Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

Jenny Merkesvik

Towards genotype—phenotype association

Leveraging multiple-source microbial data and genome annotations to infer trait attributes for the uncultured microbiome

Master's thesis in Biotechnology
Supervisor: Daniel Machado
Co-supervisor: Miguel Teixeira
May 2022

Jenny Merkesvik

NTNU
Kunnskap for en bedre verden

Jenny Merkesvik

Towards genotype—phenotype association

Leveraging multiple-source microbial data and genome annotations to infer trait attributes for the uncultured microbiome

Master's thesis in Biotechnology
Supervisor: Daniel Machado
Co-supervisor: Miguel Teixeira
May 2022

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

NTNU
Kunnskap for en bedre verden

Preface

This Master thesis is conducted under the Department of Biotechnology and Food Science at the Faculty of Natural Sciences, of the Norwegian University of Science and Technology.

It represents my last efforts as a master student of biotechnology.

At the close of this chapter in my life, I catch myself thinking back to all the experiences I have had and how they've culminated in where I am now. I've always been fascinated by biology, ranging from the smallest of cells to animals and whole ecosystems. It was not until I started my third year of biotechnology that I think I realised why. The drive for learning about the constituents of the living world; that I've had since I was old enough to follow along with the natural sciences TV programmes, was mostly an underlying curiosity for how it all works together. The complex interactions that add one and one up to three; how you think you understand how something works, only to discover that the next chapter is double the size; and the organisation of all of this into a neat system that cannot wait to be expanded upon. After all, there is perhaps no wonder how I ended up with a specialisation is systems biology.

It has been a bittersweet journey. The numerous courses and their comprehensive curriculums, assignments, and exams; the knowledge and insight shared by lecturers and teaching assistants; and the support, understanding, and memories with friends, colleagues, and family. These years have been full of joy and laughter, but also the most challenges I've ever faced. But despite the hard times, the long hours where nothing seemed to work as intended, I cannot help to think that I wouldn't be without any of it, for it has all contributed to where I currently am. In some ways, it feels like it all has come down to this thesis. The most time and effort I've spent on any project, perhaps with the exception of the continued hunt for a full shiny Pokédex.

There are several people to which I owe thanks for being able to produce this thesis. First and foremost, to Daniel Machado. You have been as helpful and supportive as a supervisor can be. Thank you for the sharing your experience and knowledge with me. Thank you for the presence and patience you've shown throughout the past year and a half; especially in the past few months leading up to the completion of the project; and especially especially in the past few weeks by providing feedback to this written work. I must also extend my gratitude to Miguel Teixeira, who have helped me tremendously with the second part of the project, and by providing feedback to this written work. I have learned much from you both, and I appreciate the time and effort you've spent on me and this thesis. I regret to inform you that I probably will not swap away from my current operative system anytime soon, however I will remember and heed by most of your other advice for a long time.

It is with great pride that I present this finished thesis. I realise now that this is just the beginning of my journey, and I am thrilled to say that I cannot wait to see how it unfolds.

Jenny Merkesvik
Trondheim, 16.05.2022

Abstract

Microorganisms have been studied extensively since the microscope was invented. Humans depend on microbes, whether we choose to utilise them in fundamental research, industry, and food production; or whether they are an unavoidable part of our lives, for instance through health and illness. Due to their relevance, there are numerous descriptions of how these small entities work. However, there are many missing pieces within these observations. Much may be due to a long-standing issue preventing us from being able to growth most of the world's known microbes. Without their growth in the laboratory, our study of them is strictly limited. This was the case until genome sequencing was developed.

Today, unfathomable amounts of data are generated on the building blocks of the genetic material of hundreds of thousands of organisms. The aim is to interpret the genetic code to predict which phenotypic features each organism will express. This may facilitate the cultivation of microbes by suggesting its preferred nutrients and conditions. Additionally, it can be used to indicate potential applications of the microorganisms. Such genotype—phenotype association is however highly complicated due to the many levels of regulation and interactions happening in all cells. To recreate all these using models and mathematical equations is currently beyond our reach. In the meantime, we may use the knowledge we have on microbes that have been cultivated successfully to search for patterns between their genotypes and phenotypes. If a particular sequence is found often in organisms which all have one feature in common, it may imply that this sequence is related to the feature. If we then find this same sequence in the genome of another organism which we do not know as much about, perhaps it will express the same feature as the other organisms did.

The aim of this thesis is to demonstrate how existing microbial data can be leveraged to predict the features of an organism based on associations made in previous observations, and the new organism's genotype. Microbial data from ten sources is thereby assembled into a standardised trait dataset. It consists of 146,767 rows which cover about 126,000 different microbes on the strain-level, which all have information on at least one of 17 included traits. Examples are substrate, oxygen requirement, antibiotic resistance, and gram staining. Further, the genomes of a selection of organisms were functionally annotated. This marked sequences of interest though gene ontology and orthology. Genotype—phenotype association was conducted by assessing the relation between the annotations and the known outcome of the trait of gram staining. The association was quantified through Fisher's exact tests. Any association with odds ratio over 10 and p-value less than 0.001 were considered significant. This means that with a given gram staining, the odds are 10 times higher that a particular annotation exists in the genome of the organism; and that maximum 1% of the instances may erroneously assume association.

With these requirements, 4,444 annotation terms were found associated with a particular gram attribute. 2,974 terms were associated with gram-negativity, and 1,470 terms with gram-positivity. Of these, 1,562 and 159 terms were found exclusively for organisms with gram-negative and gram-positive staining, respectively. Several terms were found to represent characteristic features of each of the two cell types. These terms were tracked in the annotated genomes of three random organisms without a registered gram stain attribute in the assembled dataset. For all organisms, the correct gram staining was assumed based on the identified associations. Thus the thesis aim was achieved: the conducted methodology has demonstrated how existing data may be considered and used to form hypotheses for a broad group of microorganisms, and that the association may be applied to infer missing phenotypic observations for other microorganisms.

Sammendrag

Mikroorganismer har blitt studert helt siden mikroskopet ble oppfunnet. Mikrober er svært viktige for mennesket, enten vi velger å bruke dem innen grunnforskning, industri og matproduksjon, eller om de er en obligatorisk del av livene våre, eksempelvis gjennom helse og sykdom. Som følge av deres relevans finnes det utallige beskrivelser av hvordan de små skapningene fungerer. Det er derimot store mangler blant disse observasjonene. Mye skyldes et langvarig problem som fører til at vi kun klarer å dyrke en brøkdel av de oppdagede mikrobene. Uten vekst i laboratoriet har vi svært begrensede muligheter til å studere dem. Dette var i alle fall tilfelle frem til gensekvensering ble utviklet og utbredt.

I dag genereres det ufattelige store mengder data som gjengir byggesteinene i arvematerialet til hundretusenvis av organismer. Målet er å tolke den genetiske koden for å forutse hvilke fenotypiske egenskaper enhver organisme vil uttrykke. Det kan gjøre dyrking av mikrober lettere ved at vi kan anta hva slags næring og omgivelser de trives best i. I tillegg kan det brukes til å indikere potensielle anvendelser mikrobene kan ha. Slik genotype—fenotype-assosiasjon er derimot svært komplisert som følge av de mange nivåene av regulering og interaksjoner som skjer i alle celler. Å gjenskape alle disse med modeller og matematiske uttrykk er foreløpig utenfor vår rekkevidde. I mellomtiden kan vi bruke kunnskapen vi har om mikrober vi har klart å dyrke til å lete etter mønstre mellom deres genotyper og fenotyper. Dersom en spesifikk sekvens ofte blir funnet i organismer som alle har en egenskap til felles, kan det tyde på at sekvensen har noe med egenskapen å gjøre. Om vi så finner denne sekvensen i genomet til en organisme vi ikke vet like mye om, kan det antyde at den vil utrykke den samme egenskapen som de andre organismene.

Målet med oppgaven er å demonstrere hvordan eksisterende mikrobielle data kan brukes for å forutse organismers egenskaper basert på mønstre i tidligere observasjoner, og de nye organismenes genotype. Mikrobielle data fra ti ulike kilder er dermed blitt satt sammen til et standardisert datasett. Det består av 146,767 rader som dekker omtrent 126,000 ulike mikrober på stamme-nivå, og som alle har informasjon om minst én av 17 inkluderte egenskaper. Eksempler er substrat, oksygenkrav, antibiotikaresistens og gram-farging. Videre ble genomene til et utvalg organismer funksjonelt annotert. Dette markerte sekvenser av interesser ved hjelp av gen-ontologi og -ortologi. Genotype—fenotype-assosiasjon ble utført ved å vurdere sammenhengen mellom de identifiserte annotasjonene og det kjente utfallet av en bestemt egenskap: gram-farging. Assosiasjonen ble tallfestet gjennom Fishers eksakte tester. Enhver assosiasjon med odds-ratio over 10 og p-verdi under 0.01 ble ansett som signifikant. Dette tilsier at om en gitt gram-farging er kjent, så er oddsen 10 ganger høyere for at en spesiell annotasjon finnes i genomet til organismen, og at det kun aksepteres feilaktig antakelse av assosiasjon i maksimum 1% av tilfellene.

Med disse kravene ble 4.444 annotasjonstermer funnet assosiert med en spesiell attributt av gram-farging. 2.974 termer var assosiert med gram-negative organismer, og 1.470 termer med gram-positive. Av disse ble 1.562 og 159 termer funnet eksklusivt for organismer med henholdsvis gram-negativ og gram-positiv farging. Flere av disse termene viste seg å representere karakteristiske egenskaper for de to celletypene. Disse termene ble så etterlyst i de annoterte genomene til tre tilfeldige organismer uten registrert gram-farging i datasettet. Samtlige organismers gram-farging ble korrekt antatt basert på de identifiserte assosiasjonene. Dermed har prosjektets mål blitt oppnådd. Den gjennomførte metodikken har understreket hvordan eksisterende data kan sees i sammenheng og brukes for å forme hypoteser for en bred gruppe mikroorganismer, og at disse assosiasjonene kan anvendes for å utlede manglende fenotypiske observasjoner for andre mikroorganismer.

Contents

Preface i
Abstract iii
Sammendrag v

List of Figures ix
List of Tables xi
Abbreviations xii

1. Introduction
1.1. Motivation 1
1.2. Project aim 3

2. Background
2.1. Microorganisms in research 5
2.1.1. Terminology: traits and attributes 6
2.1.2. The great plate count anomaly 7
2.1.3. Microbial traits of interest 8
2.2. Databases 10
2.2.1. Current databases on microbial data 10
2.3. Leveraging biological big data 11
2.3.1. Sequencing and genomics 11
2.3.2. Genotype—phenotype association 12
2.3.3. Fisher's exact test 13
2.3.4. Ontologies 14

3. Methods
3.1. Microbial trait dataset 17
3.1.1. Data download and cleaning 17
3.1.2. Comparison of source datasets 26
3.1.3. Trait dataset assembly 26
3.1.4. Preparing a reduced trait dataset 27
3.2. Genome sequence annotation 28
3.2.1. Genome downloads 28
3.2.2. Functional annotation 28
3.3. Genotype—phenotype association 29

4. Results and analysis
4.1. Microbial trait dataset 33
4.1.1. Comparison of data sources 33
4.1.2. Dataset assembly 34
4.1.3. Reduced dataset for genotype—phenotype association 39
4.2. Genome sequence annotation 40
4.2.1. Genome sequence comparisons 40
4.2.2. Functional annotation 41

4. Results and analysis (continued)

4.3. Genotype—phenotype association 42
4.3.1. Clusters of orthologous genes 43
4.3.2. Gene Ontology 46
4.3.3. KEGG Orthology 47
4.3.4. Inferring gram stain attributes for new organisms 49

5. Discussion

5.1. Microbial trait dataset 54
5.1.1. Data sources are highly variable 54
5.1.2. Creating a new trait dataset was beneficial 54
5.1.3. Select data sources and traits were utilised 54
5.1.4. Comparison of datasets relies on flawed identification method 55
5.1.5. The dataset may be representative of the *known* microbial diversity ... 56
5.1.6. Strict criteria were set for inclusion in reduced trait dataset 57
5.2. Genome sequence annotation 58
5.2.1. Both GenBank and RefSeq were utilised as genome sources 59
5.2.2. EggNOG was the preferred tool for functional annotation 59
5.2.3. Annotation classes vary in coverage and diversity 60
5.3. Genotype—phenotype association 61
5.3.1. Gram staining was chosen for genotype—phenotype association 61
5.3.2. Fisher's exact tests facilitated biological interpretation 62
5.3.3. Thousands of annotation terms have significant gram associations 63
5.3.4. Claims of attribute exclusivity are most uncertain 63
5.3.5. Term generalisation yields trade-off between overview and specificity 65

