamonocytogenes
Vibriocholerae
Bifidobacteriumlongum
Enterococcusfaecalis
Streptococcusagalactiae
Clostridioidesdifficile
Mycobacteroidesabscessus
Streptococcuspyogenes
Campylobacterjejuni
Bordetellapertussis
Cutibacteriumacnes
Enterococcusfaecium
Staphylococcusepidermidis
Akkermansiamuciniphila
Klebsiellavariicola
Corynebacteriumdiphtheriae
Bacteroidesuniformis
Bifidobacteriumpseudocatenulatum
Enterobacterhormaechei
Klebsiellaaerogenes
Moraxellacatarrhalis
Streptococcusoralis
Bacteroidesfragilis
Bifidobacteriumbifidum

Figure S1: the majority of species in the datasaet had less than 50 genomes representing the species. A shows the number of species with 1-50 and more than 50 genomes, while B shows that it is a few species that make up the majority of the genomes. C shows the names and number of genomes for each species with more than 50 genomes.

Table S1: A summary the number of genome overlap between the region, body site, and host health categories.

region	body site	host_health	number of genomes
Africa	stool	unknown	26
Africa	stool	sick	27
Africa	urine	unknown	1
Africa	vagina	unknown	1
Africa	vagina	healthy	3
Africa	blood	sick	2
Africa	blood	unknown	5
Asia	stool	unknown	201
Asia	blood	unknown	20
Asia	stool	sick	35
Asia	liver	unknown	1
Asia	blood	sick	44
Asia	skin	unknown	1
Asia	throat	asymptomatic	1
Asia	urine	sick	16
Asia	throat	sick	6
Asia	sputum	sick	3
Asia	lung	sick	2
Asia	skin	sick	3
Asia	CSF	sick	1
Asia	miscgut	unknown	3
Asia	vagina	asymptomatic	1
Asia	stool	unkown	1
Asia	lung	unknown	1
Asia	vagina	unknown	1
Asia	vagina	sick	1
Asia	nose	sick	15
Asia	miscgut	sick	2
Asia	plaque	sick	1
Asia	bronchial	sick	1
Asia	nose	unknown	1
Asia	rectal	unknown	1

Asia	sputum	unknown	2
Asia	urine	unknown	17
CentralAmerica	stool	unknown	5
CentralAmerica	stool	sick	1
CentralAmerica	blood	sick	2
CentralAmerica	urine	sick	1
Europe	skin	unknown	2
Europe	blood	unknown	14
Europe	stool	unknown	80
Europe	bronchial	unknown	1
Europe	blood	sick	280
Europe	urine	sick	87
Europe	throat	unknown	7
Europe	sputum	sick	24
Europe	stool	sick	11
Europe	nose	healthy	1
Europe	nose	unknown	6
Europe	vagina	sick	13
Europe	skin	sick	2
Europe	urine	unknown	16
Europe	nose	sick	10
Europe	rectal	sick	1
Europe	rectal	unknown	6
Europe	UT	unknown	1
Europe	UT	sick	2
Europe	bone	sick	3
Europe	mouth	unknown	1
Europe	bronchial	sick	10
Europe	throat	sick	11
Europe	rectal	asymptomatic	1
Europe	stool	asymptomatic	10
Europe	stool	healthy	1
Europe	bile	unkown	1
Europe	vagina	asymptomatic	1
NorthAmerica	blood	sick	57
NorthAmerica	nose	healthy	1
NorthAmerica	urine	unknown	76
NorthAmerica	blood	unknown	139
NorthAmerica	nose	sick	2
NorthAmerica	throat	healthy	1
NorthAmerica	lung	unknown	12

NorthAmerica	stool	unknown	29
NorthAmerica	nose	unknown	368
NorthAmerica	sputum	unknown	159
NorthAmerica	stool	sick	23
NorthAmerica	bile	unknown	1
NorthAmerica	bronchial	unknown	19
NorthAmerica	mouth	unknown	1
NorthAmerica	skin	unknown	16
NorthAmerica	throat	unknown	10
NorthAmerica	UT	unknown	1
NorthAmerica	eye	unknown	1
NorthAmerica	rectal	unknown	13
NorthAmerica	CSF	unknown	1
NorthAmerica	bone	sick	2
NorthAmerica	lung	sick	1
NorthAmerica	vagina	unknown	3
NorthAmerica	skin	sick	2
NorthAmerica	bronchial	sick	1
NorthAmerica	sputum	sick	50
NorthAmerica	throat	sick	12
NorthAmerica	plaque	unknown	4
NorthAmerica	rectal	healthy	3
NorthAmerica	urine	sick	19
NorthAmerica	ear	sick	1
NorthAmerica	stool	asymptomatic	3
Oceania	blood	unknown	2
Oceania	stool	unknown	2
Oceania	urine	sick	2
Oceania	stool	sick	1
Oceania	blood	sick	6
Oceania	sputum	sick	3
Oceania	urine	unknown	2
Oceania	throat	sick	1
SouthAmerica	skin	sick	2
SouthAmerica	blood	unknown	7
SouthAmerica	lung	unknown	1
SouthAmerica	nose	healthy	1
SouthAmerica	nose	asymptomatic	1
SouthAmerica	urine	unknown	1
SouthAmerica	blood	sick	9
SouthAmerica	stool	sick	4

SouthAmerica	stool	unknown	4
SouthAmerica	urine	sick	8
SouthAmerica	bone	sick	1
SouthAmerica	CSF	sick	1
SouthAmerica	nose	sick	1
unk	throat	sick	1
unk	stool	sick	3
unk	stool	unknown	42
unk	miscgut	sick	7
unk	miscgut	healthy	2
unk	tooth	unknown	1
unk	vagina	unknown	4
unk	blood	unknown	1
unk	miscgut	unknown	3
unk	colon	unknown	1
unk	UT	unknown	42
unk	throat	unknown	1
unk	nose	unknown	4
unk	skin	unknown	191
unk	urine	unknown	2
unk	sputum	unknown	1
unk	bone	unknown	1
unk	blood	sick	1

Supplementary Data 2: hyperparameter testing and data summaries

Hyperparameter testing

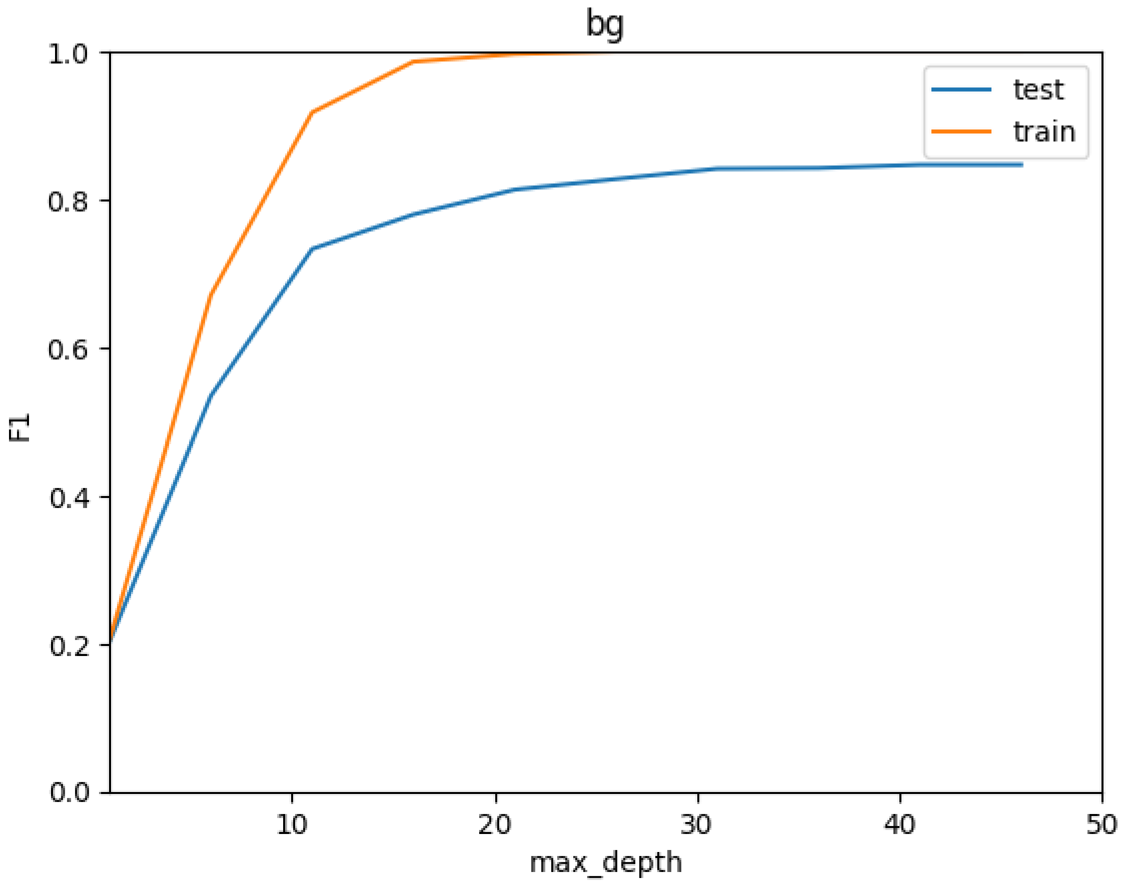

Figure S2: f1 score for max depth parameter range of 1-45.

Max depth is the maximum amount of times a tree can split. The default for sklearn's random forest classifier is none so the trees will continue to split until all 'leaves' only contain examples from one class or have reached the min_samples_split value. Our data starts to plateau around a depth of 15-20.

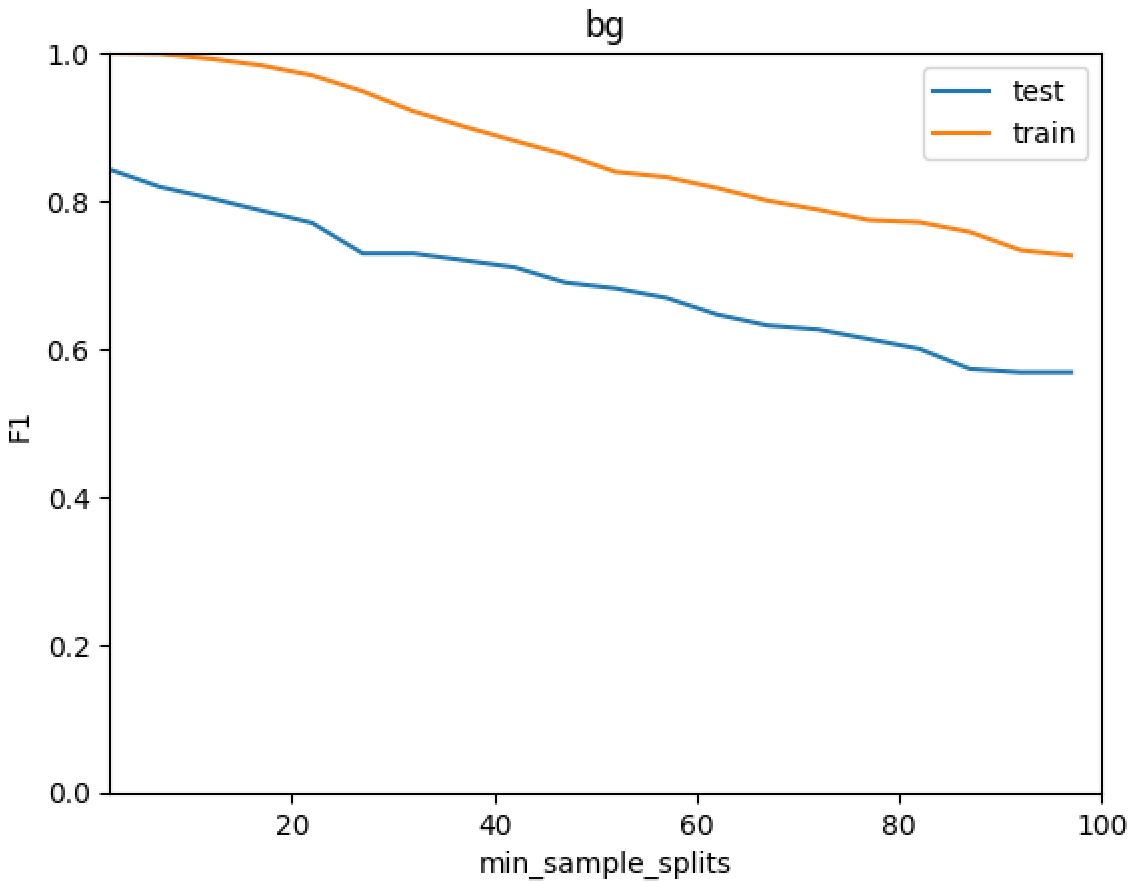

Figure S3: f1 score for min sample split parameter range of 2-100.

Min sample split is the minimum number of samples required for the tree to attempt to split the node further. The default for sklearn's random forest classifier is 2, the minimum possible number. Our results drop steadily, especially after ~10 for training data, suggesting that the optimal number is lower than that.

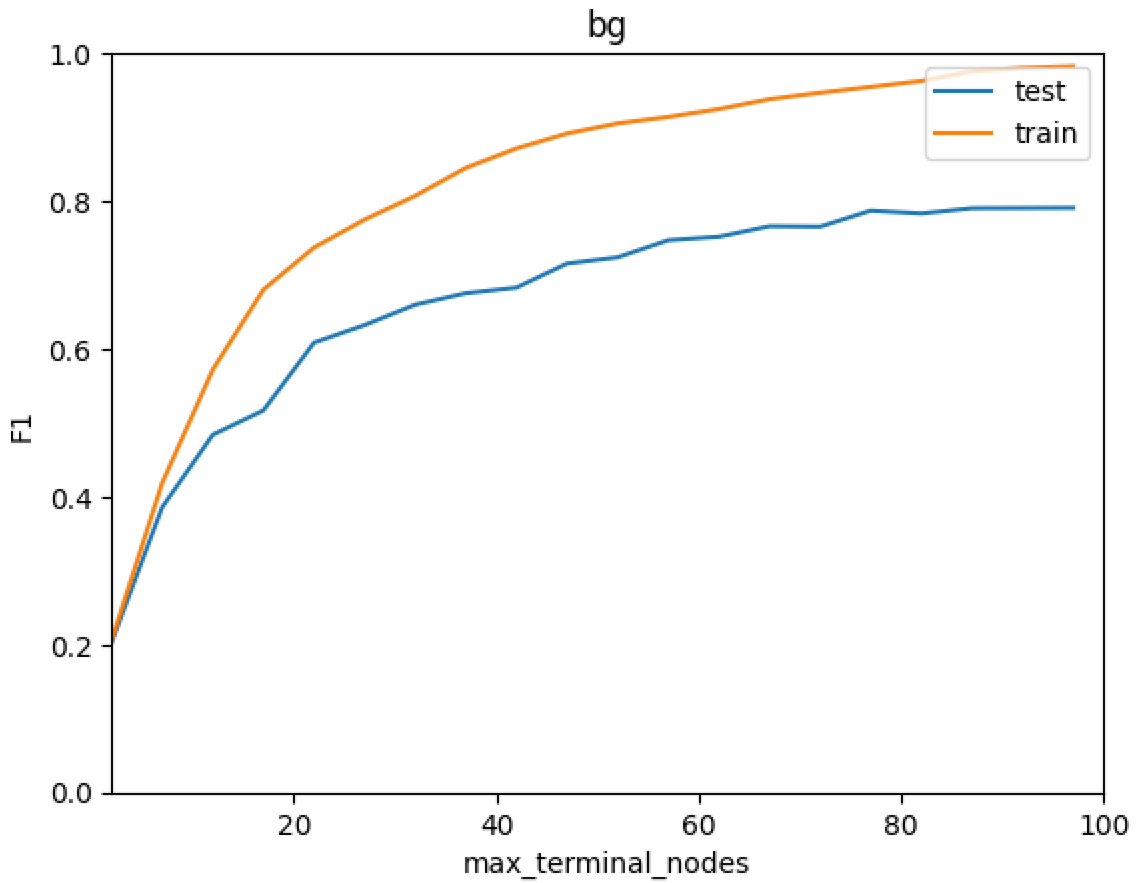

Figure S4: f1 score for max leaf nodes parameter range of 2-100.

Max leaf nodes is the number of final ‘leaves’ the trees can have. The default for sklearn’s random forest classifier is none, meaning that a tree can potentially have as many leaves as samples. The improvement to the score begins to slow at 17. Since the number of classes in the broad class set is 12, max leaf node settings below that should produce a low score as it is impossible to create leaves containing only samples of a single class if there are fewer leaves than classes.