6. Conclusion and outlook 69

References 73

Appendices

A: Supplementary information 85
B: Gene Ontology term generalisations 87
C: Test of annotation tools 95

List of Figures

1.1	Analogy of thesis motivation to recognising a pizza recipe	3
2.1	Differences between a gram-negative and a gram-positive cell envelope	9
2.2	Demonstration of a Gene Ontology network	14
4.1	Coverage comparison of the source datasets	33
4.2	Data category overview of the assembled dataset	34
4.3	Trait content overview of the assembled dataset	36
4.4	Dataset overlaps on three taxonomic levels	37
4.5	Dataset overlaps for nine data categories	38
4.6	Term frequencies across annotated genomes	41
4.7	Distribution of significantly associated annotation terms	43
4.8	COG terms significantly associated with gram stain	44
4.9	COG term categories significantly associated with gram stain	45
4.10	GO terms significantly associated with gram stain	46
4.11	KO terms significantly associated with gram stain	48
4.12	KEGG modules significantly associated with gram stain	49
B1	GO biological process significantly associated with gram-negativity	87
B2	GO cellular component significantly associated with gram-negativity	88
B3	GO molecular function significantly associated with gram-negativity	89
B4	GO biological process significantly associated with gram-positivity	90
B5	GO cellular component significantly associated with gram-positivity	91
B6	GO molecular function significantly associated with gram-positivity	92
C1	Comparison of annotation tools	95

List of Tables

3.1	Data standardisation scheme	17
3.2	Data fields from BacDive	18
3.3	Data fields from Bergey's	19
3.4	Data fields from Corkrey	20
3.5	Data fields from IJSEM	20
3.6	Data fields from JGI GOLD	21
3.7	Data fields from Kremer	22
3.8	Data fields from MediaDB	22
3.9	Data fields from TMD	23
3.10	Data fields from Pasteur	24
3.11	Data fields from PhyMet	25
3.12	Data fields from ProTraits	25
4.1	Comparison of GenBank and RefSeq genome assemblies	41
A1	GitHub repository overview	85

Abbreviations

BacDive	Bacterial Diversity Database
BP	biological process
CC	cellular component
COG	clusters of orthologous genes
CSV	comma-separated value
DNA	deoxyribonucleic acid
DOI	digital object identifier
eDNA	environmental deoxyribonucleic acid
FAA	FastA Amino Acid
FAPROTAX	The Functional Annotation of Prokaryotic Taxa
FDR	false discovery rate
FTP	file transfer protocol
GO	gene ontology
GPA	genotype—phenotype association
GPCA	great plate count anomaly
HGT	horizontal gene transfer
IJSEM	International Journal of Systematic and Evolutionary Microbiology
JGI GOLD	Joint Genome Institute Genomes Online Database
KEGG	Kyoto Encyclopaedia of Genes and Genomes
KO	KEGG ontology
MF	molecular function
NCBI	National Centre for Biotechnology Information
NGS	next-generation sequencing
OGT	optimal growth temperature
OR	odds ratio
PATRIC	Pathosystems Resource Integration Centre
PhyMet	Physiology and Metabolism of methanogens
PLP	pyridoxal 5’-phosphate
ProTraits	Atlas of Prokaryotic Traits
RefSeq	Reference Sequence
TMD	The Microbe Directory
VBNC	viable but not culturable

1. Introduction

During the late half of the 17th century, Robert Hooke published a book in which he described observations he had made when studying a selection of objects under a microscope. They included plants, hair, a feather, fleas, and a mouldy leather-bound book [1]. Upon inspecting the mould under the microscope, Hooke observed:

> *"(…) a very pretty shap'd Vegetative body, which* (…) *shot out multitudes of small long cylindrical and transparent stalks, not exactly streight, but a little bended with the weights of a round and white knob that grew on the top of each of them (…), but they seem'd most likely to be of the same nature with those that grow on Mushroms, which they did, some of them, not a little resemble."*
>
> [2, pp. 125-126]

This would later be known as the first published description of a microorganism [1]. In his next sentence, Hooke goes on to describe the taste and smell of the mould, which were "*active enough to make a sensible impression on those organs*" and were "*unpleasant and noisome*" [2, p. 126]. Further, the vegetative bodies did not seem to catch fire after passing them through the flame of a candle "*three or four times*" [2, p. 126]. Despite these actions seeming odd or potentially dangerous today, they capture the essence of human curiosity and how it drives scientific endeavours. Almost 400 years later, we still strive to learn more about what we see, and in the case of microorganisms; that which we can't see, in the world around us. Although, these days there are more regulations in place to protect the environment, health, and safety of involved parties.

Scientific curiosity is thriving and seeing a steady increase in the number of contributors. This is not limited to microbiology, but is also true for fields such as biotechnology, ecology, statistics, genomics, and bioinformatics. An important addition to the world since the days of Hooke has been computers and the Internet. They enable the generation, analysis, and storing of immense amounts of information which is shared among individuals, research groups, and institutions. As a result of the availability of additional data, experiences, and ideas from all over the world, science is accelerated beyond what was envisioned possible just a few decades ago. For instance, we are able to identify cancers by the increased presence of a protein [3]; genetically engineer cells to produce desired compounds [4]; and acquire the full genome sequence of a human in less than six hours [5]. The limit for what is possible to achieve given the time, effort, and resources to do so, is seemingly decreasing at a fast pace. Still, there remains central limitations and challenges within any field, which cannot be solved simply by generating more data; supposedly.

1.1. Motivation

Other than having pretty shapes, weird smell, and some resembling mushrooms, microorganisms display many interesting features. They may be viewed as small factories that can generate a variety of products, ranging from building materials to nutrients, medicines, and even new factories. Some estimate the existence of millions [6], billions [7, 8], or even a trillion [9] different microbial species on Earth. Nevertheless, the immense variety within the microbiota that is already known, attests to the great diversity and potential that remain undiscovered. Learning more about these microscopic entities is of great advantage for many disciplines. Medicine, environmental sciences, food production, and fundamental research are just a few examples.

Traditionally, microorganisms have been identified and classified on the basis of phenotypic descriptions of their capabilities. Taste and smell might not be as commonly used today,

but morphology, such as cell size, shape, and colour; substrates, aerobioses, and temperatures required for growth, and metabolic capabilities are all examples of microbial features of interest. The generation of these data may be a slow and uncertain process, however [10, 11]. A central issue is the instance where a microorganism fails to grow in a controlled environment, such as in the laboratory, which is required for their proper study. The great plate count anomaly (GPCA) is a notorious issue describing the phenomena that within a sample of cells, as few as 0.01% will form colonies when attempted grown by regular plating techniques [12]. With so many species still left to study, GPCA is a great limitation to the discovery of microbial species and their phenotypic features.

An important facilitator for the discovery and research of microorganisms is the field of genomics. It encompasses approaches that study whole or partial genome sequences [13, p. 240]. The field is relatively young but has over the course a century grown from the basic concept of a gene as the determinant of inheritable traits, into a sophisticated area of research with several subfields and associated disciplines [14]. With modern technology, the entire collection of genes present in the deoxyribonucleic acid (DNA) of an organism can be determined by reading the order of its constituent bases. Furthermore, the development of next-generation sequencing (NGS) techniques enables massive parallel procedures that can be utilised to infer many sequences at once both faster, cheaper, and more accurately [13, p. 240].

Thus the *genotype* of an organism is remarkably accessible, and genome data for a multitude of species is rapidly accumulating. The data may be used to supplement the lack of microbial phenotypic descriptions caused by the issues in cultivation. After all, the entire organism is designed as per the instructions stored within its genetic code. Genotype—phenotype association (GPA) denotes the efforts of relating the information contained in the genome of an organism to its expressed features [15]. However, we are still not sure how the myriad of instructions is expressed and regulated throughout the many levels of organisation and interaction present within a cell. Thus how to derive concrete phenotypic cues from the millions of bases within a genome remains mysterious. A good starting point for understanding the complex systems of cells however, is to refer back to what is already known: could we make sense of the already existing phenotypic descriptions of microbes and their genotypes? Could the patterns observed within these known species be applied to the unknown?

To make the concept of the present work more intuitive, I would like to present an analogy to a seemingly unrelated topic. Most are familiar with what a typical Hawaiian pizza is made of: dough, tomato sauce, ham, cheese, and pineapple. Suppose a recipe is found, but its title and image are torn off. The recipe calls for dough, tomato sauce, ham, cheese, and pineapple. Some may think that this must be a recipe for an inventive new smoothie. Most will correctly assume that it is a recipe for pizza, however with some creative liberty regarding the addition of fruit. They've seen enough recipes for pizza before to recognise the dish by its ingredients. Moving from the culinary into the present theme. Most are familiar with what a typical gram-negative bacterium is made of: a cell membrane, a thin peptidoglycan layer, a large periplasm, and an outer membrane. Suppose a genome is found, but we don't know what organism it is from, or what its characteristics are. The genome contains sequences known from other organisms to yield a cell membrane, a thin peptidoglycan layer, a large periplasm, and an outer membrane. Some may think this is the genome of an elephant, while most will recognise the constituents as characteristics of a gram-negative bacterium. They've seen enough such sequences to recognise this cell type by its genomic content.

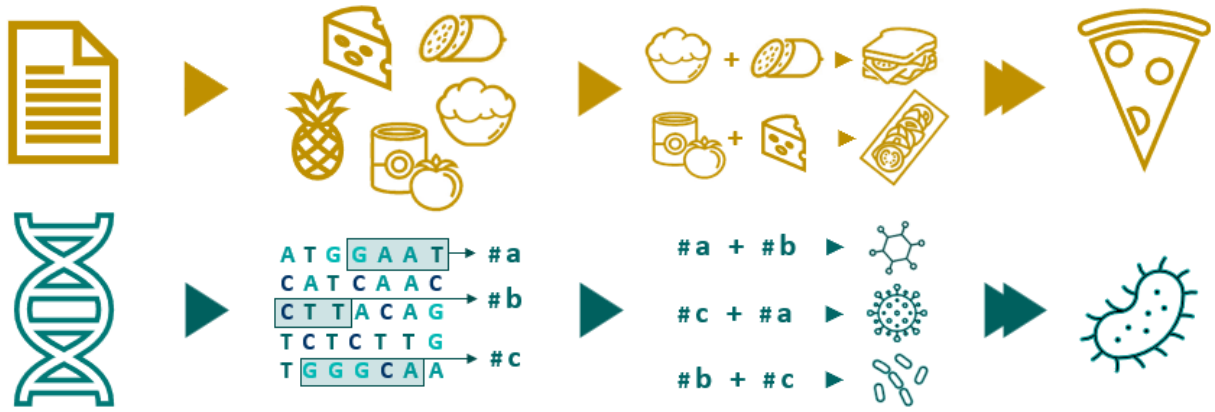


Figure 1.1: Analogy of thesis motivation to recognising a pizza recipe*: suggesting the final result of a recipe (genome) based on what its ingredients (sequences) have been known to make previously.*

To be able to make connections between what a genome contains and the phenotypic features it yields, observations from previous studies can be leveraged. Given the many contributors and hence the large amounts of data available, it could provide enough observations to infer statistically significant patterns between known features. However, caution must be expressed when choosing which ingredients to associate with the final outcome. For instance, tomato sauce can be used in many other dishes that do not quite resemble a Hawaiian pizza. Similarly, genomic contents encoding a cell membrane is not unique to gram-negative bacteria; they may be found in any type of cell.

1.2. Project aim

This project seeks to combine knowledge and methods from the fields of microbiology, systems biology, and bioinformatics, with the aim of exemplifying how existing microbial data can be leveraged to supplement the current lack of phenotypic descriptions seen for many microorganisms. By utilising known data, connections will be made between two feature levels: what is encoded in the organisms' genome, and the traits they express in their phenotypes.

Known traits of microbial organisms will be collected and regarded as phenotypic features, while genomic contents will be represented by the organisms' functionally annotated genome in the form of FastA Amino Acid (FAA) sequences. With information on the two feature levels for a large selection of microbes, possible patterns can be inferred in the presence of particular genomic contents, and the organisms' expressed traits. Any potential patterns may thus be used to suggest the manifestation of the same traits for other organisms displaying the same genomic content. These predictions may be of use in various ways, with examples of being early assessments of a cultured microorganism's possible functions, prior to initiating resource-demanding examinations; characterisation of microorganismal traits present in an environment given an environmental DNA (eDNA) sample; and trait predictions for microorganisms that are yet to be cultured successfully.

In order to make such a GPA possible, data on species that have been cultured and described will be utilised. For this purpose, a dataset will be created to gather and store trait data from a variety of resources in a uniform manner. In addition to being of direct use in the present work, the assembled dataset aims to be a comprehensive source of microbial trait data and exemplify the utility of homogenous sources of biological data.

2. Background

In an attempt outline the immense progression seen within scientific communities since the days of Hooke, this chapter will provide insight into the current statuses of fields, resources, and key techniques relevant for the present project. These include microbiology, genomics, and systems biology. Additionally, some current limitations within said fields and techniques will be identified in order to exemplify the versatility of the present work.

2.1. Microorganisms in research

As the microbiota of Earth has been subjected to study, the interest and appreciation for it has but grown. The impact microbes have on humans is immense. Hence the tireless efforts of learning more about them. Both the health and disease of humans heavily depend on microbes: some are vital to proper digestion, while others are potent pathogens. Microbes are important in agriculture, as they can produce key plant nutrients. An example is fixation of atmospheric nitrogen into ammonia. Fermentation has been used for preservation of cheeses, vegetables, and meats for thousands of years. Industrial microbiology sees the utilisation of naturally occurring microorganisms in large-scale production facilities for antibiotics, chemicals, and enzymes [16, pp. 44-45]. Hence a large diversity of microbes is already well-known and utilised today. Still, considering the estimated millions of microbial species existing on Earth [6-9], this known variety may be vanishingly narrow compared to the entirety of the undiscovered microbial diversity.

For fundamental research, simpler microorganisms are often preferred for the study of universal cellular structures and functions such as gene expression and regulation. Many microbes are also used as tools for efforts like cloning and recombination [15]. There are several reasons why microorganisms are more beneficial for these endeavours than macroorganisms. Macroorganisms in this case refers to both cells and whole individuals of plants, insects, and animals; and even eukaryotic single-celled organisms like yeast. Perhaps the most obvious difference, when regarding microbes versus whole animals, is their size. Millions of microbes can be maintained on the same area as one larger organism. Secondly, the generation times of microbes are generally much shorter than that of cells from macroorganisms. A typical bacterial cell divides in less than one hour; yeast requires about 1.5 hours; while a mammalian cell may take 18 hours to divide [15]. Hence the time required to grow a sufficient number of cells for research purposes is generally significantly higher for more complex species. There are points raised regarding the potential rights and ethics of microbes, and how humans may not use them entirely as they please without concern [17], but the ethical perspectives of using microbes in research are in general lesser priorities than those for macroorganisms. Lastly, which is of particular relevance to the present work, is the accessibility of genome sequences for many microorganisms. With some exceptions, the genomes and number of genes are usually smaller in microorganisms than those of higher organisms [13 p. 96]. Their significance for human activities means that they are commonly studied, and their genomes often sequenced [18].

The very same features that promote the use microorganisms may simultaneously present several challenges and limitations to their versatility. The fact that they are small means that there is no way of observing their phenotypic features as readily as for macroorganisms. To study them, tools such as microscopes or imaging technologies must be utilised to magnify the microorganisms. Additional procedures are then required to infer any non-morphological information, such as whether the cells tolerate and utilise oxygen [19] and what metabolites they may produce [20]. Furthermore, the fast growth of many species may even be a disadvantage. The emergence of mutations that may alter the

research subject [21] and overgrowth preventing accurate colony counts [22] are two examples. Conversely, the GPCA is a prevalent issue in microbial cultivation that currently has no clear solution. It will be regarded further in Subsection 2.1.2 (p. 7).

There are several matters of microbes that are irrelevant to macroorganisms. A central feature like this is the phenomena of horizontal gene transfer (HGT). It denotes the transfer of genes between cells through other means than the regular vertical inheritance from a mother cell to its daughter [16, p. 69]. HGT is facilitated by the presence of a mobilome: genetic elements which are not confined to the genome of an organism. Examples include plasmids, which are DNA molecules distinct from the main chromosome that usually harbour genes for non-vital functions [16, p. 40]; transposons: elements that uses transposase enzymes to relocate within and between DNA molecules; and prophages, which have sequences originating from a viral infection [16, p. 291]. Thus some sequences of DNA can relocate within the organism's genome, but they may also spread to another cell entirely. As a consequence, the genetic material of the receiving microorganism can change dramatically without the need for a single cell division.

To be able to accurately detect DNA originating from another source, the two organisms must be phylogenetically distant enough for there to exist notable differences in their genetic contents [16, p. 290]. HGT may not compromise the association of particular traits to genetic content, however. Even if an organism containing a transferred gene does not have the corresponding feature recorded, this is likely also the case for much of its vertically inherited genomic content. Rather, the prospect of HGT and derived phenotypic features as a result of it, challenges several existing definitions commonly used for macroorganisms. Thus before the remaining background on microbes in research may continue, select concepts and terminology must be approached with particular care.

2.1.1. Terminology: traits and attributes

In the simplest definitions of the word, a "trait" is a specific characteristic of an individual which may be determined by genes, environmental factors, or a combination of the two [23]. It is broad and seemingly covers all features of any organism as long as it is linked to the individual level. Hence originally, there is no room for referring to the traits of a community of species, or the traits of an environment. Still, this has been widely done throughout the past decades, for instance in the many subfields of ecology [24]. Examples include the use of "trait" for characteristics of both morphological features (individual level) and soil nutrients, canopy height, and vegetation cover (ecosystem level) [25, 26]. This is an unfavourable use of the term "trait" because the characteristic of the immediate environment is rarely due to the contributions of just the organism in question. While an organism, be it a plant or a bacterium, may affect its environment for instance by the depravation of some nutrients and the excretion of others, there are likely other organisms present that do the same through their own traits. By attributing all these ecosystem changes as a "trait" of just one organism, its capabilities are incorrectly suggested, thereby compromising its accurate description. Thus the term "trait" should be reserved for the characteristics limited to the organism itself, and not to any higher level of organisation.

Furthermore, a distinction is made for a particular subcategory of traits. Already in 1859, Charles Darwin described that inherited variations of structure are important, whether they result in slight or considerable physiological changes [27, p. 12]. He goes on to suggest that not all features ("*contrivances*") seen in nature will suit our "*ideas of fitness*" [27, p. 472]. This is the basic notion of "functional traits". These affect the physiology of an organism in such a way that it may change its ability to survive, namely its fitness relative

to other individuals [28]. Functional traits are therefore often of particular interest when regarding an organism's phenotype, as it provides a direct way of measuring fitness.

Despite these definitions generally being readily applicable, there are limitations to their adoption for microorganisms. Of the most pressing is that of HGT. As described in Section 2.1 (p. 5), this feature enables microbes to derive new functionalities through lateral gene transfers from other cells. This challenges the current terminology due to the possibility of these additional features impacting the fitness of the receiving organism. For instance, a bacterium will increase its fitness if it acquires genes for antibiotic resistance through HGT [30]. Thus the organism's inherited variations and its implied fitness is altered, effectively changing which traits should be considered functional within the course of just one cell generation and independently of vertical gene transfer.

A potential solution to the issue of functional microbial traits is provided by Lajoie *et al*. [28]. They highlight that no trait is constant, attesting the need for further specification of the terminology. An example utilising their suggestion is that of a plant, which does not always have the trait of "green leaves". The leaves are non-existent until the plant sprouts, and they may turn yellow during autumn. Still, it does not mean that the trait of having such leaves is non-existent in the plant. Rather, its current state or environment does not allow their manifestation. The same concept applies to microorganisms. A bacterium may be capable of nitrogen fixation, but the momentary phenotype of the cell will depend on whether there is atmospheric nitrogen available. If none is present, no fixation will occur; but the organism is still able to metabolise nitrogen if the environment changes. Hence a trait will manifest itself differently throughout an organism's lifetime and in varying conditions. To account for the variations seen for any trait, the term "attribute" can be used to refer to its particular modality at any given time or place [28]. As a result, any derived functionality of a microorganism through HGT may merely pose a change in the attribute of a trait, for instance from "antibiotic-susceptible" to "antibiotic-resistant".

Summarising the possible terminology suitable for regarding the phenotypic characteristics of microorganisms, "trait" may be defined as a characteristic of an individual determined by its genotype and/or environmental factors. "Functional traits" are those that impact an organism's ability to survive. Lastly, an "attribute" is the value or modality taken by a trait at a given time or place and may vary. Thus the study of microbial traits may be free of terminological limitations, so that its practical issues may be of focus once more.

2.1.2. The great plate count anomaly

Recalling the benefits of using microbes in research and other human activities, most such endeavours require the organisms of interest to grow. Examples include the isolation of potential food-borne pathogens from a product; maintaining a strain used in fundamental research; or growing colonies which can be used for phenotypic characterisation of a novel species. Failed attempts at cultivation of an organism of interest might be due to incorrect plating techniques; contamination and out-competing by other species [15]; or other human errors. Other instances of lack of growth are more difficult to explain.

The GPCA is a phenomenon characterising the difficulty in growing microbes in controlled environments. It is estimated that only 1% of microbes are culturable [29], and even then, a sample containing a million cells may not yield more than 100 colonies [12]. There are several suggested reasons to the GPCA. A central one regards the conditions provided by the growth medium and how it may not successfully mirror the natural environment of the organism. This could for instance be by lacking essential nutrients and growth factors [30]. Still, there are documented instances where previously cultivated species present in a

sample still fail to grow into cultures [31, 32]. Thus despite the use of an appropriate medium, no growth may be achieved. Another possible reason to the GPCA is the presence of damaged or viable but non-culturable (VBNC) cells. These may be detected though various nucleic acid-based techniques but will not to form colonies when incubated [33].

Nevertheless, the lack of growth as a result of GPCA limits our ability to study the expressed traits of many organisms. This issue diminishes discoveries of many microbes' possible applications and utility. Introducing minuscule changes to the utilised media through trial-and-error in an attempt to adjust the conditions to yield growth, is a time-consuming and resource-demanding process. It may not be feasible at all if the microorganismal source is limited due to only a few cells being isolated from an inconvenient location. Thus alternative approaches have been adopted to circumvent the issue of GPCA. Attempting to learn more about an organism without requiring its prior cultivation may, in some cases, eliminate the need for cultivation entirely. However, it is of most use if the inferred features could aid future efforts in cultivation. The next Subsection will thus regard some microbial traits that are of particular interest to investigate through culture-independent techniques.

2.1.3. Microbial traits of interest

By applying the terminology and considering the issue of GPCA presented in the previous Subsections, it is possible to more accurately infer microbial traits that are of particular interest. This is of direct relevance to the present work through the aim of creating a dataset of microbial traits. However, the dataset must contain other data categories, such as organism name, taxonomic classification, and other supporting information. These must not be confused with "traits". Additionally, because traits are limited to the individual level, any characteristics detailing the features of the isolation source of an organism is not relevant in the present work. The isolation source itself is not a trait either, seeing as it is not determined by the organism's genes or environment. Still, it is an important piece of supporting information that may be used to describe the currently known microbial diversity. Hence its relevance for the present work and its dataset. For the same reason, growth temperature has also been regarded as relevant supporting information.

As suggested in Chapter 1 (p. 1), the motivation of the thesis is to exemplify the use of existing data to supplement the current lack of phenotypic descriptions due to issues in cultivation. Traits central to growth are therefore in high regard. An example is the nutrients an organism uses. These may be represented both by the substrates it needs for its main metabolic tracks, and essential compounds present in the culture medium, such as vitamins and cofactors [15]. Further, an organism's tolerance and requirement with respect to oxygen is also pivotal, seeing as erroneous oxygen supply may prevent the growth of the desired species. By gathering information on these traits relevant for growth, the dataset may facilitate the identification of patterns between known genotypes and traits on growth requirements. The patterns may then be applied to the genotypes of uncultured organisms in order to make educated suggestions as to their preferred growth conditions; directly facilitating cultivation efforts and the further study of the organisms.

In addition to facilitating their cultivation, the GPA methodology may also be utilised to infer on other traits not related to growth. Such trait attributes could supplement traditional assays for determining phenotypic features. Potential traits include metabolic capabilities such as trophy (generalising the organism's source of carbon, nitrogen, energy, etc.), enzyme activity, and produced compounds; growth rates, and clinically relevant traits such as pathogenicity and antibiotic susceptibility. Learning more about these traits may indicate

whether the microbes have any particularly interesting features, thus aiding the decision of whether substantial efforts should be made to develop their cultivation procedures.