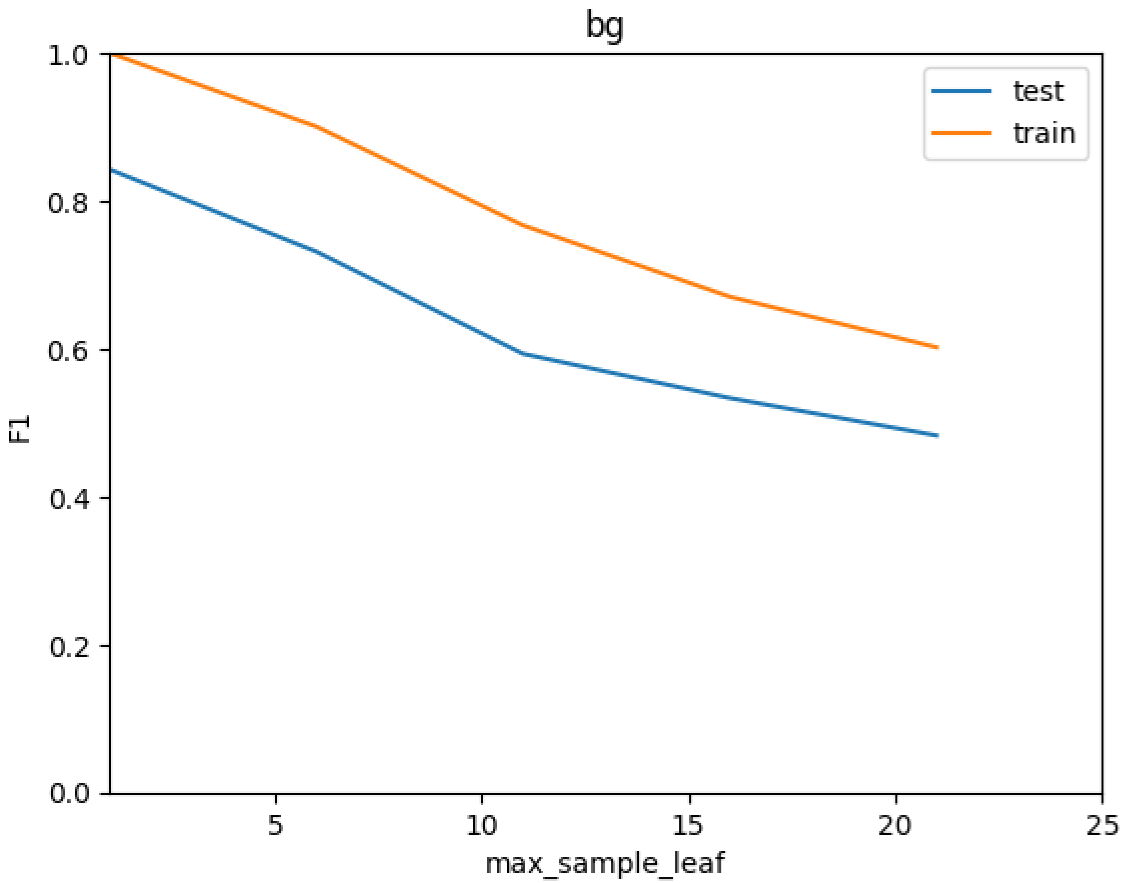

Figure S5: f1 score for min sample leaf parameter for ranges 1-21.

Min sample leaf is the minimum number of samples required for a node to be considered a leaf. The default for sklearn's random forest classifier is 1, meaning that a tree will continue splitting nodes into leaves until there are only samples from the same class in a leaf unless this conflicts with the min sample split parameter.

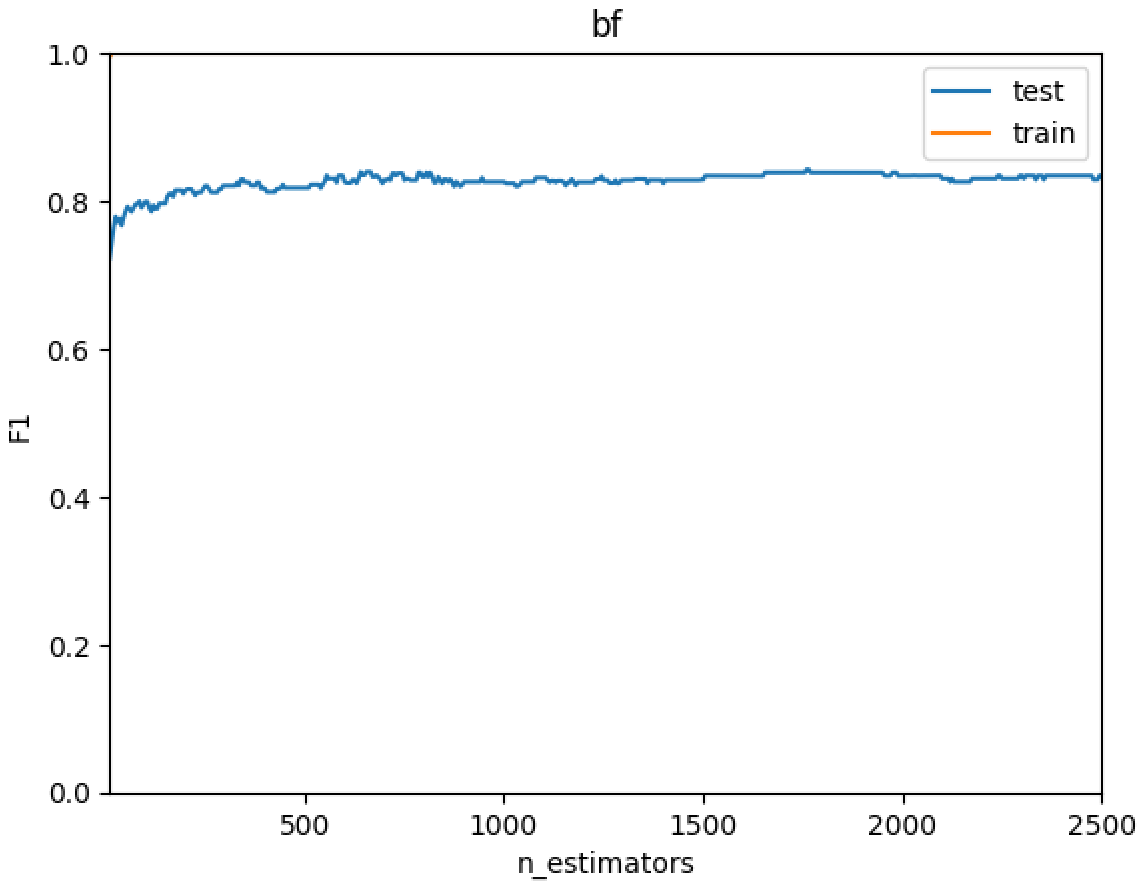

Figure S6: f1 score for the n estimators parameter for ranges 10-2500.

The n estimators parameter determines the number of trees in the forest. The default for the sklearn random forest classifier is 100. Our data showed that the score increased until ~500 trees, and the performance of the forests became more consistent at higher numbers of trees.

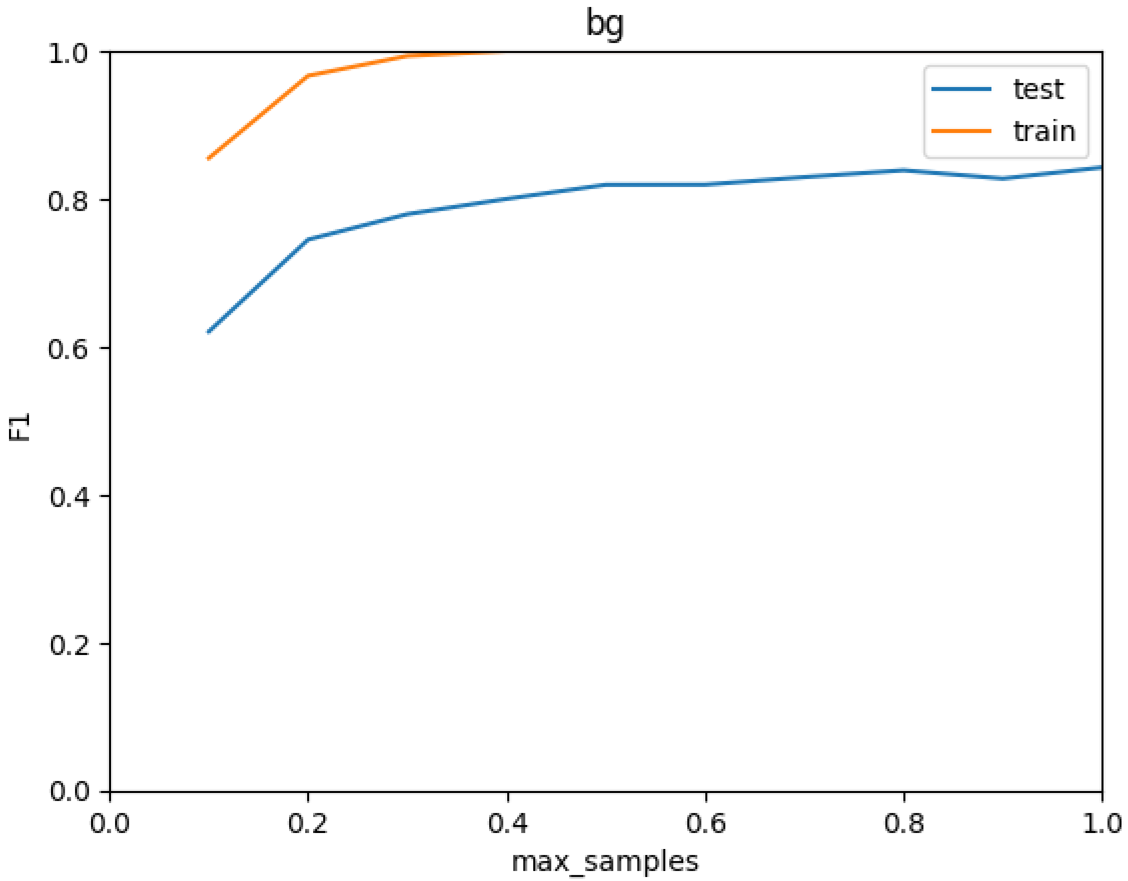

Figure S7: f1 score for the max samples parameter for a range of 0.1-1.0.

The max samples parameter controls what percentage of the training dataset is used to train each tree. Random forests do not always use every sample in the training dataset to train the trees and instead train each tree on a different random selection of the training data. This helps avoid overfitting by having the trees built with slightly different training sets, and allows for an approximation of accuracy by testing each tree on the data that it doesn't use instead of separating out a testing set to assess the forest. The default for sklearn's random forest classifier is none so it uses all the samples to train every forest. Out data shows that the score continues to improve as more data is used to train the tree but the training data is learned with 30% of the samples being used to train each tree and the score for the test set begins to plateau at 50% of the data being used for each tree.

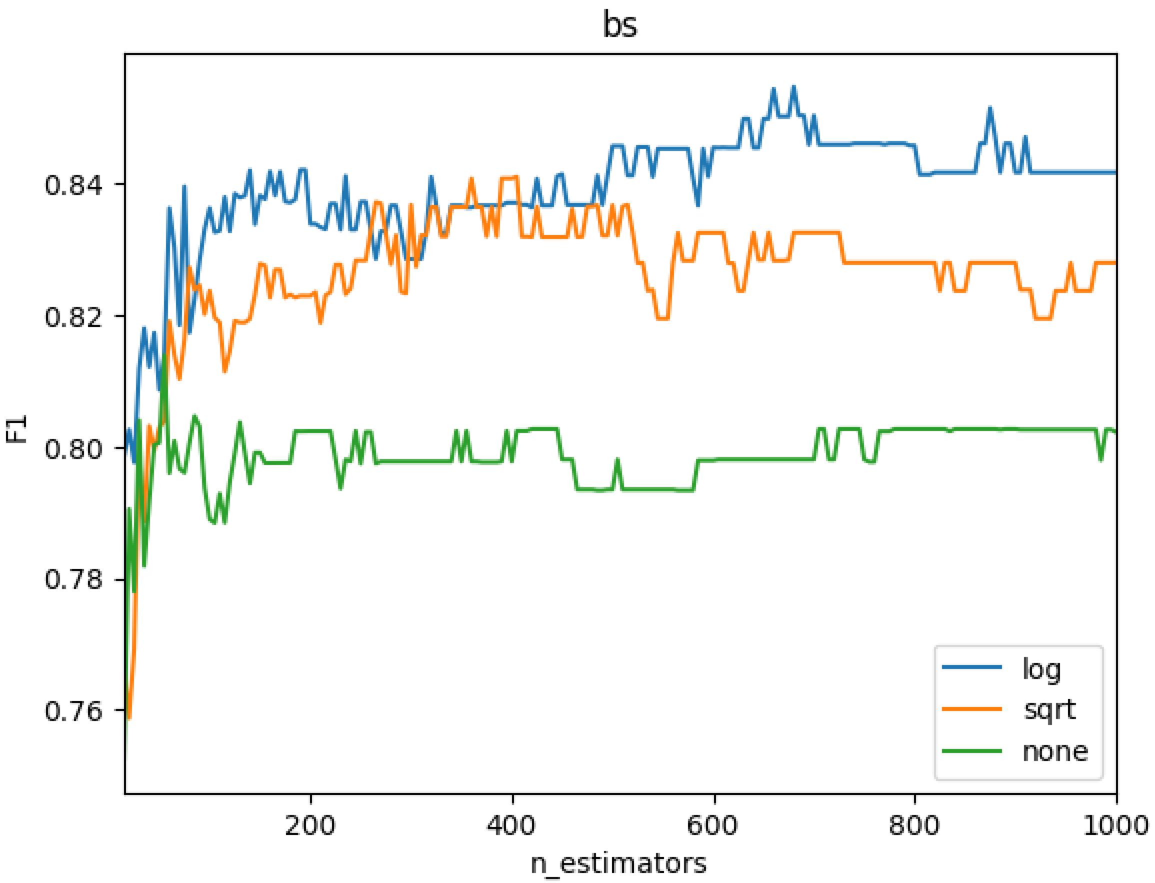

Figure S8: f1 score of three max features parameters; log2 of the number of features (log), square root of the number of features (sqrt) or all features (none)

The max features parameter controls the number of features that are considered when a tree attempts to split a node. None means that it will look through every feature while sqrt will use the square root of the number of features and log2 will use the number of features.

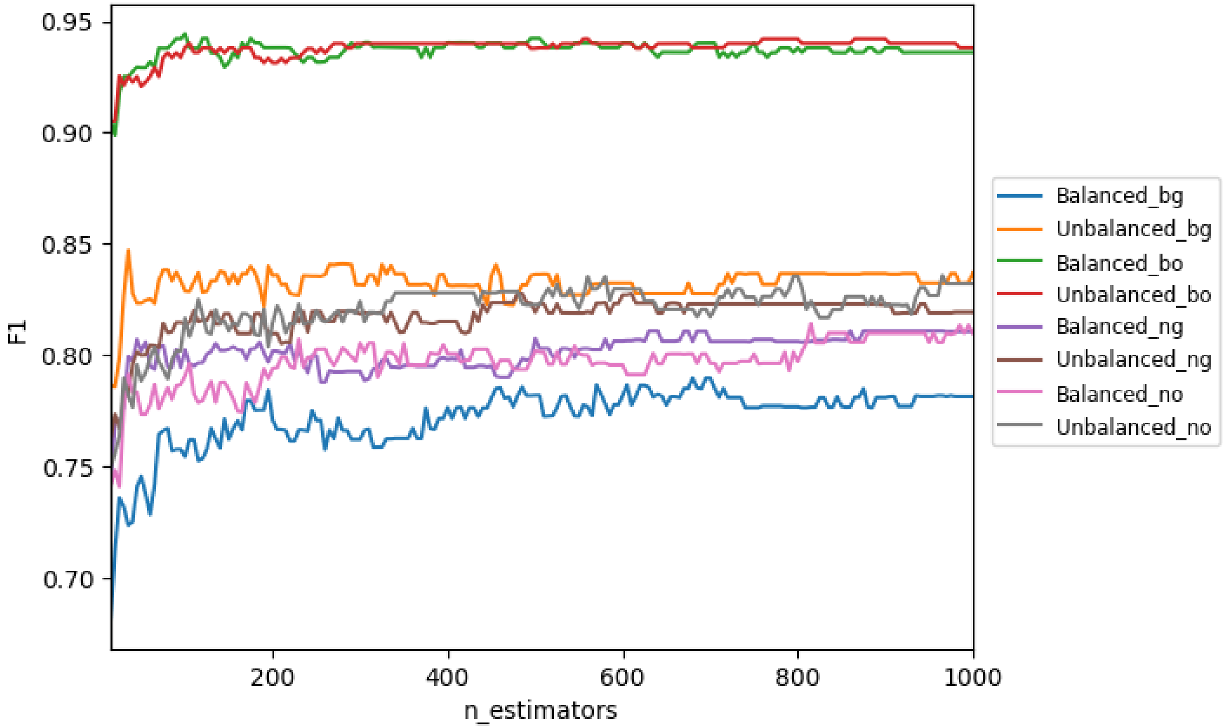

Figure S9: comparison of Unbalanced Random Forest, Random Forest, and SMOTE on both a balanced and an unbalanced dataset. The legend lists the methods applied. Variables with 'Balanced' used the Balanced Random Forest Classifier from imbalanced learn and unbalanced uses the regular Random Forest Classifier from sklearn. n' used the Narrow classes which were balanced, while variables beginning with 'b used the Broad classes which combined some Narrow classes into an unbalanced class set. 'o' denotes models that used the oversampling technique SMOTE from imbalance learn, while 'g' denotes models that did not implement oversampling.

The balanced random forest by imbalance learn is a method of reducing the issues caused by an imbalance dataset by downsampling the majority classes (Chao Chen 2004) while the Synthetic Minority Oversample Technique (SMOTE) does exactly what it says and creates synthetic examples of the minority class to oversample the minority class (Chawla et al. 2011). Balanced random forest had little effect on the f1-score, but using SMOTE() to oversample the minority classes.