In addition to traits on growth and metabolism, a good candidate trait for GPA is gram staining. It is a common way of classifying bacteria, enabling the differentiation between two main types of cellular envelopes based on their reaction to certain reagents. Being able to tell whether a cell is of either type reveals several details on their structure and functionality. A cell responding to the staining may be gram-negative or gram-positive. Fig. 2.1 conveys the central characteristics that differ gram-negative from gram-positive bacterial cell walls. Notably, both types contain an inner membrane and some peptidoglycan. This is a polymer structure providing mechanical strength to the cell. It is made by chained units of glycan tetrapeptide, which constitutes of two alternating residues made from glucose, in addition to various amino acids. The strength provided by peptidoglycan is due to peptide cross-links between the chains. Exactly how these links are formed varies between the two gram-stain attributes as well [16, p. 79].

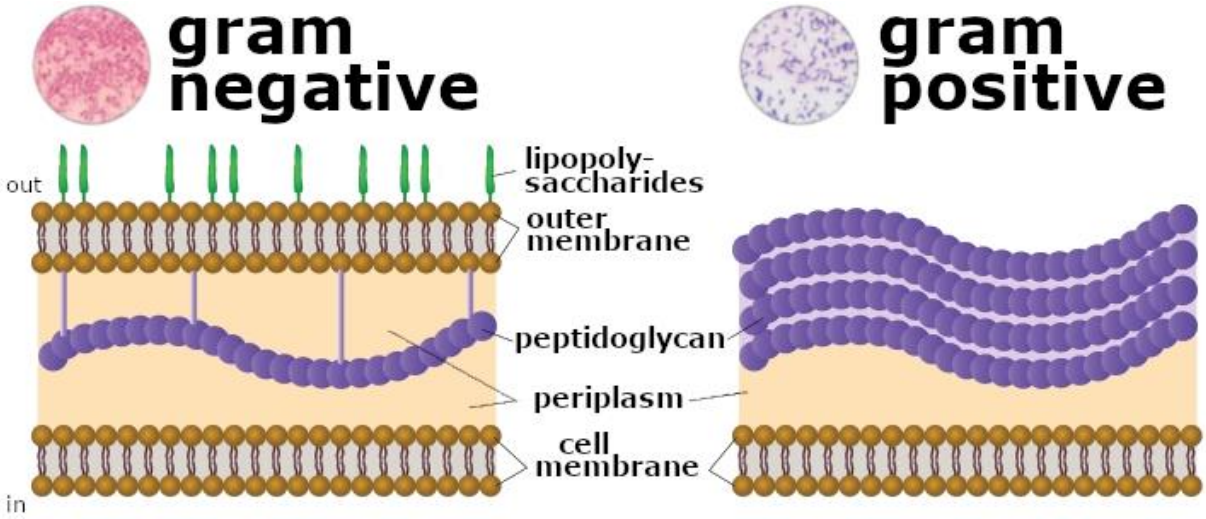


Figure 2.1: Differences between a gram-negative and a gram-positive cell envelope*. The circles above the illustrations depict how each cell type appears after gram staining, as a consequence of their differing cell wall structures. Adapted from [75, 76].*

Gram-positive cells usually contain several layers of peptidoglycan. The layers cross-link and yield an even stronger barrier between the cell's interior and its environment. As a consequence, the space between the cell membrane and the subsequent layers is usually smaller in gram-positives. An additional feature of gram-positive cell walls is that they often contain teichoic acids embedded within them (not depicted in Fig. 2.1). These have several functions, for instance to anchor the prominent peptidoglycan layer to the underlying cell membrane [16, p. 80]. On the other hand, the gram-negative envelope contains other components that are absent from its gram-positive counterpart. The outer membrane is the most prominent such structure, acting as a second phospholipid bilayer with additional inserted components. Lipopolysaccharides is one such type of component, providing additional features to the cell, such as toxicity to many animals [16, p. 82].

A benefit of using gram stain to describe cells is the generalisation of a large group of organisms into either "positive" or "negative". However, not all species have a gram-stain, and some do so variably, thus complicating the generalisation. Still, gram staining is a

highly unambiguous phenotypic feature. Few other traits can be interpreted as clearly. This is often due to a trait's high number of possible attributes and the lack of a well-established system for organising them. There are potential solutions to this limitation, which will be explored further in Subsection 2.3.4 (p. 14).

2.2. Databases

With the exponential increase in available data on microorganisms, the past decades have seen the creation of several repositories for storing biological data. They have central roles in archiving, maintaining, and sharing information derived in the myriad of research efforts performed around the world. Although genome sequencing was a major motivator for the creation of the first big biological databases [34], more and more disciplines see the need for more sophisticated solutions for storing their growing collection of data.

There are thousands of databases and datasets on biological information. They range from datasets of particular focus, such as single genera [35, 36] or metabolic functions [37]; to repositories aiming to collect data on a wide variety, such as features for many different types of organisms [38-40]. Thus the diversity of content available is immense, which is beneficial for the great variety of applications such data may have. However, a prominent issue is the varying ways the data may be presented. To be able to relate the information present in one database with that of another, it is central that an organism is identified similarly. The use of standardised naming conventions and identifiers, such as taxonomy identifiers, help this effort. Although not all sources or data entries follow these standards. This is especially true for organisms with sub-species classifications such as strain, biovar, and serovar. It is also affected by non-semantic differences, such as the inclusion of several strain designations or different symbols in the organism names [41, 42].

Another factor that may diminish the versatility of databases, is their varying data formats. This might be true both for continuous data categories, such as temperatures and growth rates; categorical fields, like trophy and metabolism; and some fields which are usually unstandardised, with isolation source and medium composition as examples. Subsection 2.3.4 (p. 14) will present a potential solution to the issue of unstandardised biological data.

2.2.1. Current databases on microbial data

As the present work seeks to exemplify the use of existing microbial data for GPA, large amounts of phenotypic data must be accessed. A comprehensive overview of biological databases is provided by Nucleic Acids Research [43]. Of notable mention is also works such as that of Madin *et al*. [44], who in 2020 considered and integrated bacterial and archaeal trait data from 26 sources. This Subsection mentions a few prominent databases of microbial data in order to demonstrate how repositories may be structured and utilised.

The Bacterial Diversity database (BacDive) [38] promotes itself as a leading database for standardised prokaryotic data. It utilises different data sources and methodologies for integrating microbial data, and thus contains many entries and a variety of information and trait fields. On October 11th 2021, the database contained 82,892 entries; a total which increased by an additional 6,653 entries until May 5th 2022. It is updated regularly and continues to grow to maintain its status as a prominent data source. Its data is readily available through a web interface. The data download process is less accessible, however.

Bergey's Manual of Systematics of Archaea and Bacteria [45] has been one of the main providers of descriptions on the known microbial diversity for almost a century. Nevertheless, the information is embedded in text descriptions for each organism and is

thus not readily accessible for computers and automated processes [44]. There have been previous efforts of extracting data from the manual [44, 46]. Still, such approaches do not capture all the information stored within the text-based entries. Until better text-mining approaches are developed, or human curation is applied, much of Bergey's information is out of reach for simple data downloads and automated processes.

The Joint Genome Institutes (JGI) Genomes OnLine Database (GOLD) [47] is a maintained collection of genome projects. JGI reports on more than 380,000 organisms from over 350,000 sequencing projects and 45,000 studies [47]. Its main focus is genome sequences. However, the entries also contain metadata. These include select trait fields and supporting information, with gram staining, oxygen requirements, and isolation sources as examples.

The Microbe Directory (TMD) [39, 48] is a relatively young repository for microbial information. It does not yet compete with the larger databases in terms of species coverage. However, the trait fields it reports are highly standardised and computer-friendly. Several categories have only two attributes and are reported with binary values. This trade-off from particularity to generalisations may not be suitable for all endeavours. For the present work, it facilitated custom attribute selection and merging of trait fields.

2.3. Leveraging biological big data

The previous Sections highlighted the use of microorganisms in research, and how large amounts of their phenotypic data is made available through various repositories. This Section first draws attention to the genotypic information acquired through genome sequencing techniques. Further, it seeks to highlight some of the scientific fields and approaches that utilise this information to learn more about the cell as a complex biological system. Lastly, some central tools for the use and analysis of big biological data are detailed, and their relevance to the present work is disclosed.

2.3.1. Sequencing and genomics

Since the introduction of the concept of genes and genomes in 1909 and 1920, respectively [14]; and the determination of the double-helical structure of DNA in 1952 [49], the field of genomics has revolutionised our perception of the word and its living constituents. The fundamental processes of life are studied extensively. A few examples are genome replication for cell division; transcription and translation of genes into proteins following the central dogma of molecular biology [13, p. 79]; and the concepts that challenge the central dogma, such as reverse transcription [16, p. 312] and functional ribonucleic acids [13, p. 85]. Utilising the concepts and components from these basic cellular functions, techniques have been developed to remake genomes *in vitro* and infer the order of its building blocks, namely the nucleotides adenine, cytosine, guanine, and thymine.

Following the age of the initial Sanger technique, new methods for sequencing have been developed. They have massively improved the time, cost, accuracy, and lengths of the inferred sequence reads. Next generation sequencing (NGS) such as Illumina [50] and Ion Torrent [51] allow for massive parallel sequencing using multiple copies of the fragmented target DNA. So-called third-generation techniques, with Nanopore [52] and SMRT [53] as examples, further improve on these techniques and can sequence a whole DNA strand in one go [13, p. 264]. Combined with development in data science and bioinformatics for processing and analysis of the generated data, the acquisition and study of genomes is now simpler and more accessible than ever.

With the study of genomes as a whole, rather than studying one single gene at a time, genomics seeks to deduce the functions of and interactions between stretches of the genome sequence. Examples include searches for related sequences in the genomes of different organisms; identifying coding and non-coding regions; and determining consensus sequences for common motifs, such as promoters and binding sites. The complexity of these concepts and their interactions have resulted in entire new research fields, with metagenomics and systems biology as prominent examples [13, p. 285-6]. The relevance of the latter field for the present project is explored in the next Subsection.

2.3.2. Genotype—phenotype association

As may be inferred, genomics has come a long way in unravelling how "life works". Still, there are many questions left unanswered. One such pivotal question is whether the acquired genotypes can be used to predict an organism's phenotype. A long sought-after approach has been to systematically collect data in order to sufficiently simulate how a cell processes information and gives rise to specific features through its responses [14]. This approach is convoluted by the complexity of biological systems. For instance, not all changes to a gene results in an altered phenotype. Further, the presence of a gene in the genome does not guarantee its expression. The many levels of regulation and organisation in a cell result in an intricate set of interactions where it is a rarity that one component single-handedly results in one specific outcome [13, p. 85].

Appreciation for this complexity is the essence of systems biology. According to this paradigm, it is not possible to model the behaviours of a complex system by a reductionistic approach of *only* considering the functions of individual constituents. Interactions between the components yield emerging properties which cannot be inferred from any one component alone [54]. Notably, this does not eliminate the prerequisite of knowledge on the system components [55]. Applied to the case of GPA, this entails that simply studying the genes present in an organism's genome will not successfully capture the intricacy of its resulting phenotype. Studying biological systems is therefore no easy feat, and some refute the possibility of modelling such complexity altogether [56]. In essence, this might be true: a model may never equal the concept which it is based on, but this does not render them futile. Rather, they help increase understanding for the system until additional pieces of the jigsaw puzzle is uncovered; either that be information on the components themselves or the interactions between them [55].

Until these figurative jigsaw puzzle pieces have been located, the present work seeks to utilise data that already has been systematically collected to infer phenotypic features based on genomic contents. In its simplest approach, such an attempt at GPA does not seek to uncover the mechanisms that ultimately determine how an organism's genotype results in its observed traits. Rather, it only needs to identify patterns between genotypes and phenotypes that have already been observed and reported on. These originate from the biological system and thus have applied these mechanisms themselves, independently of our understanding of them. Recalling the Hawaiian pizza analogy from Chapter 1 (p. 3), there was no mention of the instructions of the recipe. Acknowledged, the instructions are vital to a successful dish, and they would have helped identify the recipe. Still, the existing pattern of "pizza" inferred from the listed ingredients was so apparent that the instructions were unnecessary. Similarly, a GPA may be attempted on the basis of so many observations that the pattens between genotype and phenotype become equally apparent. The next Subsection details whether it is possible to ensure the correctness of any such inferred patterns.

2.3.3. Fisher's exact test

The motivation of the present work is to infer GPA patterns between genomic contents and phenotypic features of microorganisms. For this, Fisher's exact testing can be used. The test is based on a hypergeometric distribution in which a population of individuals is divided depending on categorical data [57]. Given a feature which either *is* or *isn't* present within an individual, the population is divided in two groups: those *with* (A) and those *without* ($\bar{A}$) the feature. Another categorical feature is present in the same population, again dividing the population in two groups (B and $\bar{B}$). Considering both features at once, the population has effectively been divided into four: those with both features (AB), those with just the first ($A\bar{B}$), those with just the second ($\bar{A}B$), and lastly, the individuals with none of the features ($\bar{A}\bar{B}$). The exact test may be used to see whether the distribution of individuals among the four groups indicate any dependencies between the two features. For instance, it will indicate whether an organism with known feature A is more likely to also display feature B [58]. The null hypothesis h_0 states that there is no such pattern, while the alternative hypothesis h_1, states that there is [59].