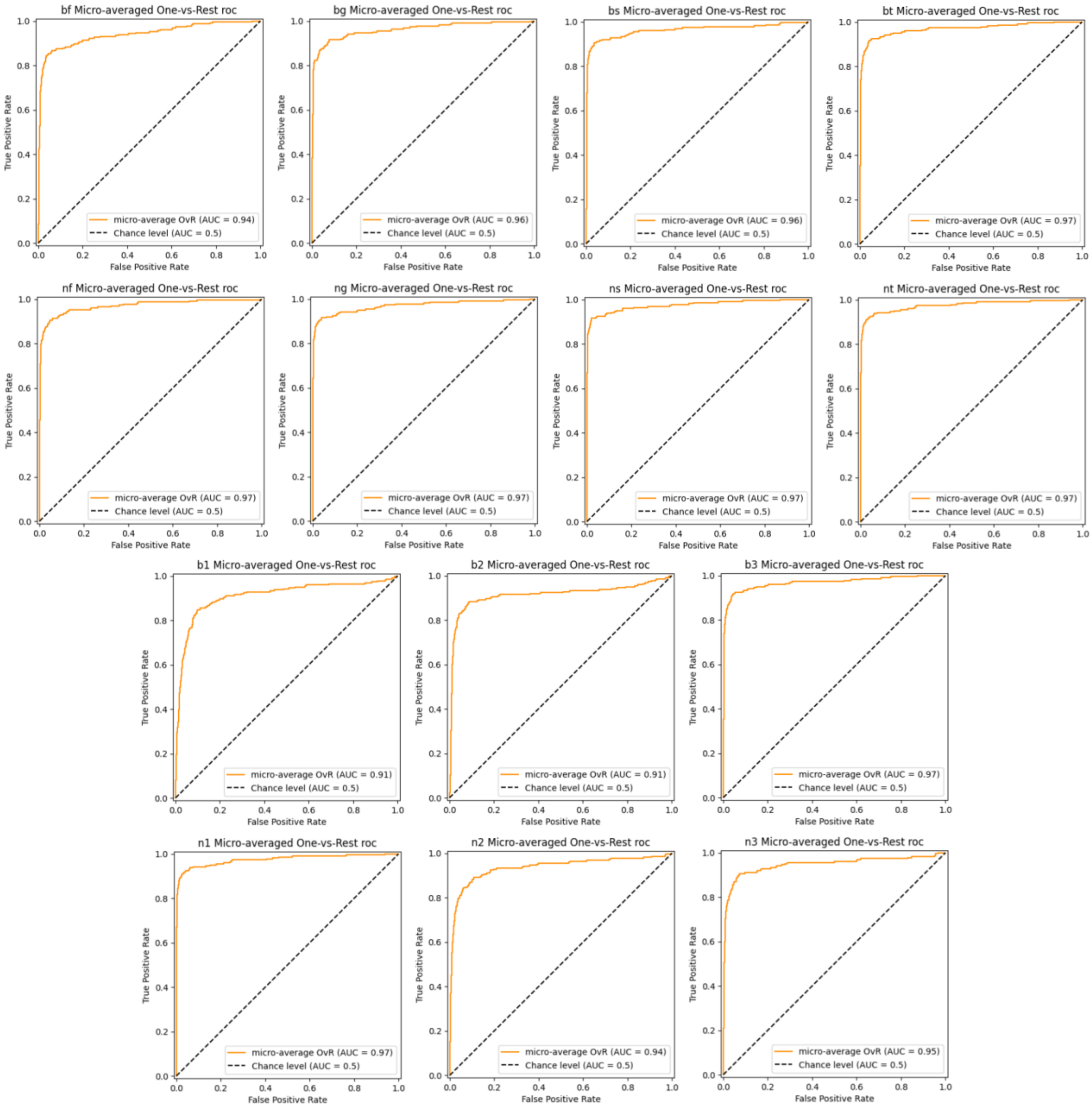

Figure S10: ROC-AUC curves for each model, averaged over each class. The top two rows are the taxonomic profiles, in the order of family, genus, species, strain, and with top row being the models trained with the broad class set, while the second row are the models trained with the narrow class set. The third row is the broad models for the functional subsystem levels 1-3, while the fourth row is their narrow class counterparts.

Nmds plots

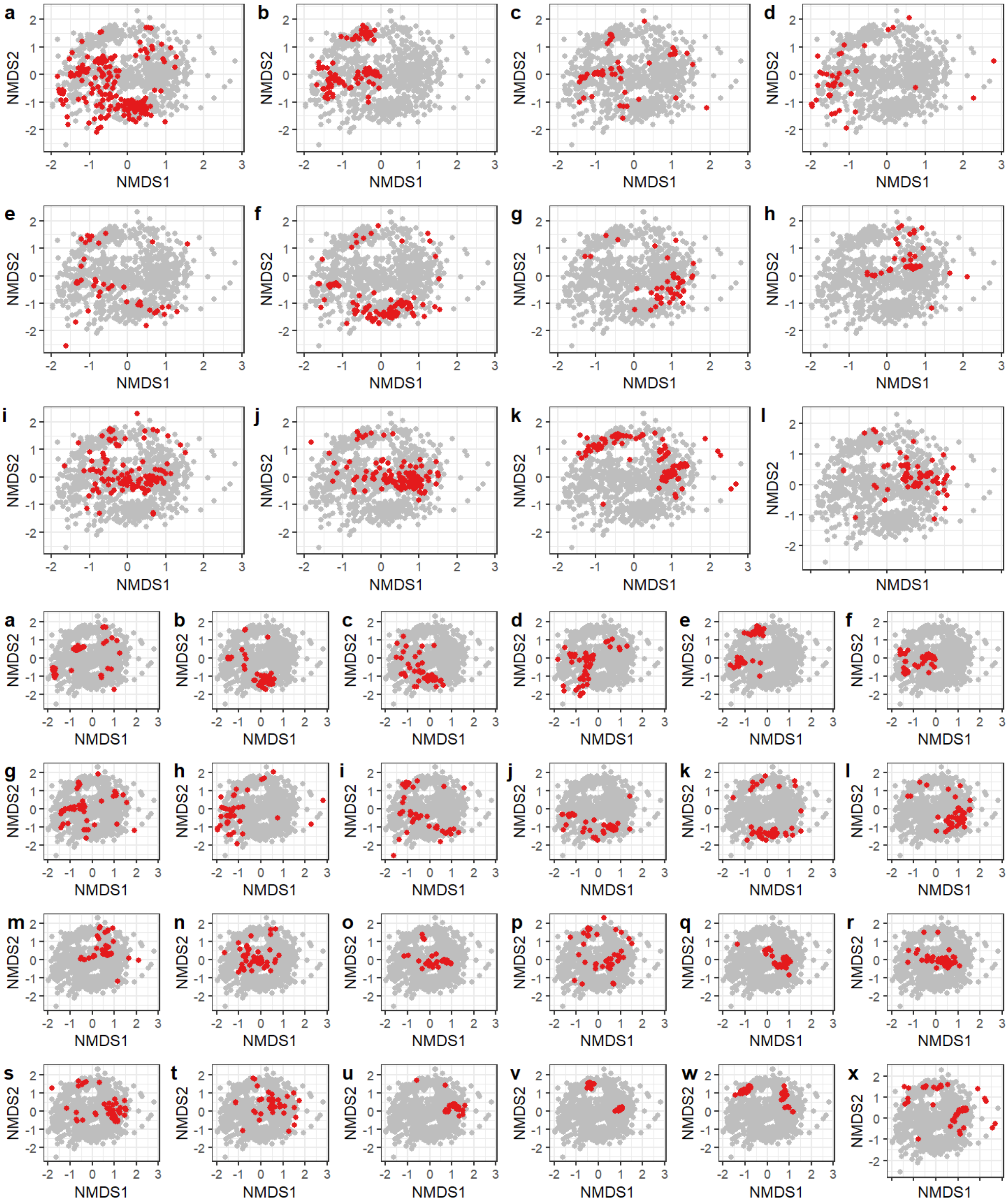

Figure S11: Nmds plots of the strain level FOCUS results split by isolation source classes. The top panel shows the Broad classes: a- animal, b- human respiratory, c- human skin, d- human

*other, e- human oral, f- human gastrointestinal, g- wastewater, h- built environment, i- marine, j- freshwater, k-soil, l- plant. The bottom panel shows narrow classes: a- animal_arthropod, b- animal_bird, c- animal_mammal, d- animal_other, e- human respiratory_lung, f- human respiratory_sputum, g- human skin, h- human other, i- human oral, j- human gastrointestinal_miscgut, k- human gastrointestinal_stool, l- wastewater, m- built environment, n- marine_, o- marine_coastal, p- marine_sediment, q- freshwater_sediment, r- freshwater_lake, s- freshwater_, t- plant_, u- plant_rhizosphere, v- soil_rhizosphere, w- soil_agricultural, x- soil_*

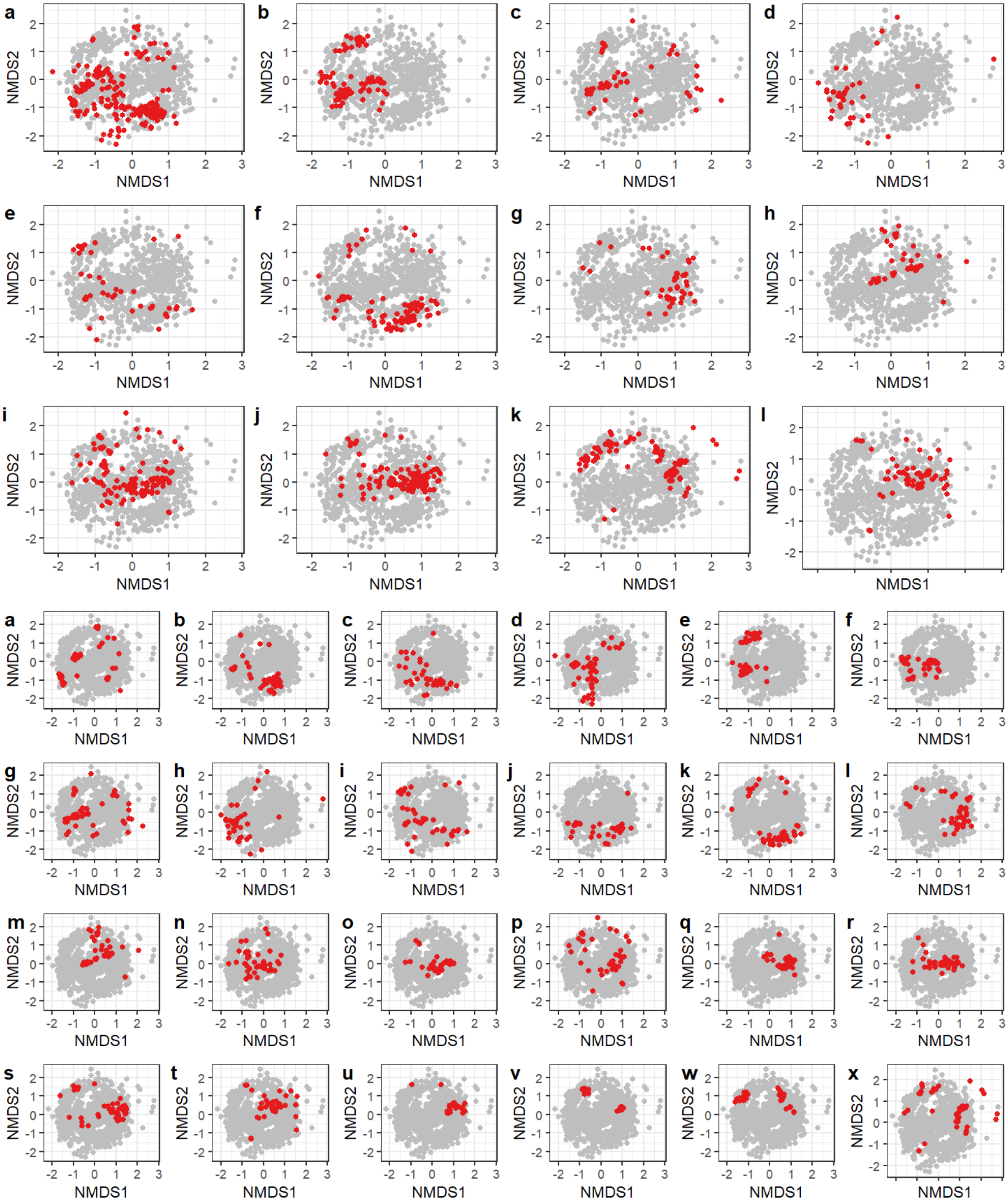

Figure S12: Nmds plots of the species level FOCUS results split by isolation source classes. The top panel shows the Broad classes: a- animal, b- human respiratory, c- human skin, d- human other, e- human oral, f- human gastrointestinal, g- wastewater, h- built environment, i- marine, j- freshwater, k-soil, l- plant. The bottom panel shows narrow classes: a- animal_arthropod, b- animal_bird, c- animal_mammal, d- animal_other, e- human

*respiratory_lung, f- human respiratory_sputum, g- human skin, h- human other, i- human oral, j- human gastrointestinal_miscgut, k- human gastrointestinal_stool, l- wastewater, m- built environment, n- marine_, o- marine_coastal, p- marine_sediment, q- freshwater_sediment, r- freshwater_lake, s- freshwater_, t- plant_, u- plant_rhizosphere, v- soil_rhizosphere, w- soil_agricultural, x- soil_*

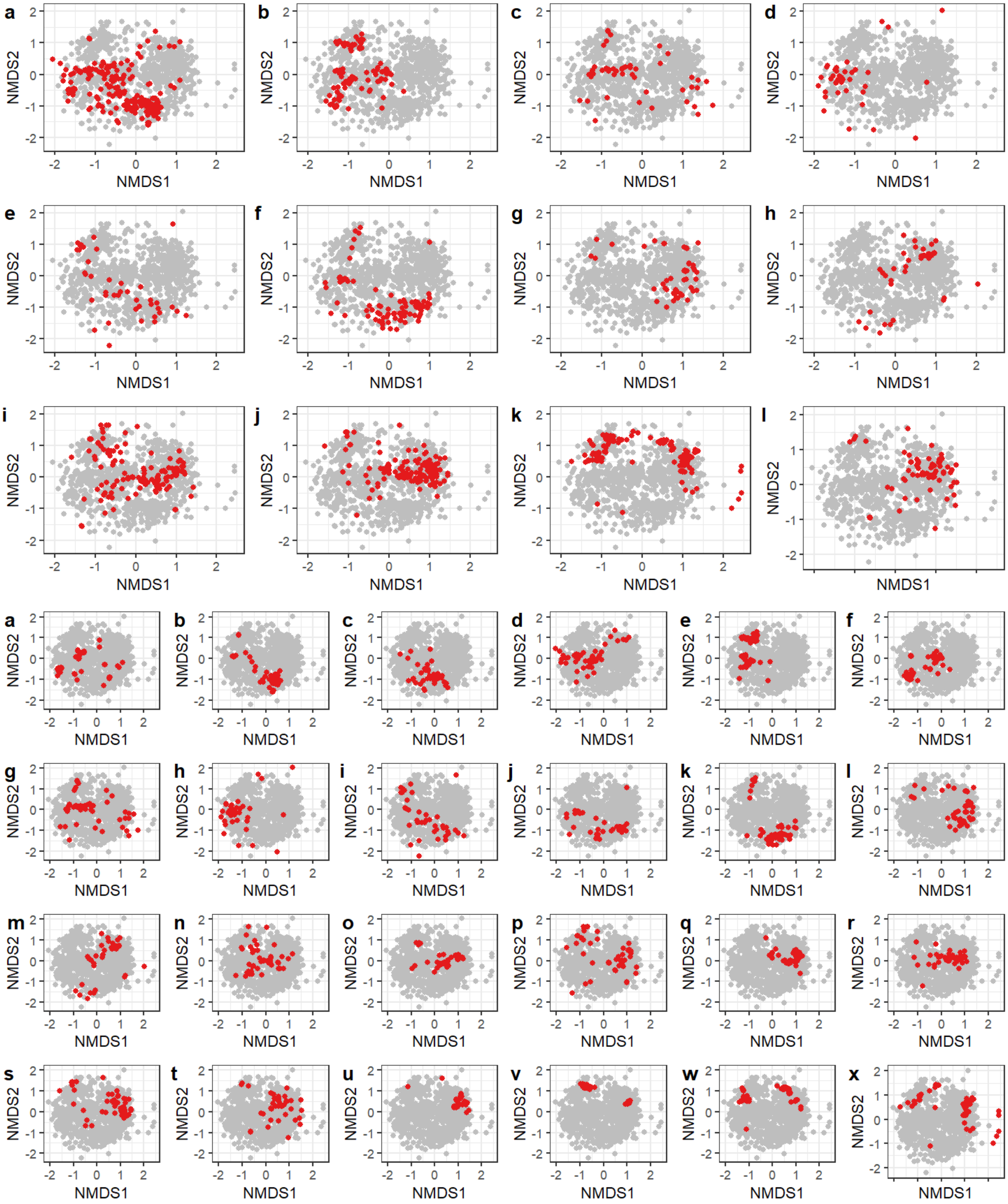

Figure S13: Nmds plots of the genus level FOCUS results split by isolation source classes. The top panel shows the Broad classes: a- animal, b- human respiratory, c- human skin, d- human other, e- human oral, f- human gastrointestinal, g- wastewater, h- built environment, i- marine, j- freshwater, k-soil, l- plant. The bottom panel shows narrow classes: a- animal_arthropod, b- animal_bird, c- animal_mammal, d- animal_other, e- human respiratory_lung, f- human

*respiratory_sputum, g- human skin, h- human other, i- human oral, j- human gastrointestinal_miscgut, k- human gastrointestinal_stool, l- wastewater, m- built environment, n- marine_, o- marine_coastal, p- marine_sediment, q- freshwater_sediment, r- freshwater_lake, s- freshwater_, t- plant_, u- plant_rhizosome, v- soil_rhizosphere, w- soil_agricultural, x- soil_*