Odds ratio (OR) is an unconditional maximum likelihood estimate measuring association between an exposure ($A, \bar{A}$) and an outcome ($B, \bar{B}$). It is calculated as shown in Eq. 1 [60]:

$$OR = {ab/\bar{a}b}\Big/{a\bar{b}/\bar{a}\bar{b}} \tag{1}$$

where a, $\bar{a}$, b, and $\bar{b}$ denote the number of individuals within the groups A, $\bar{A}$, B, and $\bar{B}$, respectively. The resulting OR informs on both the direction and magnitude of the association. It may be normalised by $\log_{10}$-transformation (further denoted as $\widetilde{OR}$) for more intuitive interpretation. Eqs. 2-4 [60] present the meaning of the test's possible outcomes:

$$OR < 1 \leftrightarrow \widetilde{OR} = \log(OR) < 0\text{: } exposure\ associated\ with\ lower\ odds\ of\ outcome \tag{2}$$
$$OR = 1 \leftrightarrow \widetilde{OR} = \log(OR) = 0\text{: } exposure\ does\ not\ affect\ odds\ of\ outcome \tag{3}$$
$$OR > 1 \leftrightarrow \widetilde{OR} = \log(OR) > 0\text{: } exposure\ associated\ with\ higher\ odds\ of\ outcome \tag{4}$$

Thus the normalised $\widetilde{OR}$ indicate the direction of the association by its sign, and the magnitude by its absolute value. Not all associations may be of interest, however, even with a non-null magnitude. To only consider the strongest associations, a minimum threshold such such as $||\widetilde{OR}|| > 1$ may be set. This means that for an association to be regarded, the odds of a particular outcome must be at least 10 times higher when a particular exposure is given. Whether the null hypothesis should be rejected in favour of the alternative, depends on the significance probability (p-value) of the association. Only if the chance of erroneously rejecting the null hypothesis with a similarly extreme distribution is sufficiently low, should the inferred association pattern be accepted [59]. The p-value is corrected by false discovery rate (FDR) to account for multiple testing [61].

For the present work, the population is represented by the microorganisms in the dataset. The categories dividing the population into groups is the presence of a given trait attribute, and the presence of a particular genomic sequence. In the case of gram staining, and by disregarding the gram-variable and non-gram staining organisms, the population is divided into gram-negatives and gram-positives. The four resulting groupings of microorganisms are thus gram-negatives *with* and *without* a particular genomic sequence; and gram-positives *with* and *without* said genomic sequence.

As an example, consider a dataset with 1000 organisms. 600 are gram-negative, and 400 are gram-positive. Their genotypes reveal that 300 contain a sequence known to represent

a particular family of enzymes. Of these 300; 70 are gram-negative and 230 are gram-positive. The distribution, represented in a contingency table, will be as follows:

	$gram-positive$	$gram-negative$	
$has\ sequence$	230	70	$\rightarrow \quad \widetilde{OR} = \log\left(\frac{230}{70}/\frac{170}{530}\right) \approx 1.011$
$lacks\ sequence$	170	530	

In the example, the odds of an organism having a genotype which includes the enzyme family in question is over 10 times higher when it is known that the organism is gram-positive. The FDR-corrected p-value for this distribution is less than 0.001, thus the association is considered significant. This enzyme family may therefore be important for the phenotypic outcome of gram-positive staining. Thud the same sequence is found within any other genotype; it may indicate that this organism is gram-positive as well.

The provided example highlights how known data for 1000 organisms could be utilised to associate genotype with a phenotypic trait. In the case gram staining, the interpretation of the test results is relatively straightforward due to only two unambiguous gram stain attributes being utilised. This may not be the case for other microbial traits, as was explored previously (Subsections 2.1.3 p. 8, and 2.2 p. 10). The next Subsection details a possible solution to this issue: namely with systems for organising biological concepts.

2.3.4. Ontologies

As the insight into molecular biology, genomics, and related fields have deepened through decades of unrelenting research, the amount of available biological information has increased exponentially. Along with the need for large databases to store the acquired data, efforts have also been needed to coordinate the acquired knowledge and concepts. Without some level of order, the myriad of information is hard to navigate and thus of limited use. Ontologies provide a way of systemising knowledge through hierarchies of terms: from generalised concepts and down branch nodes representing the most particular biological details. One such system is Gene Ontology (GO), which is exemplified in Fig. 2.2.

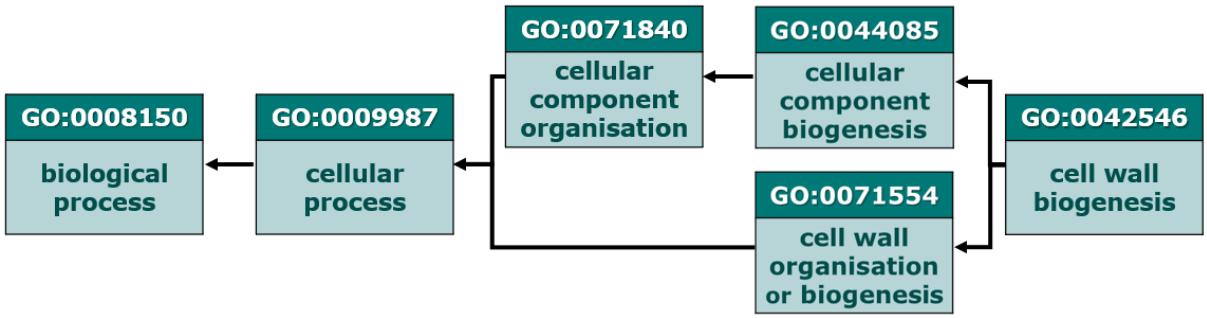


***Figure 2.2: Demonstration of a Gene Ontology network**, using the example child term "cell wall biogenesis" (right), which is generalised up to the top node "biological process" (left) through several levels in the hierarchy. Adapted from [62].*

GO serves as a controlled vocabulary and framework of biological concepts focused on the roles of genes and proteins [63]. In practice, it constitutes of three separate hierarchies: biological process (BP), cellular component (CC), and molecular function (MF). As of May 13th 2022, there are just over 28,000 BP, 11,000 MF, and 4,000 CC terms in GO. Thus it is a well-established and extensive ontology, providing terminology which can be used to unambiguously refer to particular processes, functions, and components within cells, independently of which organism is regarded. The hierarchical structure also facilitates generalisation of concepts by referring to parent nodes. Of particular note is that a GO term may have several parent terms, as is true for "cell wall biogenesis". GO is thus a

"loose" hierarchy. This entails that a term may appear at different distances from the top node depending on which path is taken. For instance, moving along the top path in Fig. 2.2 (p. 14) leaves "cell wall biogenesis" at four edges away from the top node of BP. The bottom path counts only three edges. As a consequence, the generalisation may differ depending on the chosen network path.

Other frameworks for conceptualising biological information are the Kyoto Encyclopaedia of Genes and Genomes (KEGG) Orthology (KO) [64, 65], and Clusters of Orthologous Genes (COG) [66, 67]. As their names indicate, they are not true ontologies. Orthology refers to the phenomenon in which genes found in different species have the same function, and where this is because they are descended from the same gene in the organisms' last common ancestor [16, p. 427]. Because the genes have the same function, they can be described by the same term independently of which organism they are found in.

KO terms focus on molecular functions that have been manually defined based on KEGG networks. It is a hierarchical system with several top nodes that branch down into over 58,000 particular functions found within the database [64]. Notably, KO is a strict hierarchy, meaning that a node has no more than one direct parent. Additionally, there is only four levels to the KO hierarchy, which yield fewer and more generalised parent categories compared to those of GO. The COG system is even more extreme on this matter. It is the smallest terminology hierarchy of the three mentioned, with the official database listing about 5,000 terms. They are loosely organised into 26 parent categories, yielding just two levels in the hierarchy. Any generalisation will thus see the direct application of one of the 26 top nodes. Despite offering a simple way of summarising COG terms, all details are lost when considering that the vast diversity of cellular functions is generalised into only 26 terms; the largest two of which are "Function unknown" and "General function prediction only". Thus the COG system is less established than the two former hierarchies.

The ability to clearly define a vocabulary of terms simplifies the task of describing biological features. Miscommunication is limited and automated processes for data management is facilitated. Several commonly reported microbial data categories would benefit from the implementation of ontologies. Examples include isolation source, substrate, and growth medium. Many databases often see rather creative descriptions of these largely string-based data fields. Examples include the reported isolation sources: "*wastewater of an acidic water neutralization facility, Water temperature, salinity and pH of the wastewater sample were 18°C, absence of salinity and pH 7.0, respectively*" [68], and "*a bacterial mat dominated by Epsilonproteobacteria growing on a black smoker hydrothermal chimney within the Loki's Castle hydrothermal vent system at a depth of 2350 m*" [69]. Such entries are very informative when read and curated by humans. However, because they do not follow any structure, the information is largely unavailable for other approaches. Additionally, there is no simple way of comparing such reports across data sources. With an ontology for isolation source however, these descriptive retellings could be represented by lists of standardised terms, increasing their accessibility.

There are ontologies developed for reporting traits, such as Ontology of Biological Attributes [70], ecoCore [71], and Ontology of Microbial Phenotypes [72]. However, the benefits of established terminology may only apply if it is agreed upon and actively utilised by relevant parties. Currently, no such trait ontology is widely applied by microbial data repositories [72]. Thus until such ontologies are employed by the wider microbiology community, those seeking to gather homogenous microbial data must usually conduct any necessary standardisation procedures or conversions into ontologies themselves.

3. Methods

This chapter details the methodology conducted in the present work. The first Section describes the creation of a microbial trait dataset. It includes the description of the downloading and preparation of multiple sources of microbial information; and their assembly into one homogenous dataset. The second Section regards how the genomes of a selection of the organisms included in the trait dataset, were accessed and annotated. The final Section regards the GPA analysis of the collected data and its demonstrated application for inferring phenotypic attributes for other microorganisms.

3.1. Microbial trait dataset

In order to circumvent the issue of microbial trait data being scattered across multiple sources and data formats, the first major task of the present work was to gather trait data from various data sources and present them in a homogenous manner. The following Subsections detail the procedures of accessing, cleaning, and reviewing the content of 19 sources for microbial trait information. Further, a selection of these were assembled into one homogenous dataset. The final Subsection details how a selection of the organisms of this dataset was extracted to a reduced dataset for the subsequent GPA methodology.

3.1.1. Data download and cleaning

To assess which sources of microbial trait information to utilise in the present work, the number of organisms (coverage) and the relative number of data reports of select categories (completeness) in various data sources were compared. All relevant data were therefore downloaded and partially cleaned so that their formats aligned sufficiently for this comparison. In addition to the conducted procedure detailed for each data source, select categories were standardised to ensure homogeneity across datasets. Table 3.1 summarises the terminology applied to the datasets of the present work; listing the possible values each field could contain in their original datasets, and the standard to which they were formatted.

Table 3.1: Data standardisation scheme. *Select fields of reported data were changed to the presented standard to ensure homogeneity across datasets.*

Fields	Original data values	New data values
Type strain, extremophile, antibiotic susceptibility	yes, y, 1	yes
	no, n, 0	no
Gram stain	positive, positiv, pos, p, +, 1	positive
	negative, neg, n, -, 0	negative
	variable, var, +/-	variable
	Indecisive	NaN
Oxygen requirement	(…)-ic (…)-be	(…)-ic
Trophy	(…)-troph, (…)-trophic	(…)-troph
Growth temperature	x-y	x,y
	x°C x°	x
Incubation period	x days	x
	x-y	x,y
	>x	x,
	<x	,x

The raw datasets from the utilised sources are all available in Supplementary information 1 (Appendix A, p. 85), while Supplementary information 3 (App. A) contains the prepared

datasets. Scripts utilised in this Subsection are included in Supplementary information 2. If nothing else is specified, the datasets were prepared using terminal commands and Microsoft Excel (ver. 2203) [73] and its built-in functions. Semicolons are used as column delimiters in all files produced throughout the project work.