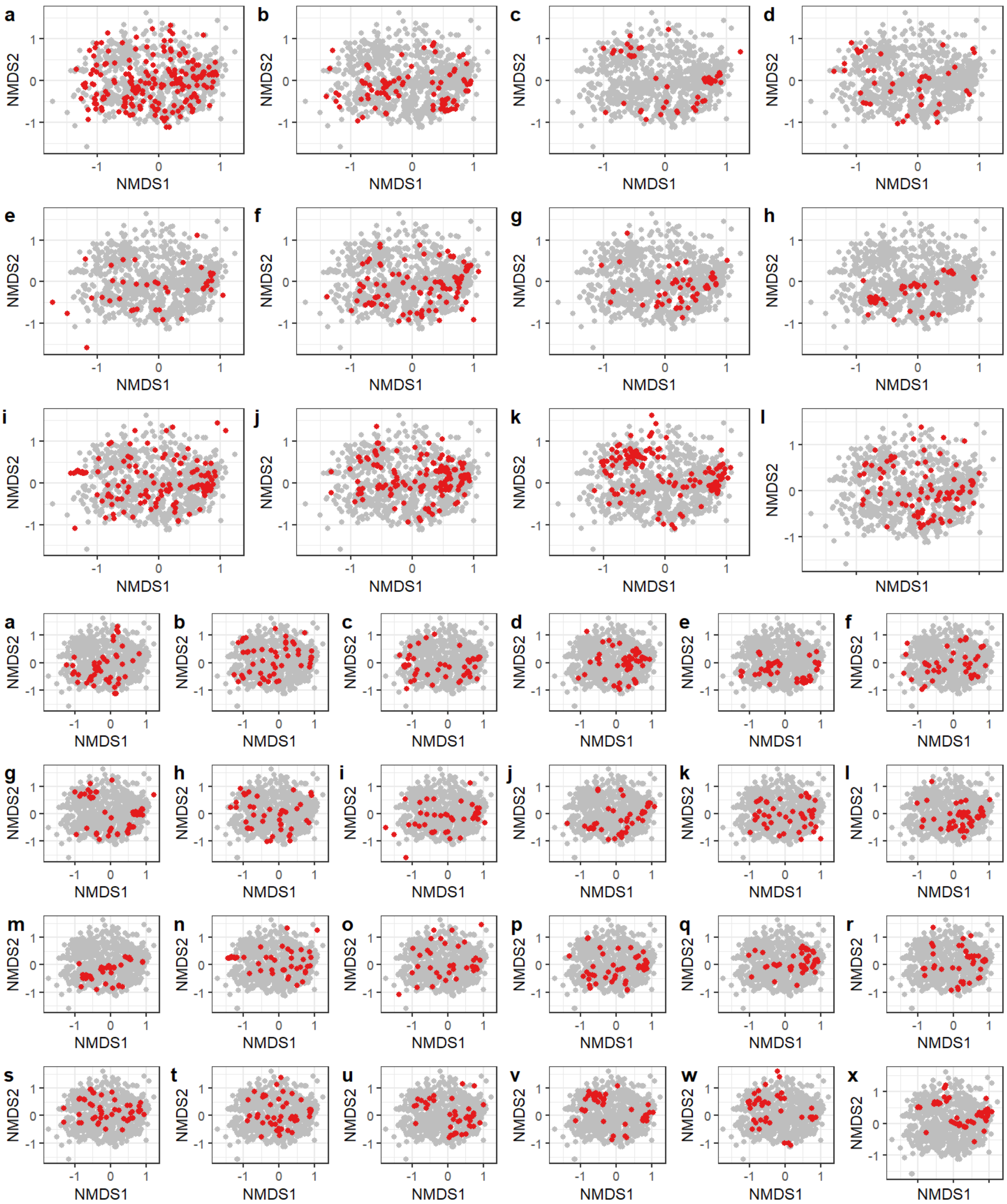

Figure S14: Nmds plots of the family level FOCUS results split by isolation source classes. The top panel shows the Broad classes: a- animal, b- human respiratory, c- human skin, d- human other, e- human oral, f- human gastrointestinal, g- wastewater, h- built environment, i- marine, j- freshwater, k-soil, l- plant. The bottom panel shows narrow classes: a- animal_arthropod, b- animal_bird, c- animal_mammal, d- animal_other, e- human respiratory_lung, f- human

*respiratory_sputum, g- human skin, h- human other, i- human oral, j- human gastrointestinal_miscgut, k- human gastrointestinal_stool, l- wastewater, m- built environment, n- marine_, o- marine_coastal, p- marine_sediment, q- freshwater_sediment, r- freshwater_lake, s- freshwater_, t- plant_, u- plant_rhizosphere, v- soil_rhizosphere, w- soil_agricultural, x- soil_*

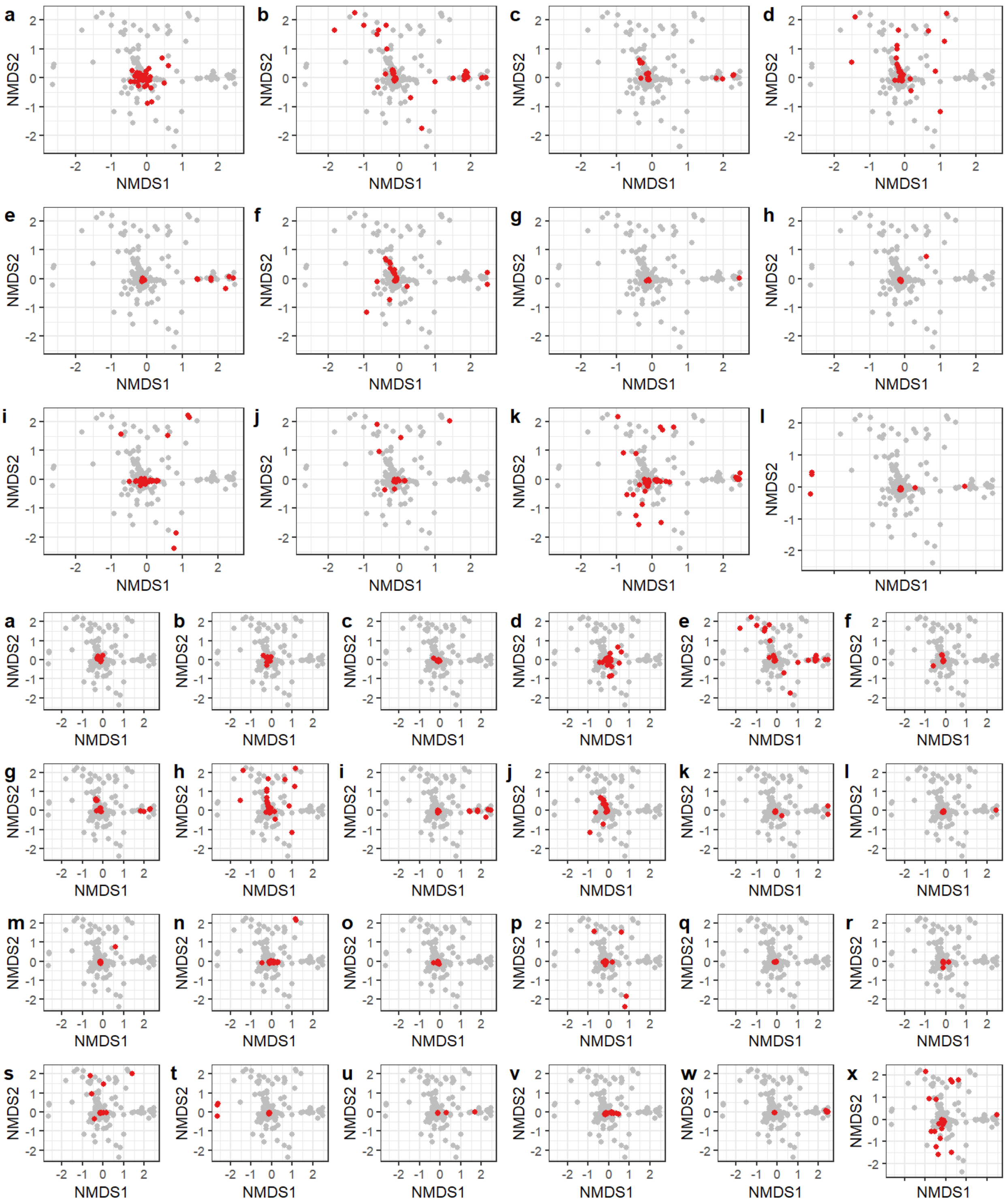

Figure S15: Nmds plots of the subsystem_level_1 level SUPERFOCUS results split by isolation source classes. The top panel shows the Broad classes: a- animal, b- human respiratory, c- human skin, d- human other, e- human oral, f- human gastrointestinal, g- wastewater, h- built environment, i- marine, j- freshwater, k-soil, l- plant. The bottom panel shows narrow classes: a- animal_arthropod, b- animal_bird, c- animal_mammal, d- animal_other, e- human respiratory_lung, f- human respiratory_sputum, g- human skin, h- human other, i- human oral, j-

*human gastrointestinal_miscgut, k- human gastrointestinal_stool, l- wastewater, m- built environment, n- marine_, o- marine_coastal, p- marine_sediment, q- freshwater_sediment, r- freshwater_lake, s- freshwater_, t- plant_, u- plant_rhizosphere, v- soil_rhizosphere, w- soil_agricultural, x- soil_*

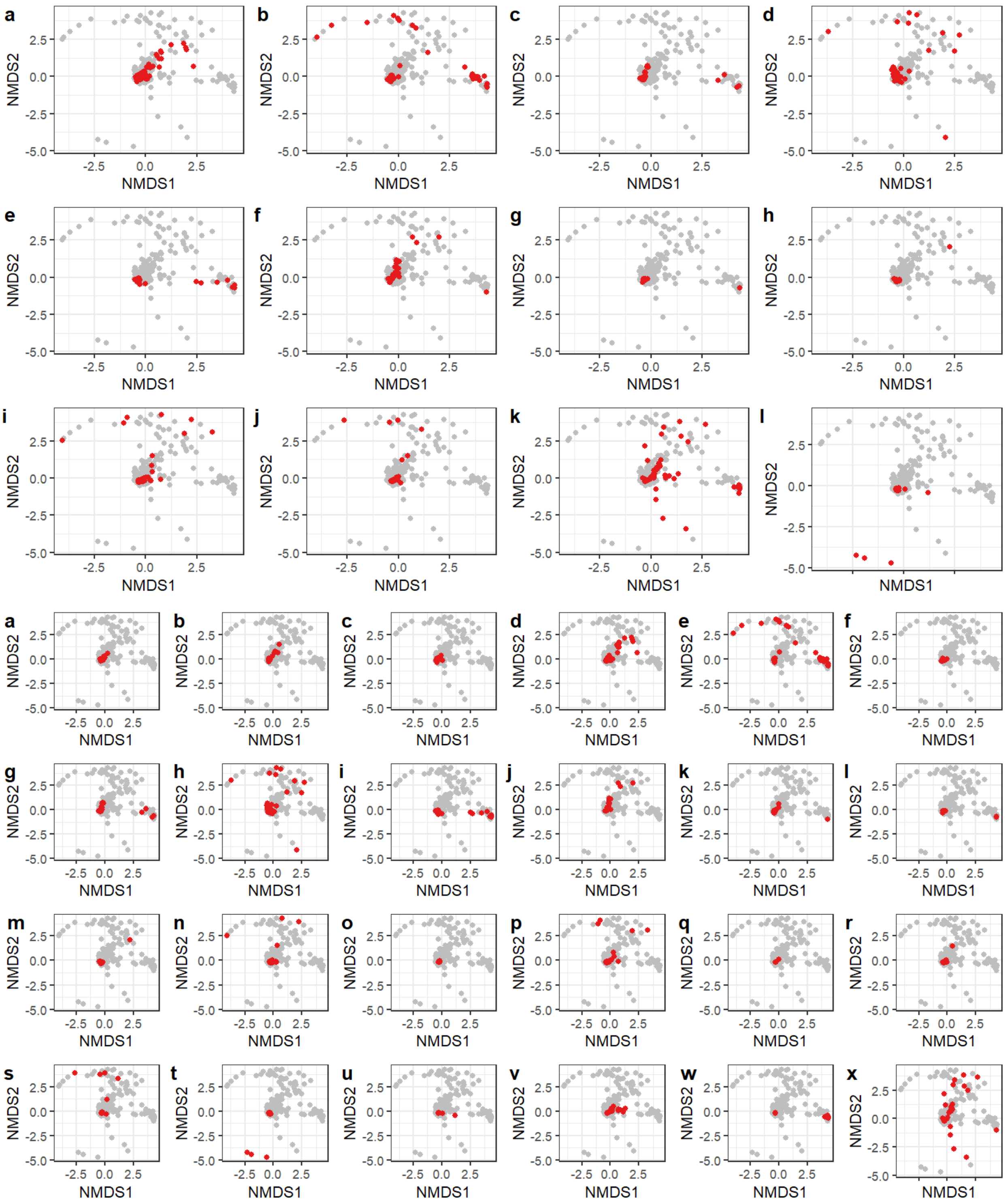

Figure S16: Nmds plots of the subsystem_level_2 level SUPERFOCUS results split by isolation source classes. The top panel shows the Broad classes: a- animal, b- human respiratory, c- human skin, d- human other, e- human oral, f- human gastrointestinal, g- wastewater, h- built environment, i- marine, j- freshwater, k-soil, l- plant. The bottom panel shows narrow classes: a- animal_arthropod, b- animal_bird, c- animal_mammal, d- animal_other, e- human

*respiratory_lung, f- human respiratory_sputum, g- human skin, h- human other, i- human oral, j- human gastrointestinal_miscgut, k- human gastrointestinal_stool, l- wastewater, m- built environment, n- marine_, o- marine_coastal, p- marine_sediment, q- freshwater_sediment, r- freshwater_lake, s- freshwater_, t- plant_, u- plant_rhizosome, v- soil_rhizosphere, w- soil_agricultural, x- soil_*

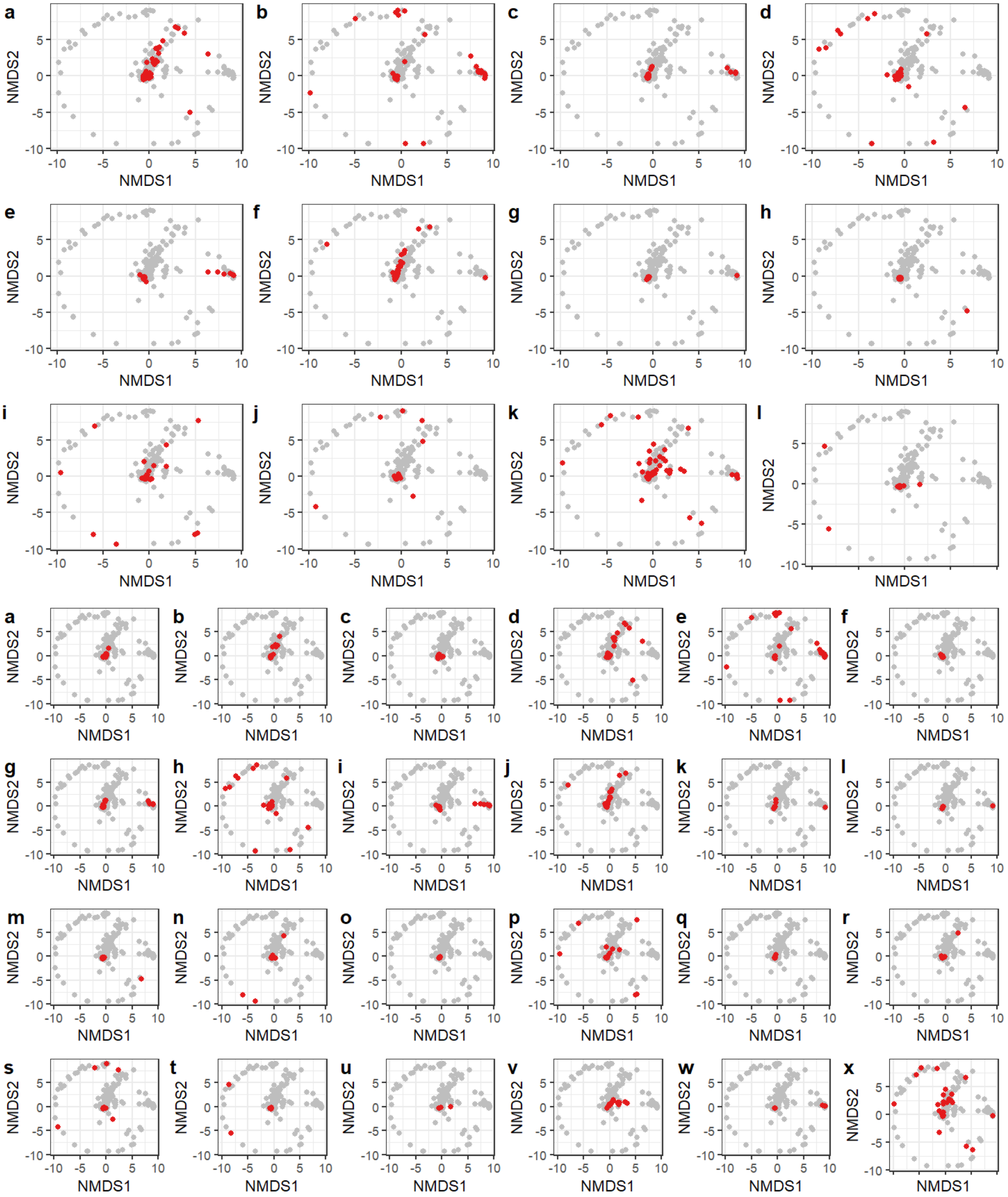

Figure S17: Nmds plots of the subsystem_level_3 level SUPERFOCUS results split by isolation source classes. The top panel shows the Broad classes: a- animal, b- human respiratory, c- human skin, d- human other, e- human oral, f- human gastrointestinal, g- wastewater, h- built environment, i- marine, j- freshwater, k-soil, l- plant. The bottom panel shows narrow classes: a- animal_arthropod, b- animal_bird, c- animal_mammal, d- animal_other, e- human

*respiratory_lung, f- human respiratory_sputum, g- human skin, h- human other, i- human oral, j- human gastrointestinal_miscgut, k- human gastrointestinal_stool, l- wastewater, m- built environment, n- marine_, o- marine_coastal, p- marine_sediment, q- freshwater_sediment, r- freshwater_lake, s- freshwater_, t- plant_, u- plant_rhizosphere, v- soil_rhizosphere, w- soil_agricultural, x- soil_*

Confusion matrices

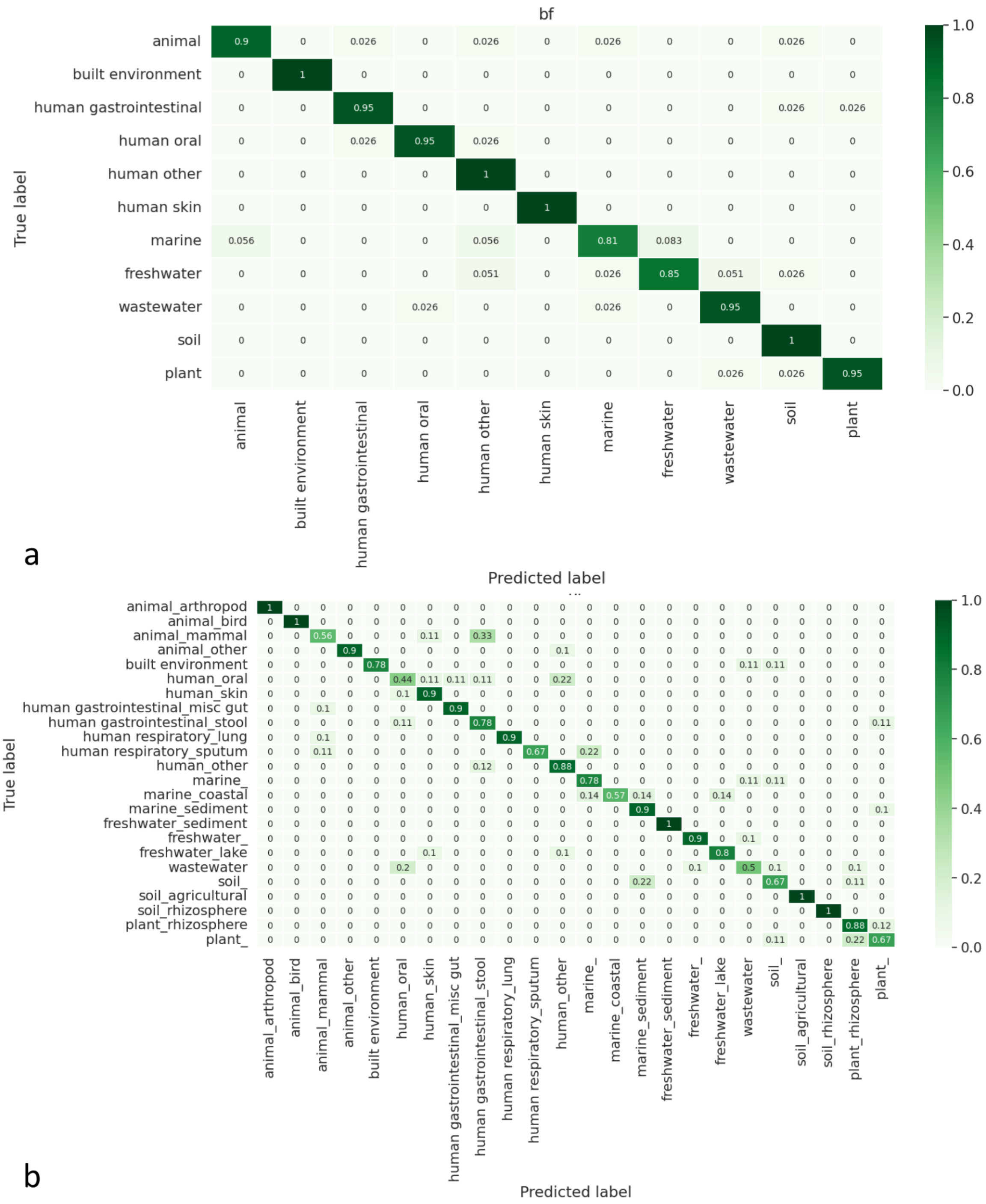

Figure S18: confusion matrices for a- Broad_family and b- Narrow_family

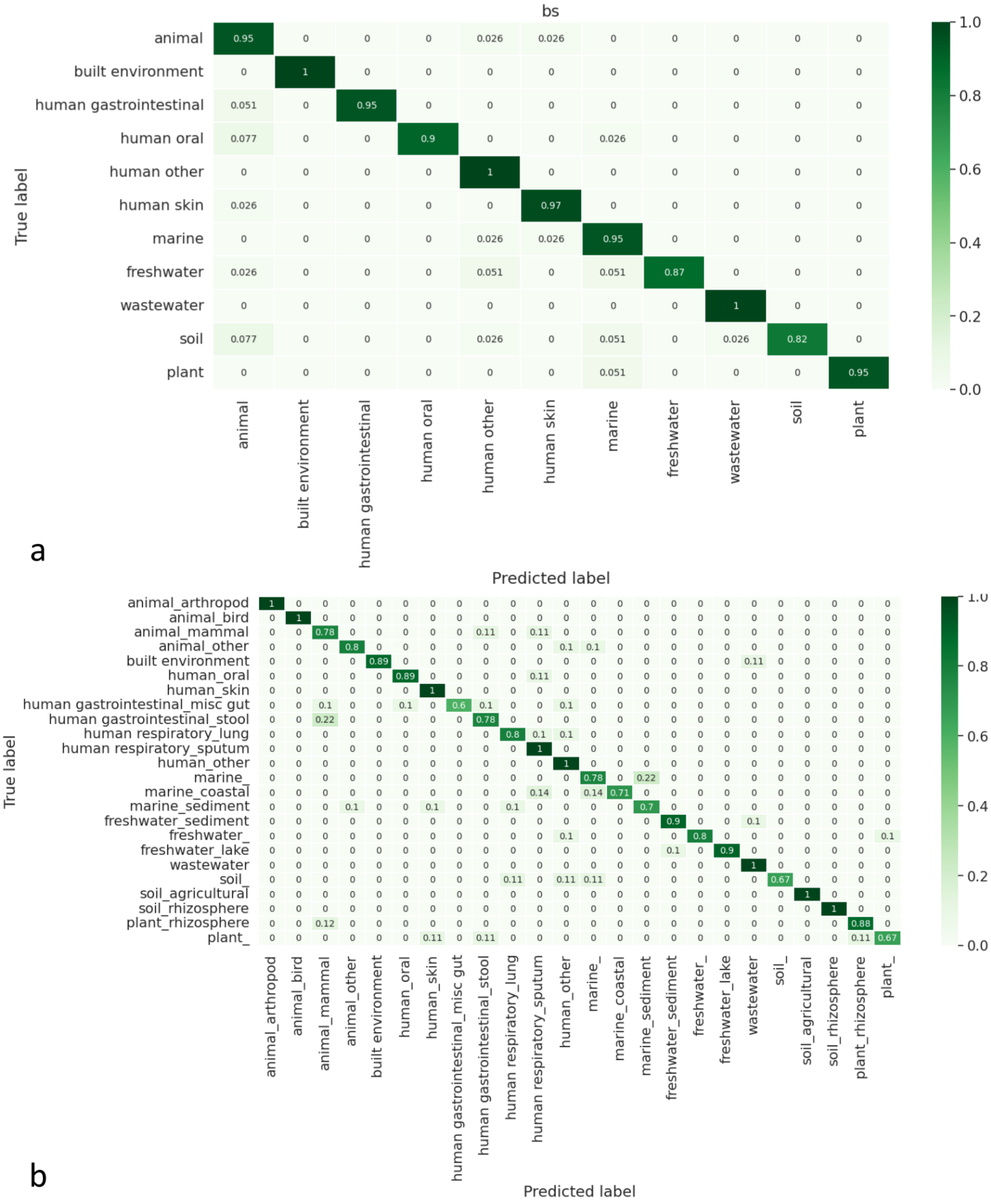

Figure S19: confusion matrices for a- Broad_species and b- Narrow_species

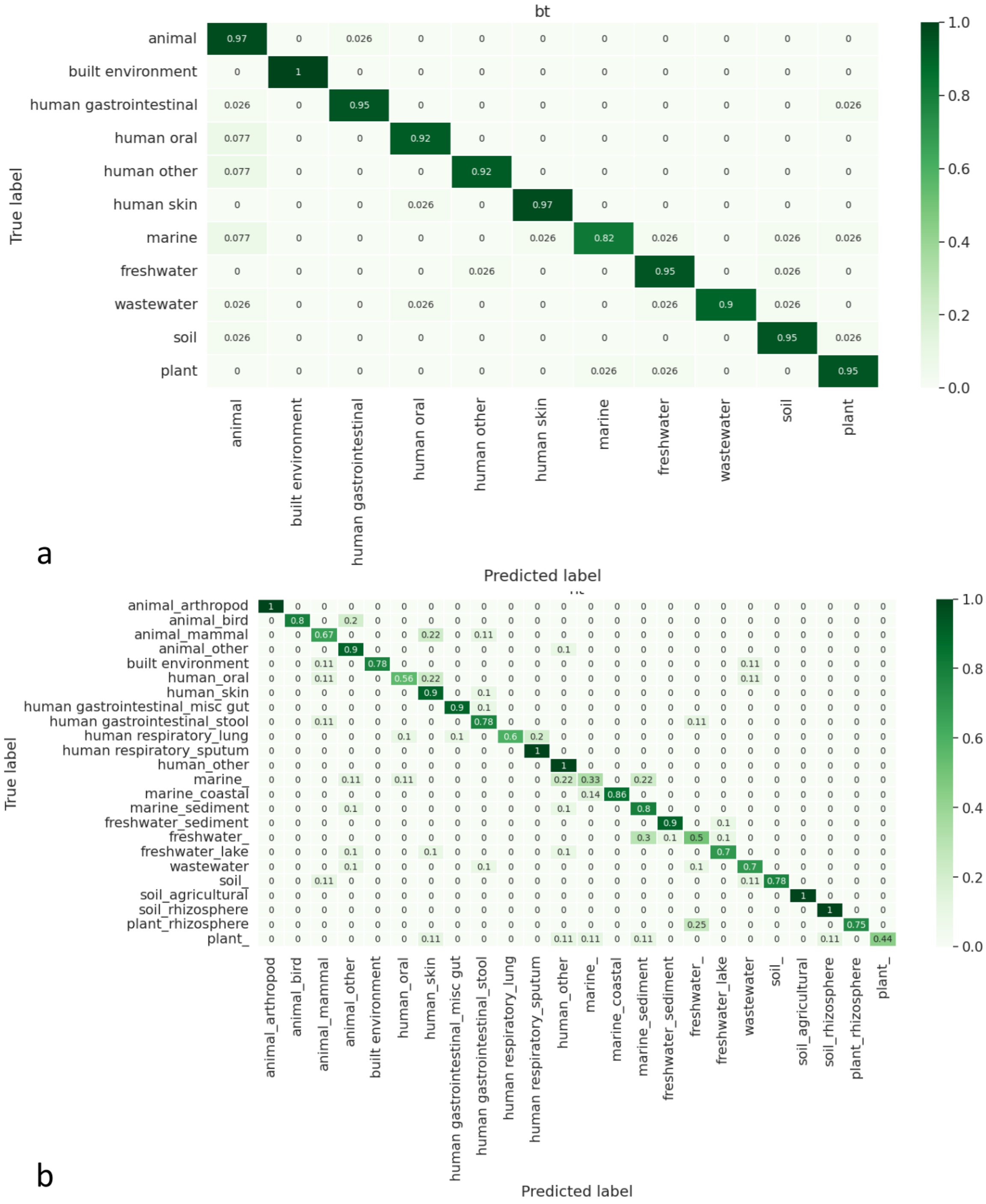

Figure S20: confusion matrices for a- Broad_strain and b- Narrow_strain

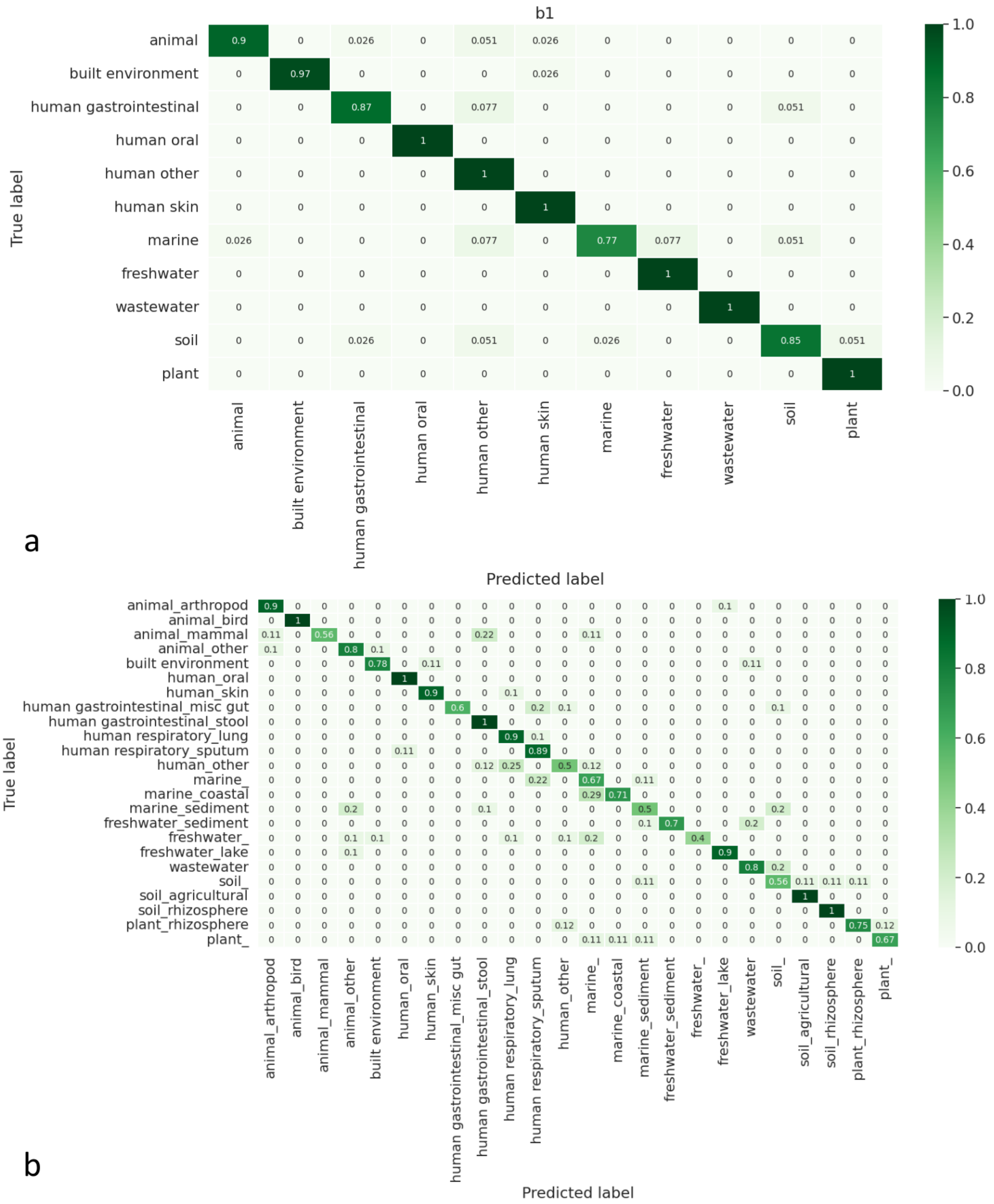

Figure S21: confusion matrices for a- Broad_subsystem_1 and b- Narrow_subsystem_1

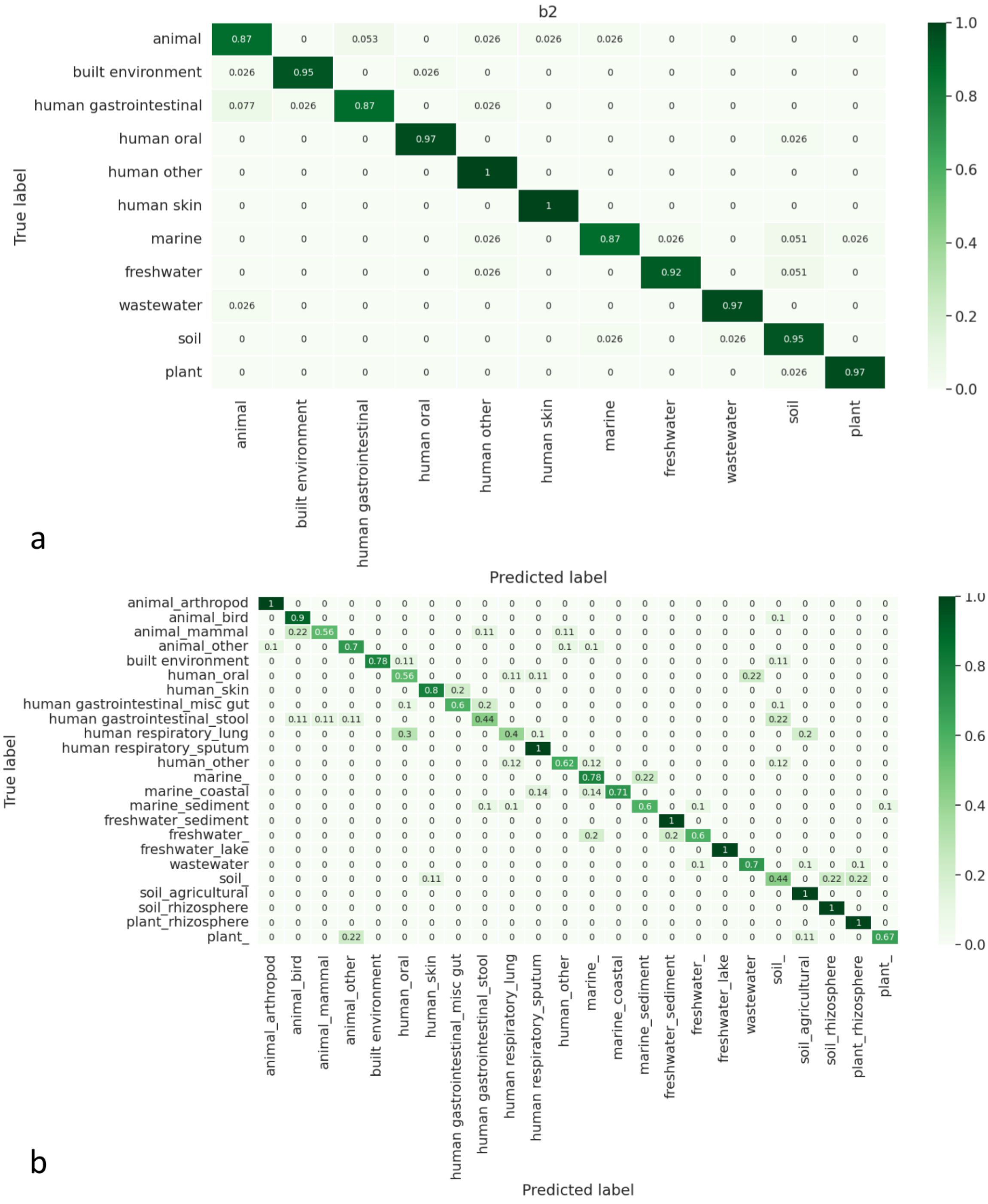

Figure S22: confusion matrices for a- Broad_subsystem_2 and b- Narrow_subsystem_2