3.1.1.1. BacDive

BacDive [38] enables downloading of select data through their website interface. In lack of a setting for selecting all entries, a search querying "*Type strain: no"* or "*Type strain: yes*" in the Advanced Search was conducted. All entries (82,892 as of October 11th 2021) were then returned and could be added to the Download Selection in chunks. Custom export comma-separated value (CSV) files were generated by checking the fields in the Download Section of BacDive as indicated in Table 3.2. The files are included in Supplementary information 1 (App. A, p. 85).

***Table 3.2: Data fields from BacDive**, which were downloaded and cleaned for assessment.*

Category	Field	Renamed field
Name and taxonomic classification	ID_strains	bacdiveID
	genus	genus
	Species name	species
	Strain designation	strain
	Type Strain	typeStrain [H]
Morphology	Incubation period	incubationDays [H]
Culture and growth conditions	Culture medium	medium
	Temperature	growthTemp [H]
Physiology and metabolism	Name of produced compound	producedComp
	Metabolite (production)	
	Production	
	oxygen tolerance	oxygen [H]
	Nutrition type	trophy [H]
	Metabolite (utilization)	substrate
	Metabolite (physiological)	assay
	Enzyme, Enzyme activity [R]	reducedEnzAct increasedEnzAct variableEnzAct
Isolation, sampling, and environmental information	Sample type/isolated from	source
Sequence information	Genome seq. accession no.	genomeAccNo
External links	Culture collection no.	cultureNo

[H] homogenised [R] reformatted

From the output files, the data columns mentioned in Table 3.2 were extracted. A new column with the full name was created from the combined genus, species, and strain name. Where no strain name was available, the culture collection number was used. To ensure homogeneity, data values in select categories (marked [H] in Table 3.2) were changed to follow the standard described in Table 3.1 (p. 17).

Based on the BacDive entry IDs, rows representing the same organism were merged. For string-based fields (isolation source, growth medium, and produced compounds), all unique field values across rows with the same ID were combined with a vertical slash ("|") delimiter. For instance, three entries of *Acetobacter orleanensis* (ID 8) had reports of respective isolation sources "beer", "beer", and "belgian bottle beer". The combined cell thus reads "beer|belgian bottle beer". For numerical fields such as incubation duration and

growth temperature, the merged cell displays the range of values represented in each column. For *A. orleanensis*, four rows reporting growth temperatures (in °C) of 28, 30, 25, and 30 yields a merged cell reading "25,30". Notably, this is different from a report of "25|30", which does not indicate a temperature range.

The data fields *enzyme* and *enzyme activity* (marked [R] in Table 3.2, p. 18) were reformatted. In their original structure, one enzyme and its respective activity level was indicated per dataset row, yielding near duplicate entries for organisms registered with several enzymes. To reduce the number of rows to one per organism entry, and simultaneously eliminate dependency across columns, separate lists were made for the names of enzymes with each type of activity level: decreased, increased, and variable. For instance, *Actinotignum schaalii* (BacDive ID 145) had several original enzyme entries and activity levels, such as "decreased alkaline phosphatase", "increased alpha-galactosidase", and "variable pyrazinamidase". The *reducedEnzAct* column for this organism thus reads "alkaline phosphatase"; the column *increasedEnzAct* contains "alpha-galactosidase"; while "pyrazinamidase" is found within the *variableEnzAct* column.

Missing data entries were left as empty cells. Lastly, all instances of semicolons were substituted by commas as to not interfere with column delimitation.

3.1.1.2. Bergey's

Microbial trait data from Bergey's Manual of Systematic Bacteriology [45] were retrieved through the prepared data file "bergeys.csv" (Supplementary information 1, App. A, p. 85), originally acquired from Madin *et al.* [44]. The dataset was reported to have been extracted from text-based entries of the original repository. Table 3.3 details the data categories from the mentioned file utilised in the dataset of the present work.

Table 3.3: Data fields from Bergey's *considered in the present work.*

Field	Renamed field
tax_id	taxID
genus	genus
species_name	speciesStrain [M]
source	source
doubling_h	doublingH [H]
metabolism	oxygen [H]

[H] homogenised [M] modified

All other data fields than those mentioned in Table 3.3 were removed. For 77 entries, part of the organism's name was enclosed by square brackets to indicate classification uncertainty. The brackets were removed and a new column, named *misclassified* and data points set to "yes", was created to indicate the uncertainty for the relevant entries. The full organism names from the *speciesStrain* field were utilised to create two new columns for the species and strain names, respectively. Data values in select fields (marked [H] in Table 3.3) were changed to follow the standard described in Table 3.1 (p. 17). Lastly, missing data entries across all columns were changed from "NA" strings to empty cells.

3.1.1.3. Campedelli

In Campedelli *et al*. of 2018 [35], growth conditions for 196 type strains of the *Lactobacillus* genus were studied. Madin *et al.* [44] manually extracted the growth data from the article and produced a standardised dataset, "campedelli.csv" (Supplementary information 1, App. A, p. 85). From this prepared file, data on organism names and oxygen requirements

was extracted. A separate field for the genus was created from the full organism names, and the oxygen requirements were ensured to follow the set standard of Table 3.1 (p. 17).

3.1.1.4. Corkrey

In 2016, Corkrey *et al*. [74] reported growth conditions and kinetics of 661 microorganisms. Their data was accessed and prepared by Madin *et al.* [44], yielding a single dataset for the taxonomic, growth, and reference information for each entry. Table 3.4 summarises the fields extracted from this prepared file and utilised in the present work.

***Table 3.4: Data fields from Corkrey** utilised in the current dataset.*

Field	Renamed field
tax_id	taxID
org_name	speciesStrain
growth_temp	growthTemp [H]
metabolism	oxygen [H]
trophy	trophy [H]
doubling_H	doublingH [H]
reference	ref

[H] homogenised

A separate field was created for the genus names extracted from the full organism names. The values in the four trait columns (marked [H]) were standardised according to Table 3.1 (p. 17) to ensure homogeneity. Lastly, missing data points were left as empty cells.

3.1.1.5. FAPROTAX

The Functional Annotation of Prokaryotic Taxa (FAPROTAX) (ver. 1.2.4) [77, 78] was downloaded and reformatted through the script in the Jupyter Notebook (ver. 3.6.0) [79] "faprotax_reformat.ipynb" (Supplementary information 2, App. A, p. 85). Operations include the separation of metabolic function, organism names, and references from the original single column into three; adding and removing groups of organisms from select functions according to database entry instructions; and removing entries without taxonomic information at the species level. From the full organism names, separate columns for genus; genus and species; and stain names were created. Missing data entries were left as empty cells, and all semicolons were substituted by commas.

3.1.1.6. IJSEM

The International Journal of Systematic and Evolutionary Microbiology (IJSEM) phenotypic database (ver. 3) [80] was downloaded, and the data columns in Table 3.5 were extracted.

***Table 3.5: Data fields from IJSEM** utilised in the current dataset.*

Field	Renamed field
Genus name	genus [M]
species name	species [M]
strain name	strain [M]
Habitat	source
article doi	ref
oxygen preference	oxygen [H]
Metabolism assays	assays
Sole carbon substrate use	substrate
reference	ref

[H] homogenised [M] modified

In addition to data standardisation in accordance with Table 3.1 (p. 17), other modifications of the raw file were needed to ensure homogeneity in the dataset. Firstly, four organism names were changed. One entry with a digital object identifier (DOI) as its genus name, "10.1099/ijs.0.006320-0" , was changed to "*Bradyrhizobium"* [81]. Another entry had genus, species, and strain names of "10.1099/ijs.0.02424-0", "*Thalassolituus*", and "*oleivorans*", respectively. The latter two were shifted to their correct taxonomic level, while the strain name was left blank. A third entry had genus and species names "*Mycobacterium gordonae*" and "*Mycobacterium paragordonae*", respectively. The latter name was kept as per the entry reference [82]. Finally, the name of one entry was changed from "*Plasticicumulans* not yet known; article proposed: *Plasticicumulans lactativoran* YDT (=DSM 25287T=NCCB 100398T)" to the proposed name.

Secondly, the genus name was included prior to the species name in the *species* column. Any mention of the genus and species names was removed from the *strain* column, leaving only the strain name. A separate column for the full organism names was created by combining the genus, species, and strain names. Based on the full organism names, matching rows were merged with "|" as a delimiter between unique trait data. Missing data points were left as empty cells, and semicolons within cells were substituted by commas.

3.1.1.7. JGI GOLD

JGI GOLD [47] in its entirety was provided by its authors to Madin *et al*. [44], yielding the file "gold.csv" in Supplementary information 1 (App. A, p. 85). Table 3.6 summarises the select trait categories from this prepared file that were utilised in the present work.

***Table 3.6: Data fields from JGI GOLD** utilised in the current dataset.*

Field	Renamed field
tax_id	taxID [S]
org_name	species
STRAIN	strain
metabolism	oxygen [H]
gram_stain	gram [H]
isolation_source	source

[H] homogenised [S] supplemented

Entries on *Homo sapiens* and viruses were removed, as were entries without information on at least one of the three trait categories included in Table 3.6. Three rows (organisms *Pseudomonas aeruginosa* Habs 0, *Stenotrophomonas maltophilia* 810-2, and *Stenotrophomonas maltophilia RH 1168*) in the dataset were corrupted. Here, no taxonomy ID was reported, and other data points did not separate correctly due to the use of different delimiters. Taxonomy IDs were supplemented [83, 84] and the data points were manually distributed to the appropriate columns. Fields for genus, strain and full organism names were created. Using the full organism names as key, rows on the same organism were merged and unique datapoints joined in one cell was separated by "|" delimiters. Data values in the columns *oxygen* and *gram* were changed according to Table 3.1 (p. 17) for homogeneity across datasets. Finally, missing data points were changed to empty cells, and all semicolons were substituted with commas to avoid further delimitation issues.

3.1.1.8. Kremer

In 2017, Kremer *et al*. [85] quantified and reported growth conditions and kinetics for 194 phytoplankton species, including growth rates at specified temperatures. The dataset "lno10523-sup-0008-suppinfo8-1.csv" ("kremer_data.csv", Supplementary information 1,

App. A, p. 85) contains several data categories, with accompanying descriptions in the file "lno10523-sup-0005-suppinfo5-1" ("kremer_description.txt", Supplementary information 1). The fields specified in Table 3.7 were extracted from the data file.

Table 3.7: Data fields from Kremer *considered for the current dataset.*

Field	Renamed field
Name	speciesStrain
Environment	source
Temperature	growthTemp
R	growthRate
Ln_rT	growthRateLnCorr

A separate column was created for the genus names. Using the full organism names as key, rows were merged and observations for each column were joined in one cell separated by "|". No data points were missing; thus any particular set of growth rates and temperatures have the same index in their respective cells.

3.1.1.9. Mason

Growth rates of various bacterial species was reported by Mason [86]. From the original publication, Madin *et al.* [44] extracted the data and updated the organism names to the current standard of National Centre for Biotechnology Information (NCBI) (raw file in Supplementary information 1, App. A, p. 85). For the present work, the updated names were mapped to the original Mason dataset. Where no updated NCBI name was found, the original report was used. Data on growth medium, temperature, generation time, and entry references were extracted and changed to follow the standard of Table 3.1 (p. 17). The full organism names were used to create separate columns for genus; genus and species; and strain names. The generation times (minutes) were reformatted to doubling time (hours) by division with 60. Identical organism entries (true for two organisms in the dataset, *Corynebacterium pseudodiphtheriticum* and *Pseudomonas syringae*) were merged with "|" as a separator between unique trait attributes. Missing data points were left as empty cells.

3.1.1.10. MediaDB

From the "growth_data" table in the MediaDB database [87, 88], the columns described in Table 3.8 was extracted.

Table 3.8: Data fields from MediaDB *considered for the current dataset.*

Field	Renamed field
Genus	genus
Species	species [M]
Strain	strain
Growth_Rate	growthRate
Temperature_C	growthTemp [H]

[H] homogenised [M] modified

The genus name was added prior to the species name in the *species* column. A new column combining the genus, species, and strain names was created. With the full organism names as key, rows for the same organism were merged and unique data values in each column were joined in one cell with "|" as a separator. One entry, *Escherichia coli* B834(DE3), was deleted due to having no reports for either trait. Growth temperature data were changed

to follow the same standard as previous datasets (see Table 3.1, p. 17). Missing data points were left as empty cells.

3.1.1.11. Moore

The growth data from the study of Moore *et al.* [36] was provided to Madin *et al.* [44], in which the observations were standardised (Supplementary information 1, App. A, p. 85). From this prepared dataset, organism names and doubling time (hours) were extracted for the present work. From the full organism names, separate columns were created for genus; genus and species; and strain names.

3.1.1.12. The Microbe Directory

From the full database of TMD [39, 48], the categories listed in Table 3.9 were extracted.