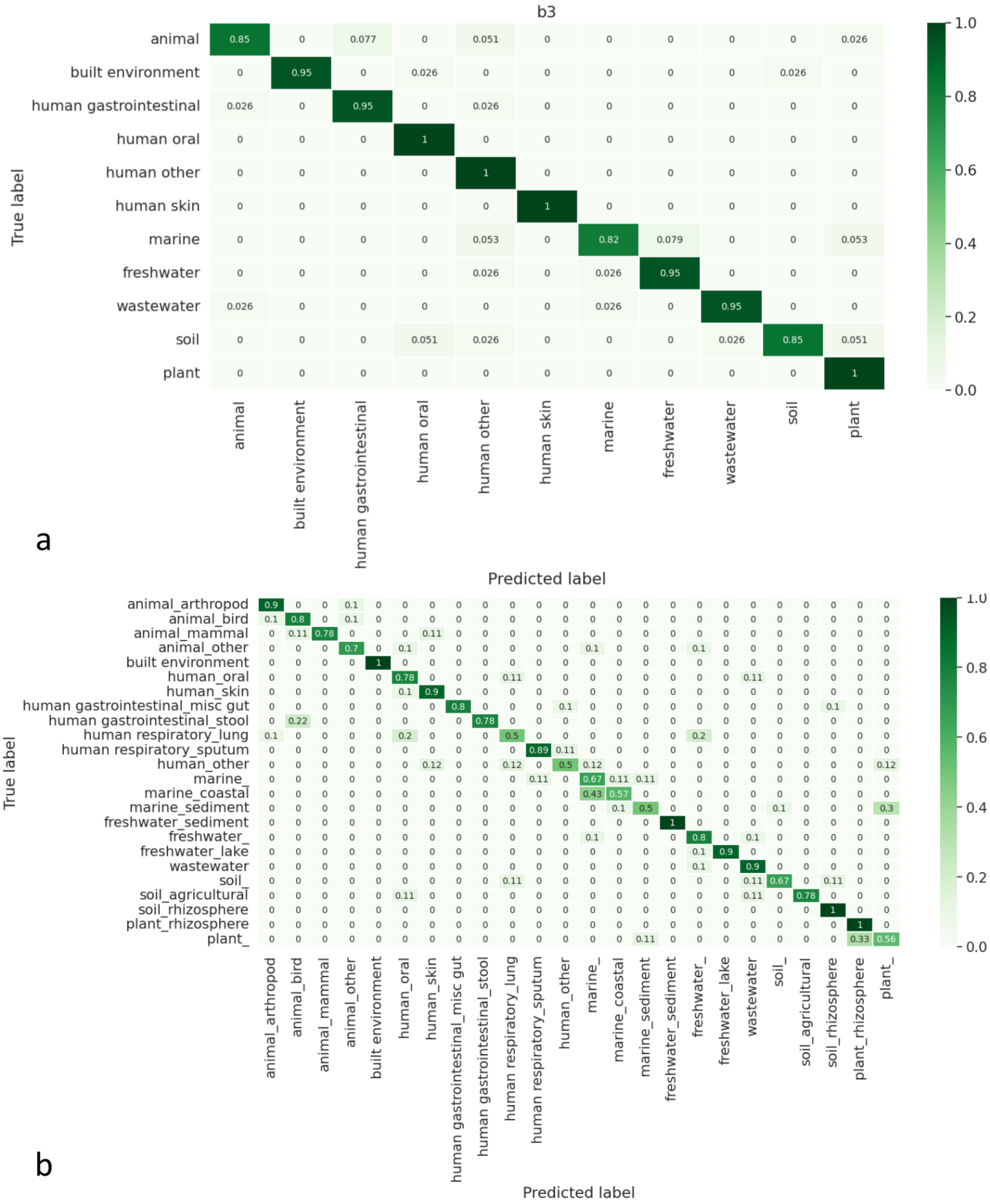

Figure S23: confusion matrices for a- Broad_subsystem_3 and b- Narrow_subsystem_3

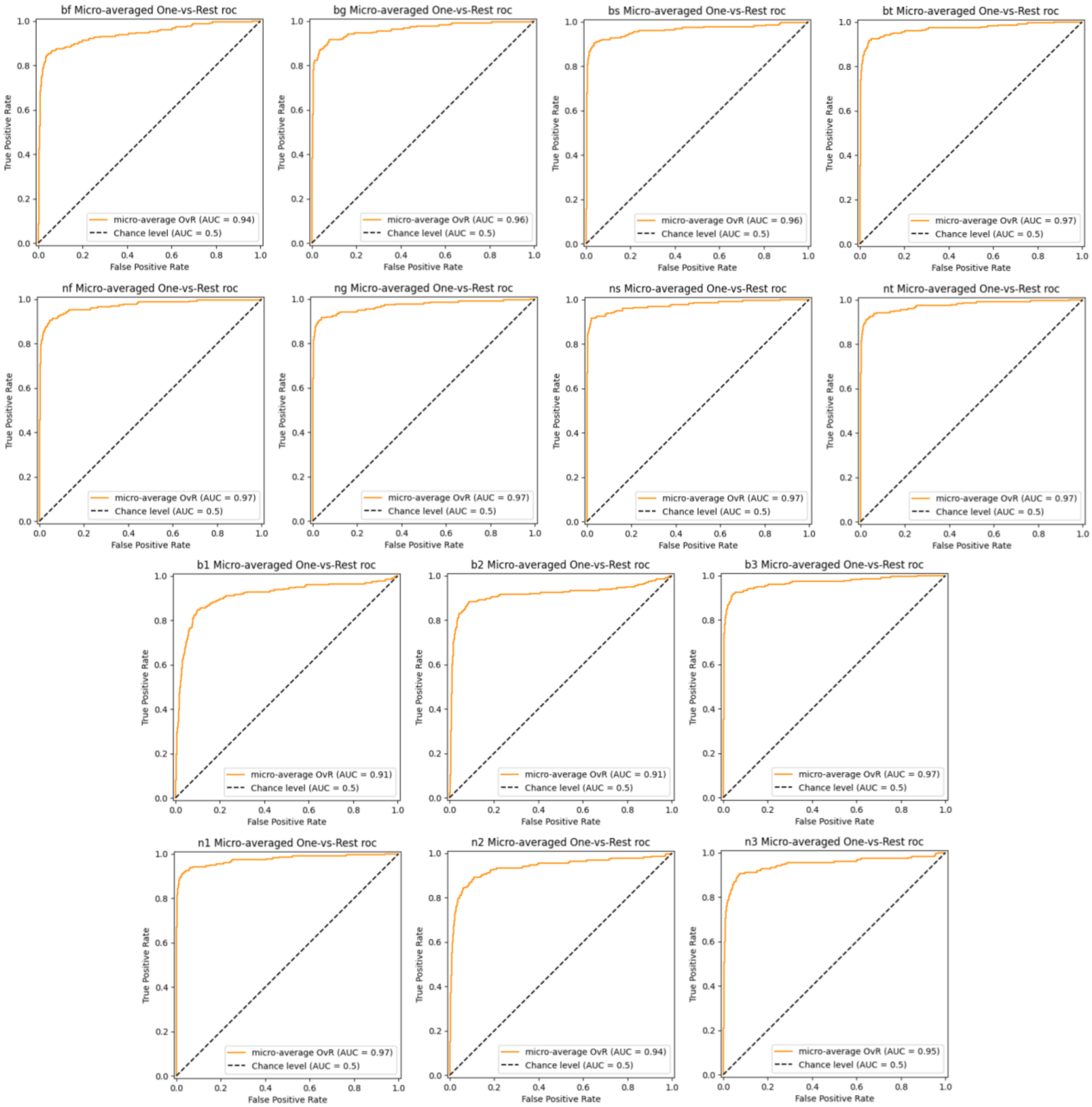

Figure S24:ROC-AUC curves for each model. From left to right: Broad_Family, Broad_Genus, Broad_Species, Broad_strain, Narrow_Family, Narrow_Genus, Narrow_Species, Narrow_strain, Broad_Subsystem1, Broad_Subsystem2, Broad_Subsystem3, Narrow_Subsystem1, Narrow_Subsystem2, Narrow_Subsystem3.

Supplementary Data 3: Data availability

The datasets created/analysed in Results Chapter 1 are available at:
2a- NCBI Genome Assemblies Summary Archive 20220601
https://doi.org/10.25451/flinders.22299664.v2
2b- Prophage predictions https://doi.org/10.25451/flinders.c.6629843
2c- Archive of the PATRIC Metadata from 20220601 (Wattam et al. 2017)
https://doi.org/10.25451/flinders.22299655.v2
2d- Prophages in humans (this paper), https://doi.org/10.25451/flinders.24564379.v3

The datasets created/analysed in Results Chapter 2 are available at:
2e- Inglis, Laura (2025). AMRfinder+ results. Flinders University. Dataset.
https://doi.org/10.25451/flinders.28282106.v1
2f- Inglis, Laura (2025). List of unique antimicrobial resistance genes found in prophage regions. Flinders University. Dataset. https://doi.org/10.25451/flinders.28282070.v1

The datasets created/analysed in Results Chapter 3 are available at:
2g- Inglis, Laura (2025). Training and test sets. Flinders University. Dataset.
https://doi.org/10.25451/flinders.29312828.v1
2h- Inglis, Laura (2025). Prediction probabilities for unseen test set. Flinders University. Dataset.
https://doi.org/10.25451/flinders.29312792.v1

Supplementary Data 4: Code

```
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix,
precision_score, recall_score, ConfusionMatrixDisplay, f1_score
from sklearn.model_selection import RandomizedSearchCV, train_test_split
from scipy.stats import randint
from sklearn.tree import export_graphviz
from sklearn import tree
from IPython.display import Image
import graphviz
from matplotlib import pyplot as plt
from matplotlib.collections import LineCollection
from sklearn import manifold
from sklearn.decomposition import PCA
from sklearn.metrics import euclidean_distances
from sklearn.metrics import classification_report
from scipy.spatial.distance import pdist
from sklearn.inspection import permutation_importance
from sklearn.linear_model import LogisticRegression
from collections import OrderedDict
from matplotlib import pyplot
from imblearn.ensemble import BalancedRandomForestClassifier
from sklearn.model_selection import cross_validate
from sklearn.model_selection import RepeatedStratifiedKFold
import seaborn as sns
import sys
from sklearn.metrics import roc_auc_score
from sklearn.metrics import auc, roc_curve
from sklearn.preprocessing import LabelBinarizer
from sklearn.metrics import RocCurveDisplay
from imblearn.over_sampling import SMOTE
import xgboost as xgb
from sklearn import preprocessing
from xgboost import XGBClassifier
from sklearn.linear_model import Lasso
from sklearn.model_selection import GridSearchCV, KFold
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import precision_recall_curve

#props to datacamp for the tutorial
BGdata = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'broadgenus')
NGdata = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'narrowgenus')
BFdata = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'broadfam')
NFdata = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'narrowfam')
BSdata = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'broadspecies')
NSdata = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'narrowspecies')
BTdata = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'broadstrain')
NTdata = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'narrowstrain')
N1data = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'narrow1')
B1data = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'broad1')
N2data = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'narrow2')
B2data = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'broad2')
N3data = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'narrow3')
```

```
B3data = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'broad3')
#split into features (X) and data (Y)
BGX=BGdata.drop('classifier', axis = 1)
BGY=BGdata['classifier']
#split the data into training and test- can change the size of this
overG = SMOTE()
OGX, OGY = overG.fit_resample(BGX, BGY)
bgx_train, bgx_test, bgy_train, bgy_test = train_test_split(OGX, OGY,
test_size=0.2, random_state=1, stratify=OGY)
#training the random forest- might need tweaking if parameters arent too
accurate
#defaults to 100 trees in the forest
#gini critereon
#mininum of 2 samples needed to split a node and minimum of 1 sample needed
for a branch to end (with only 50 samples per category it is possible that
only one from a cat gets randomly added to tree, so maybe leave this)
BRF = RandomForestClassifier(n_estimators=2500, bootstrap=True,
warm_start=True, max_features="log2")
BRF.fit(bgx_train, bgy_train)

#run the prediction
BGprediction = BRF.predict(bgx_test)
#checking run accuracy
BGaccuracy = f1_score(bgy_test, BGprediction, average='weighted')
print(f"Broad_genus =", BGaccuracy)

#narrow genus
NGX=NGdata.drop('classifier', axis = 1)
NGY=NGdata['classifier']
ngx_train, ngx_test, ngy_train, ngy_test = train_test_split(NGX, NGY,
test_size=0.2, random_state=2, stratify=NGY)
BNG = RandomForestClassifier(n_estimators=2500, bootstrap=True,
warm_start=True, max_features="log2")
BNG.fit(ngx_train, ngy_train)
NGprediction = BNG.predict(ngx_test)
NGaccuracy = f1_score(ngy_test, NGprediction, average='weighted')
print(f"Narrow_genus =", NGaccuracy)

# broad fam
BFX=BFdata.drop('classifier', axis = 1)
BFY=BFdata['classifier']
overF = SMOTE()
OFX, OFY = overF.fit_resample(BFX, BFY)
bfx_train, bfx_test, bfy_train, bfy_test = train_test_split(OFX, OFY,
test_size=0.2, random_state=3, stratify=OFY)
BBF = RandomForestClassifier(n_estimators=2500, bootstrap=True,
warm_start=True, max_features="log2")
BBF.fit(bfx_train, bfy_train)
```

```
BFprediction = BBF.predict(bfx_test)
BFaccuracy = f1_score(bfy_test, BFprediction, average='weighted')
print(f"Broad_family =", BFaccuracy)

# narrow fam
NFX=NFdata.drop('classifier', axis = 1)
NFY=NFdata['classifier']
nfx_train, nfx_test, nfy_train, nfy_test = train_test_split(NFX, NFY,
test_size=0.2, random_state=4, stratify=NFY)
BNF = RandomForestClassifier(n_estimators=2500, bootstrap=True,
warm_start=True, max_features="log2")
BNF.fit(nfx_train, nfy_train)
NFprediction = BNF.predict(nfx_test)
NFaccuracy = f1_score(nfy_test, NFprediction, average='weighted')
print(f"Narrow_family =", NFaccuracy)

# broad species
BSX=BSdata.drop('classifier', axis = 1)
BSY=BSdata['classifier']
overS = SMOTE()
OSX, OSY = overS.fit_resample(BSX, BSY)
bsx_train, bsx_test, bsy_train, bsy_test = train_test_split(OSX, OSY,
test_size=0.2, random_state=5, stratify=OSY)
BBS = RandomForestClassifier(n_estimators=2500, bootstrap=True,
warm_start=True, max_features="log2")
BBS.fit(bsx_train, bsy_train)
BSprediction = BBS.predict(bsx_test)
BSaccuracy = f1_score(bsy_test, BSprediction, average='weighted')
print(f"Broad_species =", BSaccuracy)

# narrow species
NSX=NSdata.drop('classifier', axis = 1)
NSY=NSdata['classifier']
nsx_train, nsx_test, nsy_train, nsy_test = train_test_split(NSX, NSY,
test_size=0.2, random_state=6, stratify=NSY)
BNS = RandomForestClassifier(n_estimators=2500, bootstrap=True,
warm_start=True, max_features="log2")
BNS.fit(nsx_train, nsy_train)
NSprediction = BNS.predict(nsx_test)
NSaccuracy = f1_score(nsy_test, NSprediction, average='weighted')
print(f"Narrow_species =", NSaccuracy)


#broad strain
BTX=BTdata.drop('classifier', axis = 1)
BTY=BTdata['classifier']
overT = SMOTE()
OTX, OTY = overT.fit_resample(BTX, BTY)
```

```
btx_train, btx_test, bty_train, bty_test = train_test_split(OTX, OTY,
test_size=0.2, random_state=7, stratify=OTY)
broadA = RandomForestClassifier(n_estimators=2500, bootstrap=True,
warm_start=True, max_features="log2")
broadA.fit(btx_train, bty_train)
predictionA = broadA.predict(btx_test)
accuracyA = f1_score(bty_test, predictionA, average='weighted')
print(f"Broad_strain =", accuracyA)

#narrow strain
NTX=NTdata.drop('classifier', axis = 1)
NTY=NTdata['classifier']
ntx_train, ntx_test, nty_train, nty_test = train_test_split(NTX, NTY,
test_size=0.2, random_state=8, stratify=NTY)
narrowT = RandomForestClassifier(n_estimators=2500, bootstrap=True,
warm_start=True, max_features="log2")
narrowT.fit(ntx_train, nty_train)
predictionT = narrowT.predict(ntx_test)
accuracyT = f1_score(nty_test, predictionT, average='weighted')
print(f"Narrow_strain =", accuracyT)

#broad subsystem level 1
B1X=B1data.drop('classifier', axis = 1)
B1Y=B1data['classifier']
over1 = SMOTE()
O1X, O1Y = over1.fit_resample(B1X, B1Y)
b1x_train, b1x_test, b1y_train, b1y_test = train_test_split(O1X, O1Y,
test_size=0.2, random_state=9, stratify=O1Y)
broad1 = RandomForestClassifier(n_estimators=2500, bootstrap=True,
warm_start=True, max_features="log2")
broad1.fit(b1x_train, b1y_train)
predictionB1 = broad1.predict(b1x_test)
accuracyB1 = f1_score(b1y_test, predictionB1, average='weighted')
print(f"Broad_subsystem 1 =", accuracyB1)

#narrow subsystem level 1
N1X=N1data.drop('classifier', axis = 1)
N1Y=N1data['classifier']
n1x_train, n1x_test, n1y_train, n1y_test = train_test_split(N1X, N1Y,
test_size=0.2, random_state=10, stratify=N1Y)
narrow1= RandomForestClassifier(n_estimators=2500, bootstrap=True,
warm_start=True, max_features="log2")
narrow1.fit(n1x_train, n1y_train)
predictionN1 = narrow1.predict(n1x_test)
accuracyN1 = f1_score(n1y_test, predictionN1, average='weighted')
print(f"Narrow_subsystem 1 =", accuracyN1)

#broad subsystem level 2
B2X=B2data.drop('classifier', axis = 1)
```

```
B2Y=B2data['classifier']
over2 = SMOTE()
O2X, O2Y = over2.fit_resample(B2X, B2Y)
b2x_train, b2x_test, b2y_train, b2y_test = train_test_split(O2X, O2Y,
test_size=0.2, random_state=11, stratify=O2Y)
broad2 = RandomForestClassifier(n_estimators=2500, bootstrap=True,
warm_start=True, max_features="log2")
broad2.fit(b2x_train, b2y_train)
predictionB2 = broad2.predict(b2x_test)
accuracyB2 = f1_score(b2y_test, predictionB2, average='weighted')
print(f"Broad_subsytem 2 =", accuracyB2)

#narrow subsystem level 2
N2X=N2data.drop('classifier', axis = 1)
N2Y=N2data['classifier']
n2x_train, n2x_test, n2y_train, n2y_test = train_test_split(N2X, N2Y,
test_size=0.2, random_state=12, stratify=N2Y)
narrow2 = RandomForestClassifier(n_estimators=2500, bootstrap=True,
warm_start=True, max_features="log2")
narrow2.fit(n2x_train, n2y_train)
predictionN2 = narrow2.predict(n2x_test)
accuracyN2 = f1_score(n2y_test, predictionN2, average='weighted')
print(f"Narrow_subsytem 2 =", accuracyN2)

#broad subsystem level 3
B3X=B3data.drop('classifier', axis = 1)
B3Y=B3data['classifier']
over3 = SMOTE()
O3X, O3Y = over3.fit_resample(B3X, B3Y)
b3x_train, b3x_test, b3y_train, b3y_test = train_test_split(O3X, O3Y,
test_size=0.2, random_state=13, stratify=O3Y)
broad3 = RandomForestClassifier(n_estimators=2500, bootstrap=True,
warm_start=True, max_features="log2")
broad3.fit(b3x_train, b3y_train)
predictionB3 = broad3.predict(b3x_test)
accuracyB3 = f1_score(b3y_test, predictionB3, average='weighted')
print(f"Broad_subsystem 3 =", accuracyB3)

#narrow subsystem level 3
N3X=N3data.drop('classifier', axis = 1)
N3Y=N3data['classifier']
n3x_train, n3x_test, n3y_train, n3y_test = train_test_split(N3X, N3Y,
test_size=0.2, random_state=14, stratify=N3Y)
narrow3 = RandomForestClassifier(n_estimators=2500, bootstrap=True,
warm_start=True, max_features="log2")
narrow3.fit(n3x_train, n3y_train)
predictionN3 = narrow3.predict(n3x_test)
accuracyN3 = f1_score(n3y_test, predictionN3, average='weighted')
print(f"Narrow_subsystem3 =", accuracyN3)
```

```
#precision
print("precision")
precisionNG = precision_score(ngy_test, NGprediction, average='weighted')
print(f"Narrow_genus =", precisionNG)
precisionBG = precision_score(bgy_test, BGprediction, average='weighted')
print(f"broad_genus =", precisionBG)
precisionNF = precision_score(nfy_test, NFprediction, average='weighted')
print(f"Narrow_family =", precisionNF)
precisionBF = precision_score(bfy_test, BFprediction, average='weighted')
print(f"broad_family =", precisionBF)
precisionNS = precision_score(nsy_test, NSprediction, average='weighted')
print(f"Narrow_species =", precisionNS)
precisionBS = precision_score(bsy_test, BSprediction, average='weighted')
print(f"broad_species =", precisionBS)
precisionNT = precision_score(nty_test, predictionT, average='weighted')
print(f"Narrow_strain =", precisionNT)
precisionBT = precision_score(bty_test, predictionA, average='weighted')
print(f"broad_strain =", precisionBT)
precisionN1 = precision_score(n1y_test, predictionN1, average='weighted')
print(f"Narrow_subsystem1 =", precisionN1)
precisionB1 = precision_score(b1y_test, predictionB1, average='weighted')
print(f"broad_subsystem1 =", precisionB1)
precisionN2 = precision_score(n2y_test, predictionN2, average='weighted')
print(f"Narrow_subsystem2 =", precisionN2)
precisionB2 = precision_score(b2y_test, predictionB2, average='weighted')
print(f"broad_subsystem2 =", precisionB2)
precisionN3 = precision_score(n3y_test, predictionN3, average='weighted')
print(f"Narrow_subsystem3 =", precisionN3)
precisionB3 = precision_score(b3y_test, predictionB3, average='weighted')
print(f"broad_subsystem3 =", precisionB3)

#recall
print(" ")
print('---')
print(" ")
print("recall")
recallNG = recall_score(ngy_test, NGprediction, average='weighted')
print(f"Narrow_genus =", recallNG)
recallBG = recall_score(bgy_test, BGprediction, average='weighted')
print(f"broad_genus =", recallBG)
recallNF = recall_score(nfy_test, NFprediction, average='weighted')
print(f"Narrow_family =", recallNF)
recallBF = recall_score(bfy_test, BFprediction, average='weighted')
print(f"broad_family =", recallBF)
recallNS = recall_score(nsy_test, NSprediction, average='weighted')
print(f"Narrow_species =", recallNS)
recallBS = recall_score(bsy_test, BSprediction, average='weighted')
print(f"broad_species =", recallBS)
```

```
recallNT = recall_score(nty_test, predictionT, average='weighted')
print(f"Narrow_strain =", recallNT)
recallBT = recall_score(bty_test, predictionA, average='weighted')
print(f"broad_strain =", recallBT)
recallN1 = recall_score(n1y_test, predictionN1, average='weighted')
print(f"Narrow_subsystem1 =", recallN1)
recallB1 = recall_score(b1y_test, predictionB1, average='weighted')
print(f"broad_subsystem1 =", recallB1)
recallN2 = recall_score(n2y_test, predictionN2, average='weighted')
print(f"Narrow_subsystem2 =", recallN2)
recallB2 = recall_score(b2y_test, predictionB2, average='weighted')
print(f"broad_subsystem2 =", recallB2)
recallN3 = recall_score(n3y_test, predictionN3, average='weighted')
print(f"Narrow_subsystem3 =", recallN3)
recallB3 = recall_score(b3y_test, predictionB3, average='weighted')
print(f"broad_subsystem3 =", recallB3)

#f1
print(" ")
print("---")
print(" ")
print("f1")
f1NG = f1_score(ngy_test, NGprediction, average='weighted')
print(f"Narrow_genus =", f1NG)
f1BG = f1_score(bgy_test, BGprediction, average='weighted')
print(f"broad_genus =", f1BG)
f1NF = f1_score(nfy_test, NFprediction, average='weighted')
print(f"Narrow_family =", f1NF)
f1BF=f1_score(bfy_test, BFprediction, average='weighted')
print(f"broad_family =", f1BF)
f1NS = f1_score(nsy_test, NSprediction, average='weighted')
print(f"Narrow_species =", f1NS)
f1BS = f1_score(bsy_test, BSprediction, average='weighted')
print(f"broad_species =", f1BS)
f1NT = f1_score(nty_test, predictionT, average='weighted')
print(f"Narrow_strain =", f1NT)
f1BT = f1_score(bty_test, predictionA, average='weighted')
print(f"broad_strain =", f1BT)
f1N1 = f1_score(n1y_test, predictionN1, average='weighted')
print(f"Narrow_subsystem1 =", f1N1)
f1B1 = f1_score(b1y_test, predictionB1, average='weighted')
print(f"broad_subsystem1 =", f1B1)
f1N2 = f1_score(n2y_test, predictionN2, average='weighted')
print(f"Narrow_subsystem2 =", f1N2)
f1B2 = f1_score(b2y_test, predictionB2, average='weighted')
print(f"broad_subsystem2 =", f1B2)
f1N3 = f1_score(n3y_test, predictionN3, average='weighted')
print(f"Narrow_subsystem3 =", f1N3)
f1B3 = f1_score(b3y_test, predictionB3, average='weighted')
```

```
print(f"broad_subsystem3 =", f1B3)

Yp=[precisionNG, precisionBG, precisionNF, precisionBF, precisionNS,
precisionBS, precisionNT, precisionBT, precisionN1, precisionB1, precisionN2,
precisionB2, precisionN3, precisionB3]
Ya=[NGaccuracy, BGaccuracy, NFaccuracy, BFaccuracy, NSaccuracy, BSaccuracy,
accuracyT, accuracyA, accuracyN1, accuracyB1, accuracyN2, accuracyB2,
accuracyN3, accuracyB3]
Yr=[recallNG, recallBG, recallNF, recallBF, recallNS, recallBS, recallNT,
recallBT, recallN1, recallB1, recallN2, recallB2, recallN3, recallB3]
Yf=[f1NG, f1BG, f1NF, f1BF, f1NS, f1BS, f1NT, f1BT, f1N1, f1B1, f1N2, f1B2,
f1N3, f1B3]

Yy = Yf + Ya + Yp + Yr
space = [0.01, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
Xo = ('narrow_genus', 'broad_genus', 'narrow_family', 'broad_family',
'narrow_species', 'broad_species', 'narrow_strain', 'broad_strain',
'narrow_subsystem 1', 'broad_subsystem 1', 'narrow_subsystem 2',
'broad_subsystem 2', 'narrow_susbystem 3', 'broad_subsystem 3')
Yo = {'Precision': Yp, 'Recall': Yr, 'Accuracy': Ya, 'F1': Yf}
w = 0.2
m = 0
x = np.arange(len(Xo))


fig, ax = plt.subplots(layout='constrained')
for title, value in Yo.items():
  offset = w * m
  rects = ax.bar(x + offset, value, w, label=title)
  #ax.bar_label(rects, padding=3, size=10)
  m += 1

#plt.bar(Xo, Yo)

plt.xticks(x + 0.25, Xo, rotation=90, size=10)
plt.xlabel('model')
#plt.set_color_cycle(['blue', 'black', 'red', 'yellow'])
#plt.legend(loc='lower left')
plt.ylabel('score')
plt.show()

#trying to make three figures, one for recall, precision and f1, with score
on the y axis and model/site on the x axis

NSdict= classification_report(nsy_test, NSprediction, output_dict=True)
BSdict= classification_report(bsy_test, BSprediction, output_dict=True)

NGdict= classification_report(ngy_test, NGprediction, output_dict=True)
BGdict= classification_report(bgy_test, BGprediction, output_dict=True)
```

```
NFdict= classification_report(nfy_test, NFprediction, output_dict=True)
BFdict= classification_report(bfy_test, BFprediction, output_dict=True)

NTdict= classification_report(nty_test, predictionT, output_dict=True)
BTdict= classification_report(bty_test, predictionA, output_dict=True)

N1dict= classification_report(n1y_test, predictionN1, output_dict=True)
B1dict= classification_report(b1y_test, predictionB1, output_dict=True)

N2dict= classification_report(n2y_test, predictionN2, output_dict=True)
B2dict= classification_report(b2y_test, predictionB2, output_dict=True)

N3dict= classification_report(n3y_test, predictionN3, output_dict=True)
B3dict= classification_report(b3y_test, predictionB3, output_dict=True)

DOOP = ['BSdict', 'BGdict', 'BFdict', 'BTdict', 'B1dict', 'B2dict', 'B3dict',
]
NOOP = ['NSdict', 'NGdict', 'NFdict', 'NTdict', 'N1dict', 'N2dict', 'N3dict']
#dict looks like (A: (b: c), A: (b: c))
#dict shows A= category, b= score measurement, c= score value, DOOP= model
name
xax = [] #A_DOOP
w = 0.1
o = ['animal', 'built environemnt', 'freshwater', 'human gastrointestinal',
'human oral', 'human other', 'human respiratory', 'human skin', 'marine',
'plant', 'soil', 'wastewater']
no = ['animal_arthropod', 'animal_bird', 'animal_mammal', 'animal_other',
'builtenvironment', 'freshwater_', 'freshwater_lake', 'freshwater_sediment',
'human_other', 'human_skin', 'humangastrointestinal_miscgut',
'humangastrointestinal_stool', 'humanoral', 'humanrespiratory_lung',
'humanrespiratory_sputum', 'marine_', 'marine_coastal', 'marine_sediment',
'plant_', 'plant_rhizosphere', 'soil_', 'soil_agricultural',
'soil_rhizosphere', 'wastewater_']
m = 0

BSdictp=[]
BGdictp=[]
BFdictp=[]
BTdictp=[]
B1dictp=[]
B2dictp=[]
B3dictp=[]

BSdictr=[]
BGdictr=[]
BFdictr=[]
BTdictr=[]
B1dictr=[]
```

```
B2dictr=[]
B3dictr=[]

BSdictf=[]
BGdictf=[]
BFdictf=[]
BTdictf=[]
B1dictf=[]
B2dictf=[]
B3dictf=[]

NSdictp=[]
NGdictp=[]
NFdictp=[]
NTdictp=[]
N1dictp=[]
N2dictp=[]
N3dictp=[]

NSdictr=[]
NGdictr=[]
NFdictr=[]
NTdictr=[]
N1dictr=[]
N2dictr=[]
N3dictr=[]

NSdictf=[]
NGdictf=[]
NFdictf=[]
NTdictf=[]
N1dictf=[]
N2dictf=[]
N3dictf=[]

for thing in DOOP:
  #print(thing)

  drops = ['accuracy', 'macro avg', 'weighted avg', 'support']
  thingo = getattr(sys.modules[__name__], (thing))
  x = np.arange(len(o))
  for a, d in thingo.items():
    #print(a)
    #print(d)
    if a in drops:
      pass
    else:
 #     fig, ax = plt.subplots(layout='constrained')
      for title, value in d.items():
        if title in drops:
          pass
        elif title == 'precision':
          thingo = getattr(sys.modules[__name__], (thing + f'p'))
          thingo.append(value)
        elif title == 'recall':
          thingo = getattr(sys.modules[__name__], (thing + f'r'))
          thingo.append(value)
        elif title == 'f1-score':
          thingo = getattr(sys.modules[__name__], (thing + f'f'))
          thingo.append(value)
for thing in NOOP:
  #print(thing)

  drops = ['accuracy', 'macro avg', 'weighted avg', 'support']
  thingo = getattr(sys.modules[__name__], (thing))
  x = np.arange(len(o))
  for a, d in thingo.items():
    #print(a)
    #print(d)
    if a in drops:
      pass
    else:
 #     fig, ax = plt.subplots(layout='constrained')
      for title, value in d.items():
        if title in drops:
          pass
        elif title == 'precision':
          thingo = getattr(sys.modules[__name__], (thing + f'p'))
          thingo.append(value)
        elif title == 'recall':
          thingo = getattr(sys.modules[__name__], (thing + f'r'))
          thingo.append(value)
        elif title == 'f1-score':
          thingo = getattr(sys.modules[__name__], (thing + f'f'))
          thingo.append(value)

LOOP= {'precision': 'p', 'recall': 'r', 'f1-score': 'f'}
  #Xo={'narrow_genus': , 'broad_genus', 'narrow_family', 'broad_family',
'narrow_species', 'broad_species', 'narrow_strain', 'broad_strain',
'narrow_subsystem 1', 'broad_subsystem 1', 'narrow_subsystem 2',
'broad_subsystem 2', 'narrow_susbystem 3', 'broad_subsystem 3'}
Bp={'broad_species': BSdictp, 'broad_genus': BGdictp, 'broad_family':
BFdictp, 'broad_strain': BTdictp, 'broad_subsystem 1': B1dictp,
'broad_subsystem 2': B2dictp, 'broad_subsystem 3': B3dictp}
Br={'broad_species': BSdictr, 'broad_genus': BGdictr, 'broad_family':
BFdictr, 'broad_strain': BTdictr, 'broad_subsystem 1': B1dictr,
'broad_subsystem 2': B2dictr, 'broad_subsystem 3': B3dictr}
```

```
Bf={'broad_species': BSdictf, 'broad_genus': BGdictf, 'broad_family':
BFdictf, 'broad_strain': BTdictf, 'broad_subsystem 1': B1dictf,
'broad_subsystem 2': B2dictf, 'broad_subsystem 3': B3dictf}
Np={'narrow_species': NSdictp, 'narrow_genus': NGdictp, 'narrow_family':
NFdictp, 'narrow_strain': NTdictp, 'narrow_subsystem 1': N1dictp,
'narrow_subsystem 2': N2dictp, 'narrow_subsystem 3': N3dictp}
Nr={'narrow_species': NSdictr, 'narrow_genus': NGdictr, 'narrow_family':
NFdictr, 'narrow_strain': NTdictr, 'narrow_subsystem 1': N1dictr,
'narrow_subsystem 2': N2dictr, 'narrow_subsystem 3': N3dictr}
Nf={'narrow_species': NSdictf, 'narrow_genus': NGdictf, 'narrow_family':
NFdictf, 'narrow_strain': NTdictf, 'narrow_subsystem 1': N1dictf,
'narrow_subsystem 2': N2dictf, 'narrow_subsystem 3': N3dictf}


x = np.arange(len(o))
xo = np.arange(len(no))

#plt.figure(figsize=(16,7))
for score, splot in LOOP.items():
  m=0
  sploto = getattr(sys.modules[__name__], (f'B' + splot))
  #plt.figure(figsize=(16,7))
  fig, ax = plt.subplots(layout='constrained', figsize=(16,10))
  #plt.figure(figsize=(16,7))
  for t, v in sploto.items():
    offset = w * m
    rects = ax.bar(x + offset, v, w, label=t)
    ax.bar_label(rects, padding=3, size=1)
    m += 1
  print(score)
  plt.xticks(x + 0.3, o, rotation=90, size=20)
  plt.yticks(size=15)
  plt.xlabel('Isolation Ennvironemnt')
  plt.xlabel('Isolation source', size=20)
#plt.set_color_cycle(['blue', 'black', 'red', 'yellow'])
#  plt.legend(loc="lower left", bbox_to_anchor=(1.05, 0.0), fontsize="small")
  plt.ylabel(score, size=20)
  plt.show()

mn = 0
wn = 0.1
for score, splot in LOOP.items():
  mn = 0
  wn = 0.1
  sploto = getattr(sys.modules[__name__], (f'N' + splot))
  fig, ax = plt.subplots(layout='constrained', figsize=(16,10))
  for t, v in sploto.items():
    if len(xo) != len(v):
      print(f"shape mismatch: xo with shape {xo.shape}, v with shape
{np.array(v).shape}, skipping {t}")
      continue  # Skip to the next category
    print(v)
    offset = wn * mn
    rects = ax.bar(xo + offset, v, wn, label=t)
    ax.bar_label(rects, padding=3, size=0.1)
    mn += 1
  print(score)
  plt.xticks(xo + 0.3, no, rotation=90, size=20)
  plt.yticks(size=15)
  plt.ylabel(score, size=20)
  plt.xlabel('Isolation source', size=20)
#plt.set_color_cycle(['blue', 'black', 'red', 'yellow'])
 # plt.legend(loc="lower left", bbox_to_anchor=(1.05, 0.0), fontsize="small")
  plt.xlabel('Isolation Environment')
  plt.show()

narrow_predictionF = BNF.predict(x_testF)
narrow_f13 = f1_score(N_class1, narrow_predictionF, average='weighted')
print(f"NARROW_FAMILY SRA data ", narrow_f13)

narrow_predictionG = BNG.predict(x_testG)
narrow_f1G = f1_score(N_class1, narrow_predictionG, average='weighted')
print(f"NARROW_GENUS SRA data ", narrow_f1G)

featuresN = NSX.columns
missingN = set(featuresN) - set(dataS.columns)
for col in missingN:
  dataS[col] = 0
x_testSO = dataS[featuresN]
narrow_predictionS = BNS.predict(x_testSO)
narrow_f1S = f1_score(N_class1, narrow_predictionS, average='weighted')
print(f"NARROW_SPECIES SRA data ", narrow_f1S)

narrow_predictionT = narrowT.predict(x_testT)
narrow_f1T = f1_score(N_classT, narrow_predictionT, average='weighted')
print(f"NARROW_STRAIN SRA data ", narrow_f1T)

narrow_prediction1 = narrow1.predict(x_test1)
narrow_f11 = f1_score(N_class1, narrow_prediction1, average='weighted')
print(f"NARROW_1 SRA data ", narrow_f11)

narrow_prediction2 = narrow2.predict(x_test2)
narrow_f12 = f1_score(N_class1, narrow_prediction2, average='weighted')
print(f"NARROW_2 SRA data ", narrow_f12)

narrow_prediction3 = narrow3.predict(x_test3)
narrow_f13 = f1_score(N_class1, narrow_prediction3, average='weighted')
```

```
print(f"NARROW_3 SRA data ", narrow_f13)


#mgnify new data
famdata = pd.read_excel('/content/drive/My Drive/newsetfam.xlsx', sheet_name=
'f')
gendata = pd.read_excel('/content/drive/My Drive/newsetfam.xlsx', sheet_name=
'g')
sppdata = pd.read_excel('/content/drive/My Drive/newsetfam.xlsx', sheet_name=
's')
strdata = pd.read_excel('/content/drive/My Drive/newsetfam.xlsx', sheet_name=
'st')
sb1data = pd.read_excel('/content/drive/My Drive/newsetfam.xlsx', sheet_name=
's1')
sb2data = pd.read_excel('/content/drive/My Drive/newsetfam.xlsx', sheet_name=
's2')
sb3data = pd.read_excel('/content/drive/My Drive/newsetfam.xlsx', sheet_name=
's3')

featuresF = BFX.columns
missingF = set(featuresF) - set(famdata.columns)
for col in missingF:
  famdata[col] = 0
x_testF = famdata[featuresF]
N_classF = famdata['narrow']
B_classF = famdata['broad']
Broad_predictionF = BBF.predict(x_testF)
Broad_f1F = f1_score(B_classF, Broad_predictionF, average='weighted')
print(f"BROAD_FAMILY mgnify2 data ", Broad_f1F)

featuresG = BGX.columns
missingG = set(featuresG) - set(gendata.columns)
for col in missingG:
  gendata[col] = 0
x_testG = gendata[featuresG]
N_classG = gendata['narrow']
B_classG = gendata['broad']
Broad_predictionG = BRF.predict(x_testG)
Broad_f1G = f1_score(B_classG, Broad_predictionG, average='weighted')
print(f"BROAD_GENUS mgnify2 data ", Broad_f1G)

featuresS = BSX.columns
missingS = set(featuresS) - set(sppdata.columns)
dataS = sppdata.reindex(columns=featuresS, fill_value=0)
x_testS = dataS[featuresS]
N_classS = sppdata['narrow']
B_classS = sppdata['broad']
Broad_predictionS = BBS.predict(x_testS)
Broad_f1S = f1_score(B_classS, Broad_predictionS, average='weighted')
```

```
print(f"BROAD_SPECIES mgnify2 data ", Broad_f1S)

featuresT = BTX.columns
missingT = set(featuresT) - set(strdata.columns)
dataT = strdata.reindex(columns=featuresT, fill_value=0)
x_testT = dataT[featuresT]
N_classT = strdata['narrow']
B_classT = strdata['broad']
Broad_predictionT = broadA.predict(x_testT)
Broad_f1T = f1_score(B_classT, Broad_predictionT, average='weighted')
print(f"BROAD_STRAIN mgnify2 data ", Broad_f1T)

features1 = B1X.columns
missing1 = set(features1) - set(sb1data.columns)
for col in missing1:
  sb1data[col] = 0
x_test1 = sb1data[features1]
N_class1 = sb1data['narrow']
B_class1 = sb1data['broad']
Broad_prediction1 = broad1.predict(x_test1)
Broad_f11 = f1_score(B_class1, Broad_prediction1, average='weighted')
print(f"BROAD_1 mgnify2 data ", Broad_f11)

features2 = B2X.columns
missing2 = set(features2) - set(sb2data.columns)
for col in missing2:
  sb2data[col] = 0
x_test2 = sb2data[features2]
N_clasS2 = sb1data['narrow']
B_class2 = sb2data['broad']
Broad_prediction2 = broad2.predict(x_test2)
Broad_f12 = f1_score(B_class2, Broad_prediction2, average='weighted')
print(f"BROAD_2 mgnify2 data ", Broad_f12)

features3 = B3X.columns
missing3 = set(features3) - set(sb3data.columns)
for col in missing3:
  sb3data[col] = 0
x_test3 = sb3data[features3]
N_class3 = sb3data['narrow']
B_class3 = sb3data['broad']
Broad_prediction3 = broad3.predict(x_test3)
Broad_f13 = f1_score(B_class3, Broad_prediction3, average='weighted')
print(f"BROAD_3 mgnify2 data ", Broad_f13)

#mgnify new data NARROW
famdata = pd.read_excel('/content/drive/My Drive/newsetfam.xlsx', sheet_name=
'f')
```

```
gendata = pd.read_excel('/content/drive/My Drive/newsetfam.xlsx', sheet_name=
'g')
sppdata = pd.read_excel('/content/drive/My Drive/newsetfam.xlsx', sheet_name=
's')
strdata = pd.read_excel('/content/drive/My Drive/newsetfam.xlsx', sheet_name=
'st')
sb1data = pd.read_excel('/content/drive/My Drive/newsetfam.xlsx', sheet_name=
's1')
sb2data = pd.read_excel('/content/drive/My Drive/newsetfam.xlsx', sheet_name=
's2')
sb3data = pd.read_excel('/content/drive/My Drive/newsetfam.xlsx', sheet_name=
's3')

featuresF = BFX.columns
missingF = set(featuresF) - set(famdata.columns)
for col in missingF:
  famdata[col] = 0
x_testF = famdata[featuresF]
N_classF = famdata['narrow']
B_classF = famdata['broad']
Broad_predictionF = BNF.predict(x_testF)
Broad_f1F = f1_score(N_classF, Broad_predictionF, average='weighted')
print(f"BROAD_FAMILY mgnify2 data ", Broad_f1F)

featuresG = BGX.columns
missingG = set(featuresG) - set(gendata.columns)
for col in missingG:
  gendata[col] = 0
x_testG = gendata[featuresG]
N_classG = gendata['narrow']
B_classG = gendata['broad']
Broad_predictionG = NRF.predict(x_testG)
Broad_f1G = f1_score(N_classG, Broad_predictionG, average='weighted')
print(f"BROAD_GENUS mgnify2 data ", Broad_f1G)

featuresS = BSX.columns
missingS = set(featuresS) - set(sppdata.columns)
dataS = sppdata.reindex(columns=featuresS, fill_value=0)
x_testS = dataS[featuresS]
N_classS = sppdata['narrow']
B_classS = sppdata['broad']
Broad_predictionS = BNS.predict(x_testS)
Broad_f1S = f1_score(N_classS, Broad_predictionS, average='weighted')
print(f"BROAD_SPECIES mgnify2 data ", Broad_f1S)

featuresT = BTX.columns
missingT = set(featuresT) - set(strdata.columns)
dataT = strdata.reindex(columns=featuresT, fill_value=0)
x_testT = dataT[featuresT]
```

```
N_classT = strdata['narrow']
B_classT = strdata['broad']
Broad_predictionT =  narrowT.predict(x_testT)
Broad_f1T = f1_score(N_classT, Broad_predictionT, average='weighted')
print(f"BROAD_STRAIN mgnify2 data ", Broad_f1T)

features1 = B1X.columns
missing1 = set(features1) - set(sb1data.columns)
for col in missing1:
  sb1data[col] = 0
x_test1 = sb1data[features1]
N_class1 = sb1data['narrow']
B_class1 = sb1data['broad']
Broad_prediction1 = narrow1.predict(x_test1)
Broad_f11 = f1_score(N_class1, Broad_prediction1, average='weighted')
print(f"BROAD_1 mgnify2 data ", Broad_f11)

features2 = B2X.columns
missing2 = set(features2) - set(sb2data.columns)
for col in missing2:
  sb2data[col] = 0
x_test2 = sb2data[features2]
N_clasS2 = sb1data['narrow']
B_class2 = sb2data['broad']
Broad_prediction2 = narrow2.predict(x_test2)
Broad_f12 = f1_score(N_clasS2, Broad_prediction2, average='weighted')
print(f"BROAD_2 mgnify2 data ", Broad_f12)

features3 = B3X.columns
missing3 = set(features3) - set(sb3data.columns)
for col in missing3:
  sb3data[col] = 0
x_test3 = sb3data[features3]
N_class3 = sb3data['narrow']
B_class3 = sb3data['broad']
Broad_prediction3 = narrow3.predict(x_test3)
Broad_f13 = f1_score(N_class3, Broad_prediction3, average='weighted')
print(f"BROAD_3 mgnify2 data ", Broad_f13)

#prediction probability
Bforests = {'famdata': ['BBF'], 'gendata': ['BRF'], 'sppdata': ['BBS'],
'strdata': ['broadA'], 'sb1data': ['broad1'], 'sb2data': ['broad2'],
'sb3data': ['broad3']}
forests = {'famdata': ['BNF'], 'gendata': ['BNG'], 'sppdata': ['BNS'],
'strdata': ['narrowT'], 'sb1data': ['narrow1'], 'sb2data': ['narrow2'],
'sb3data': ['narrow3']}

for data, forest in forests.items():
  #d = getattr(sys.modules[__name__], (data))
  d = eval(data)
  title = 'sample'
  f = getattr(sys.modules[__name__], forest[0])
 # f= forest
#x = data.drop(title, axis = 1)
  ind = d[title]
  features = f.feature_names_in_
  x = d.reindex(columns=features, fill_value=0)
  cls = f.classes_

  #print(cls)
  proba = f.predict_proba(x)
  #predict = forest.predict(x)


  df = pd.DataFrame(proba, index=ind, columns=cls)
  #print(df)
  framename= data + '.csv'
  df.to_csv(framename, index=True, header=True, sep=',')
```