Table 3.9: Data fields from TMD *utilised in the current dataset.*

Field	Renamed field
genus	genus
species	speciesStrain
gram_stain	gram [H]
microbiome_location	source
extreme_environment	extremophile [H]
antimicrobial_susceptibility	abSusceptible [H]
animal_pathogen [R]	pathogenicity
plant_pathogen [R]	

[H] homogenised [M] modified [R] reformatted

All entries with names containing "virus" were removed. From the full organism names in *speciesStrain*, the genus; genus and species; and strain names were inferred and placed in three separate columns. The binary columns indicating extremophiles and antibiotic susceptibility were standardised as specified in Table 3.1 (p. 17). The columns of animal and plant pathogenicity were merged into one, and the original binary values were replaced by categorical values "plant" or "animal" in accordance with the reported pathogenicity.

3.1.1.13. Nielsen

In 2006, Nielsen [89] studied and reported on size dependency of growth rates for species of cyanobacteria and green algae. The data is not publicly available, but was provided to Madin *et al*. [44], from which a dataset was produced (Supplementary information 1, App. A, p. 85). From this file, a separate column with the organism genus names was created. Ten unique organism entries without a specified species name were removed. Lastly, rows representing the same organism were merged with "|" as a separator between data entries for each cell in the columns *reference*, *type*, *growthRate* (d^{-1}), and *size* ($\mu\mathrm{m}$). Associated data across the three columns thus kept the same index, enabling extraction of particular sets of observations from the combined rows.

3.1.1.14. Pasteur

The Biological Resource Centre of the Pasteur Institute provides a catalogue of microorganisms [40]. From their web interface, entries on Bacteria were queried and exported. The columns listed in Table 3.10 (p. 24) were extracted for use in the present work.

***Table 3.10: Data fields from Pasteur** utilised in the current dataset.*

Field	Renamed field
Taxonomic name	speciesStrain
Reference équilibre	cultureNo
Type strain	typeStrain [H]
T° of incubation in C°	growthTemp [H]
Bibliography	ref
Isolated from	source
Genotype	genes
Atmosphère incubation	oxygen [H] [M]
link medium composition	medium

[H] homogenised [M] modified

Entries with no data for any of the selected traits were removed. From the full organism names in the *speciesStrain* column, the genus and species names were inferred and copied to two new, separate columns. Changes in data values to ensure homogeneity across datasets (see Table 3.1, p. 17) were conducted in the columns on growth temperature and oxygen requirement. Additionally, data points which only informed on carbon dioxide (CO_2) conditions within the *oxygen* column were deleted. Lastly, entries with the same full organism names were merged, and unique data points from each row were joined in a single cell with "|" as data separators.

3.1.1.15. PATRIC

The Pathosystems Resource Integration Centre (PATRIC) [90] provides a dataset file named "genome_metadata" in the resource's FTP server [91]. This file was downloaded and the organism taxonomy ID, name, gram stain and oxygen requirement (columns indices 4, 2, 55, and 62, respectively) were extracted using terminal commands as listed:

```
$ wget ftp://ftp.patricbrc.org/RELEASE_NOTES/genome_metadata
$ awk -F '\t' '{print $4";"$2";"$55";"$62}' genome_metadata > patric.csv
```

The resulting raw data file "patric.csv" is provided in Supplementary information 1 (App. A, p. 85). The full organism names were cleaned for heading and tailing special characters. Names enclosed by square brackets, indicating classification uncertainty, were resolved through literature searches: [*Clostridium*] *mangenotii TR* was updated to *Clostridioides* [92]; [*Eubacterium*] *cylindroides* ATCC 27803 to *Faecalitalea* [93]; while [*Scytonema*] *hofmanni* UTEX 2349 was not updated [94]. All brackets were subsequently removed from these entries. The full organism names were utilised to create separate columns for genus; genus and species; and strain names. Gram and oxygen requirement fields were changed to follow the standard of Table 3.1 (p. 17). Missing data entries were reformatted to empty cells and lastly, rows without trait information were deleted.

3.1.1.16. PhyMet^2

From the Phylogeny and Metabolism of Methanogens (PhyMet) database (ver. 2, PhyMet^2) [37] of Michał *et al*. [95], the columns described in Table 3.11 (p. 25) were exported into a CSV file.

Table 3.11: Data fields from PhyMet² *considered for the current dataset.*

Field	Renamed field
Name	species
Type strain	strain
Gram reaction	gram [H]
Min. growth requirements	substrate
Additional growth requirements	
19 compounds * [R]	
5 environments ** [R]	source
Other environment [R]	
Main publication	ref

[H] homogenised [R] reformatted

* 1-butanol, 2-butanol, isobutanol, 2-propanol, acetate, butanol, carbon monoxide, cyclo-pentanol, dimethylamine, dimethyl sulfide, ethanol, formate collections, H2+CO2, H2+methanol, methanol, methylamine, propanol, propionate, and trimethylamine.

** Intestinal tracks, reactor, soil, volcanic, and water.

Separate columns were made for the genus and full names of the organisms. Gram stain data were homogenised as indicated in Table 3.1 (p. 17). The substrate growth requirements were merged into one column. For this purpose, the values in the 19 columns on specific nutritional compounds were first modified from "1" to the compound name. A similar substitution of "1" values by category name and subsequent column merging was conducted for the six isolation environment columns. Lastly, missing data points were changed from "no data", "not applicable", and "not indicated" to empty cells.

3.1.1.17. ProTraits

The atlas of Prokaryotic Traits (ProTraits) [96] allows downloading of its database from a web interface [97]. Table 3.12 summarises the columns extracted for use in the dataset.

Table 3.12: Data fields from ProTraits *utilised in the current dataset.*

Field	Renamed field
Organism_name	speciesStrain
Tax_ID	taxID
109 compounds * [R]	substrate
2 ecosystems * [R]	source
14 ecosystem categories * [R]	
18 ecosystem types * [R]	
20 ecosystem subtypes * [R]	
5 specific ecosystems * [R]	
44 known habitats * [R]	
*7 habitats** [R]	
11 hosts * [R]	

[R] reformatted

* see columns in the raw data file *protaits.csv* with prefix identical to each respective field name

The *speciesStrain* column was utilised to create separate fields for genus; genus and species; and strain names. A column indicating misclassification was created and assigned "yes" for 33 entries with bracket-enclosed names, from which the brackets were subsequently removed. For the column *gramStain*, instances of "1" were changed to "positive", and "0" to "negative", as per the standardisation procedure in Table 3.1 (p. 17). For other columns with binary data values, instances of "1" were substituted by the column

name, such as "ethanol" and "hostAssociated". Instances of "0" and "?" were substituted by empty cells. Lastly, columns were merged as indicated in Table l, with "|" as a separator between values from the included cells.

3.1.1.18. RefSeq

From the GenBank [98] flat files stored in the assemblies of The Reference Sequence (RefSeq) FTP (file transfer protocol) server [99, 100], Madin *et al*. [44] extracted select data categories into a CSV file (Supplementary information 1, App. A, p. 85). From this prepared file, the organisms' taxonomic identifiers, names, and isolation sources were extracted for use in the present work. Separate columns were created for the full; genus; genus and species; and strain names, and entries without isolation source data were removed.

3.1.1.19. Vieira-Silva

Vieira-Silva *et al*. [101] reported data from their study on generation times for 214 microbial species in their Supplementary Table S1, available in Supplementary information 1 (App. A, p. 85) of the present work. From the table, columns on organism names, minimum generation time (d), and optimum growth temperatures (OGT) were extracted and changed to follow the standard detailed in Table 3.1 (p. 17). The full organism names were used to generate separate columns for genus; genus and species; and strain names.

3.1.2. Comparison of source datasets

Following the download and initial cleaning of the data sources described in the previous subsection, the coverage of the 19 datasets were compared. The notebook "initial_overview.ipynb" (Supplementary information 2, App. A, p. 85) assembles an output file with the same name (Supplementary information 4) that lists the fields present in each dataset; the number of non-NaN entries for each field; and their completeness relative to the coverage of the relevant dataset. Based on a visualisation of the dataset coverages with Tableau Desktop (ver. 2021.3.1) [102], the largest datasets were selected for further utilisation.

3.1.3. Trait dataset assembly

To join the selected datasets, the notebook "dataset_assembly.ipynb" (Supplementary information 2, App. A, p. 85) was utilised. First, all column headers were standardised to the same naming convention. The datasets were concatenated, and a new column named 'database' indicated the source dataset for each entry. Four new columns were introduced to standardise the full, genus, species, and strain names of the entries. Indicators of taxonomic level ("sp.", "subsp.", and "str".), special characters (|!"#%&/()[]{}=+?'*-_.:,~^°), and spaces spaces were removed, and all letters were adjusted to lowercase.

In the same script, the content of the assembled dataset was plotted in three ways. First, the output file "assembledDataset_fieldCoverage_plottable.csv" (Supplementary information 4) was used to create a stacked bar chart to convey the total number of non-NaN data entries for each field within the dataset, and how many entries each source dataset contributed with. Secondly, the unique data values and their frequencies in each data category were inferred and saved to the file "assembledDataset_statistics.csv" (Supplementary information 4). Tableau Public was used to create the visualisations, using both pie charts and histograms. Third, the standardised organism names were used as keys to plot the overlap between datasets on genus-, species-, and strain-level organisation using the UpSetPlot package (ver. 0.6.0) [103, 104], based on the prepared file "category_coverages.csv" (Supplementary information 4). Similar figures were used to

plot overlapping trait data entries between datasets, measured on strain-level organisation. The visualisations of category overlaps were considered when a selection of trait dataset rows was merged, as will be described in the next Subsection.

3.1.4. Preparing a reduced trait dataset

To limit the scope of the subsequent GPA methodology and reduce the amount of genotypic data to gather and process, both the number of organisms and included trait categories were reduced in a copy the assembled dataset. In order to qualify for this "reduced dataset" that would be used in the second part of the project, any entry had to fulfil two requirements. Firstly, the entry had to contain data for at least two of three traits of particular interest: gram stain, oxygen requirement, and substrate usage. Secondly, the entry had to have an annotated genome in FastA amino acid (FAA) format available within the NCBI FTP server through either the GenBank [98] or the RefSeq [100] repository. Hence the column *genomeAccNo*, functioning as an identifier for a particular organism and its genome assembly, was also included in the reduced dataset.

To ensure that trait reports from all data sources were considered for each organism, select rows within the reduced dataset were merged prior to filtering the dataset by the two described conditions. Entries on the same organism from different datasets were merged based on the standardised full organism names, with unique trait values originating from different rows being merged into one cell with "|" as a delimiter. From the *genomeAccNo* column, the first GCA (GenBank) or GCF (RefSeq) accession number was extracted to a separate column to increase accessibility. No additional changes were introduced in the cells of the *substrate* column. For the trait columns on gram stain and oxygen requirement, conflicting reports from different data sources required resolving before proceeding.

As an example, three datasets reported on the gram staining for *Lactobacillus rhamnosus*: two reports of gram-positive staining, and one report of gram-negative. The value was changed to the observation which a majority of reports agreed on. In the instance of *L. rhamnosus*, the value was thus set to "positive". Where no majority was present, literature searches were conducted to decide the gram stain. For oxygen requirements, the observation with majority amongst the reports was also utilised. Where no majority was present, but all reports agreed on either aerobic- or anaerobic-related terms, the most general of the included terms was chosen. If both aerobic- and anaerobic-related terms were reported for an organism, the oxygen requirement was marked as "conflict". In the example of *L. rhamnosus*, the reported oxygen requirements were "aerobic", "microaerophilic", and "facultative aerobic". These are all terms indicating aerobicity. Hence the observation was changed to the most general term, "aerobic". Following the row merging, the first condition for inclusion in the reduced dataset was enforced by removing entries which did not contain data in at least two of the three trait columns.

The second condition of having an genome sequence in FAA format in NCBI was enforced using the genome accession numbers of the trait-filtered entries. The accessions were used to map each dataset entry to their respective assemblies by referring to NCBI's overview of prokaryotic entries within their FTP server ("ncbi_prokryotes.txt", Supplementary information 5, App. A, p. 85) [105]. The GCA FTP address matching an entry's accession number was added to a new column in the dataset. The corresponding GCF FTP address was added to a second column by copying the GCA column and substituting the links' two instances of "GCA" with "GCF". All links were given the suffix "_protein.faa.gz" so that they led to the direct download page of the assembly's zipped genome FAA file. The FTP links were extracted to text files "protein_links_all_GCA.txt" and "protein_links_all_GCF.txt"

(Supplementary information 6). With the scripts "protein_links_check_GCA.sh" and "protein_links_check_GCF.sh", the links in each file were sorted depending on whether they led to an existing genome file, with successful links being saved to the files "protein_links_GCA.txt" and "protein_links_GCF.txt" (Supplementary information 6). The dataset was thus filtered to only keep entries with successful FTP links. Finally, the entries meeting both criteria were exported to the final reduced dataset, "reducedDataset.csv" (Supplementary information 7).