Comparing Random Forest, Logistic regression and XGboost models

```
BAdata = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'broad_all')
NAdata = pd.read_excel('/content/drive/My Drive/focus results.xlsx',
sheet_name= 'narrow_all')
#NX=NAdata.drop('classifier', axis = 1)
NY=NAdata['classifier']
BY=BAdata['classifier']
dic = {r'\[':'_', r'\]':'_', r'<':'_'}

for a, b in dic.items():
  NAdata.columns = NAdata.columns.str.replace(a, b, regex=True)
  BAdata.columns = BAdata.columns.str.replace(a, b, regex=True)

NX=NAdata.drop('classifier', axis = 1)
BX=BAdata.drop('classifier', axis = 1)
# broad fam
BFX=BAdata.drop('classifier', axis = 1)
BFY=BAdata['classifier']
overF = SMOTE()
OFX, OFY = overF.fit_resample(BFX, BFY)
bfx_train, bfx_test, bfy_train, bfy_test = train_test_split(OFX, OFY,
test_size=0.2, stratify=OFY)
BBF = RandomForestClassifier(n_estimators=2500, bootstrap=True,
max_features="sqrt")
BBF.fit(bfx_train, bfy_train)
BFprediction = BBF.predict(bfx_test)
BroadRF = f1_score(bfy_test, BFprediction, average='weighted')
```

```
print(f"Narrow_family =", BroadRF)

# narrow fam
NFX=NAdata.drop('classifier', axis = 1)
NFY=NAdata['classifier']
nfx_train, nfx_test, nfy_train, nfy_test = train_test_split(NFX, NFY,
test_size=0.2, stratify=NFY)
BNF = RandomForestClassifier(n_estimators=2500, bootstrap=True,
max_features="sqrt")
BNF.fit(nfx_train, nfy_train)
NFprediction = BNF.predict(nfx_test)
narrowRF = f1_score(nfy_test, NFprediction, average='weighted')
print(f"Narrow_family =", narrowRF)
```

```
#xgboost model
label_binarizer = LabelBinarizer().fit(NY)
nbin = label_binarizer.transform(NY)
nx_train, nx_test, ny_train, ny_test = train_test_split(NX, nbin,
test_size=0.2, stratify=nbin)

label_binarizer = LabelBinarizer().fit(BY)
bbin = label_binarizer.transform(BY)
bx_train, bx_test, by_train, by_test = train_test_split(BX, bbin,
test_size=0.2, stratify=bbin)

overb = SMOTE()
OX, OY = overb.fit_resample(BX, BY)
label_binarizer = LabelBinarizer().fit(OY)
obin = label_binarizer.transform(OY)
ox_train, ox_test, oy_train, oy_test = train_test_split(OX, obin,
test_size=0.2, stratify=obin)

Bxgb = XGBClassifier()
Bxg = Bxgb.fit(bx_train, by_train)
bxped= Bxg.predict(bx_test)
Bxf1 = f1_score(by_test, bxped, average='weighted')
print(f"Broad_plain =", Bxf1)

oxgb = XGBClassifier()
oxg = oxgb.fit(ox_train, oy_train)
oxped= oxg.predict(ox_test)
oxf1 = f1_score(oy_test, oxped, average='weighted')
print(f"Broad_smote =", oxf1)

Nxgb = XGBClassifier()
Nxg = Nxgb.fit(nx_train, ny_train)
nxped= Nxg.predict(nx_test)
Nxf1 = f1_score(ny_test, nxped, average='weighted')
```

```
print(f"Narrow =", Nxf1)
```

```
# Logistic Regression with LabelEncoder
label_encoder = LabelEncoder()

# Broad data
by_train_encoded = label_encoder.fit_transform(OY)
bx_train_log, bx_test_log, by_train_log, by_test_log = train_test_split(OX,
by_train_encoded, test_size=0.2, stratify=by_train_encoded)

BAlog = LogisticRegression(max_iter=3000).fit(bx_train_log, by_train_log)
Blogprediction = BAlog.predict(bx_test_log)
Bf1 = f1_score(by_test_log, Blogprediction, average='weighted')
print(f"Broad_LOG =", Bf1)

# Narrow data
ny_train_encoded = label_encoder.fit_transform(NY)
nx_train_log, nx_test_log, ny_train_log, ny_test_log = train_test_split(NX,
ny_train_encoded, test_size=0.2, stratify=ny_train_encoded)

NAlog = LogisticRegression(max_iter=3000).fit(nx_train_log, ny_train_log)
Nlogprediction = NAlog.predict(nx_test_log)
Nf1 = f1_score(ny_test_log, Nlogprediction, average='weighted')
print(f"Narrow_LOG =", Nf1)
```