3.2. Genome sequence annotation

The previous Section described the assembly of a trait dataset of microbial phenotypic information from several data sources, and the extraction of select entries and data fields to a reduced dataset which would be used in the second part of the project. The next step of the present work was to gather data on genotypic contents. The next Subsections describe the methodology for obtaining the genomes and their functional annotations.

3.2.1. Genome downloads

With the reduced trait dataset containing direct links to the NCBI assembly genome FAA files, the scripts "get_proteins_GCA.sh" and "get_proteins_GCF.sh" (Supplementary information 6, App. A, p. 85) were utilised to download the FAA files from the FTP servers of GenBank and RefSeq, respectively. The two sets of downloaded genomes, each with files numbered from 1 to 3307, are available in Supplementary information 8. The script "compare_protein_counts.py" utilises BioPython's SeqIO package (ver. 1.76) [106] to compare the number of sequences included in the GCA- and GCF-retrieved genomes, and returns the results in the file "comparison_GCA_GCF.csv". For genomes where the sequence count was at least 15% larger for GCF-entries, the RefSeq-acquired genomes were utilised rather than their GenBank GCA equivalent. The preferred genome for each organism was kept for functional annotation. The same key used for the genomes (1-3307) was introduced to the corresponding entries in the reduced trait dataset, in order to facilitate the subsequent connection between collected traits and annotation results.

3.2.2. Functional annotation

Following the downloading of the genomes for the entries of the reduced trait dataset, functional annotation was conducted to discover genomic characteristics stored within the FAA files. The annotations were conducted using the eggNOG mapper (ver. 2.1.7) [107] based on eggNOG orthology data [108] and sequence searches utilising DIAMOND protein alignment [109]. Through submission of a job array ("slrm_prot_anno.sh" in Supplementary information 9) to NTNU's high performance computing cluster Idun [166], the annotation process was automated so that each job in the array handled one genome. Each FAA file was unzipped, submitted to eggNOG for annotation, and rezipped. The annotation output (emapper.annotations files) were placed in a separate directory (raw_annotations in Supplementary information 9) for subsequent processing.

The script "anno_extract.sh" (Supplementary information 9) was used to obtain select columns from the eggNOG-generated annotations: COG, GO, and KO terms. For each annotated genome, the term columns were extracted into three respective files along with the genome key (1-3307). The result was thus three two-column files ("COG_extract.csv", "GO_extract.csv", and "KO_extract.csv" in Supplementary information 9) listing the genome key and its associated annotation terms. The three lists were further prepared with the notebook "annotation_term_overview.ipynb". Included in the tasks within this script is the cleaning of annotation terms and column reformatting to yield one term per

dataframe row, producing the final annotation overviews "terms_COG.csv", "terms_GO.csv", and "terms_KO.csv" (Supplementary information 9). With Matplotlib (ver. 3.3.4) [110], the notebook plots the frequencies of terms across the annotated genomes.

3.3. Genotype—phenotype association

Previous Sections within this Chapter detailed the assembly of a dataset of microbial traits, and the functional annotation of genome sequences representing the genotypes of a selection of these organisms. Knowing what genomic contents are present within the organisms and which phenotypic traits they exhibit; the remaining methodology seeks to exemplify how these data can be used to infer patterns between the two feature levels.

Of focus for demonstrating GPA was the trait of gram staining. The reasoning behind this choice is based on its number of reports in the dataset (sample size) and limited number of possible attributes (simplicity). This choice is detailed further in Subsection 5.3.1 (p. 61). The notebook script "fisher_gram.ipynb" (Supplementary information 10, App. A, p. 85) was utilised to first filter the reduced trait dataset by removing organism entries with "variable" or no registered gram strain. The filtered dataset was merged with each of the three annotation files containing COG, GO, and KO terms, respectively. In the resulting dataframes, terms with frequencies in less than 5% or more than 95% of the genomes were removed. Contingency tables were subsequently created and assigned to a new column in each dataframe. For each entry, the 2x2 contingency table listed the number of entries with each possible gram stain value (positive `p` or negative `n`) that either had (`h`) or lacked (`l`) the given annotation term: $[[ph, nh], [pl, nl]]$.

To test the null hypothesis stating hypergeometric distribution among the four groups in the contingency table; meaning there is no association of a particular term for any particular gram stain trait, Fisher's exact tests were utilised. With the SciPy statistical functions module (ver. 1.6.2) [111], the contingency tables were submitted to the exact tests. Resulting OR were normalised by $\log_{10}$-transformation (denoted $\widetilde{OR}$), and values with infinite magnitudes were set to a threshold outside the $\widetilde{OR}$ range ($\|3\|$). The p-values were FDR-corrected using Statsmodels (ver. 0.12.2) [112]. Volcano visualisations (ver. 2.0.8) [113] were created for each of the three term classes, plotting $\widetilde{OR}$ against corrected p-values with respective significance thresholds of 1.0 and 0.01. The resulting dataframes ("fisher_*.csv" files) are available in Supplementary information 10 (App. A, p. 85).

Different procedures were conducted to increase the readability of the produced results, and all are contained in part in "term_visualisation.ipynb" (Supplementary information 11, App. A, p. 85). Firstly, the three term classes were all divided into four based on whether the terms were significantly associated with either gram stain attribute ("significants"); and whether they were found exclusively in genomes of organisms with a particular attribute ("exclusives"). An overview of the distribution across these four categories ("significant_count.csv", Supplementary information 11) was plotted in the same notebook.

Each COG term was generalised into its parent category using the overviews provided by NCBI's FTP COG server: "cog-20.def.tab" [114], and "fun-20.tab" [115] (both included in Supplementary information 12). The generalisation was plotted using Tableau Desktop. For GO terms, the four categories were submitted to separate queries in the tool REVIGO (ver. 16.11.21) [116]. Default options were used, except for the value representation for $\widetilde{OR}$ included in the submitted list for all significant terms for either attribute. For $\widetilde{OR}$ of gram negative-associated terms, the "Lower value is better" setting was opted. "Higher value is

better" was used for $\widetilde{OR}$ of gram positive-associated terms. The outputs files are included in Supplementary information 13. Lastly, KO terms were submitted to the KO database [117] and mapped to KEGG modules. The lists of modules and their matched KO terms were extracted ("ko_mapper.csv", Supplementary information 14), and a column with the terms' associated $\widetilde{OR}$ values ("ko_or.csv") was added. Tableau Desktop were used to plot the KO modules for all four term categories.

Based on the produced figures representing the associations between annotation terms and gram attributes, their biological connotations were inferred. The identified pattens were attempted applied to demonstrate the use of the collected data. For three random organisms which lacked gram stain reports in the assembled dataset, the genome accession numbers were used to access and download their genomes in FAA formats. EggNOG was again used to functionally annotate the genomes, and the terms indicative of the inferred GPA were searched for in the annotation files of these new organisms. The files used for this test are found in Supplementary information 15 (App. A, p. 85).

4. Results and analysis

This chapter will provide the results of the present work and their immediate analyses. Its Sections follow the structure of the methodology in the previous Chapter, starting with the assembled dataset of microbial traits. This is followed by the generation of genomic contents through functional annotation. Using the traits and functional annotations, the third and final Section presents the connections found between the two feature levels. Observed patterns were applied to test organisms from the dataset to demonstrate the use of the conducted method for inferring new trait attributes based on genomic contents.

4.1. Microbial trait dataset

The discovery of sources for microbial trait information saw the consideration of 19 databases, repositories, and datasets from different authors. The first of the following Subsections summarises the relative impressions given by these source datasets' outputs. Secondly, the condition of the dataset formed by their assembly is regarded. This includes statistics on the dataset as it is presented, and the organismal overlap seen between the utilised data sources.

4.1.1. Microbial trait data sources

An overview of the relative coverage (number of organisms) of the 19 refarded data sources for microbial traits is given in Fig. 4.1. The datasets range from 24 entries in Moore, to 58,169 in JGI. Campedelli reports on the fewest traits by only containing only oxygen requirements. On the other end of the spectrum is BacDive with 12 trait fields.

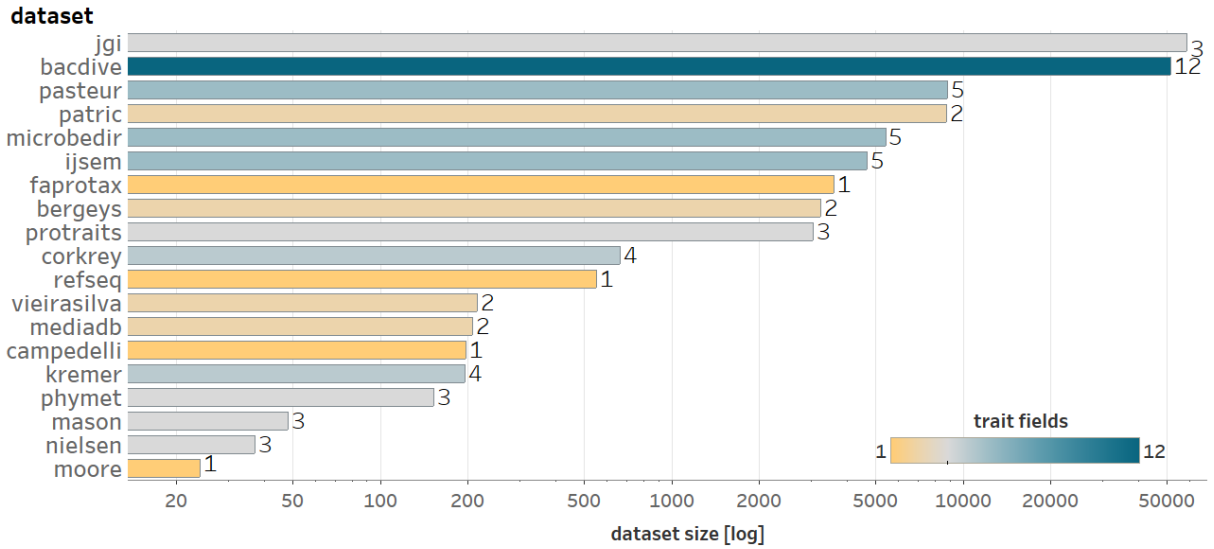


***Figure 4.1: Coverage comparison of the source datasets**. Bar lengths represent the number of entries in the source, given on a log-transformed scale. Colours and labels indicate the number of traits fields each data source contains.*

https://public.tableau.com/app/profile/jenny.merkesvik/viz/dataset_comparison/initial_overview

An observation from the overview in Fig. 4.1 is that there is no clear connection between dataset coverage and the number of traits it reports on. Rather, the diversity of data categories seems highly connected to the aim and area of focus of each respective data source. For example, FAPROTAX seeks to gather data on microbial metabolic functions, which it does utilising only one trait field. In contrast, BacDive aims to be the largest database of standardised bacterial information. Hence its broad range of data categories.

Upon reviewing sources of microbial trait data, the great variation between them was apparent. The differences are related to dataset coverage and the trait fields they include, but also in terms of format and accessibility. While some provide large datasets covering thousands of organisms, such as BacDive and JGI, others are highly specialised and only cover species within particular genera or select metabolic functions. Examples include *Lactobacillus* in Campedelli, *Prochlorococcus* in Moore, and methanogens in PhyMet[2]. Some sources provide web-interfaces for accessing and downloading data, like BacDive and Pasteur. Others require programmatical or command-line approaches, with examples being FAPROTAX and PATRIC. On the other hand, not all datasets are publicly available, such as Nielsen and Moore. The varying format and accessibility are one of the main challenges tacked by the present work and will be regarded further in the Chapter 5 (p. 54).

4.1.2. Dataset assembly

From the 19 gathered datasets, the ten largest were selected for assembly. Together, they yield a dataset with 147,676 entries ("assembledDataset.csv", Supplementary information 4, App. A, p. 85). Fig. 4.2 summarises the constituents of the assembled dataset, noting by colour the origin of the data in each included category. The elements within top bar of Fig. 4.2 are equivalent to the bars represented in Fig. 4.1 (p. 33) before log-transformation. The contrasts between data source coverages are most clear within this section of Fig. 4.2, with the combined efforts of BacDive (purple) and JGI (yellow) comprising 74.2% of the assembled dataset's total entries.

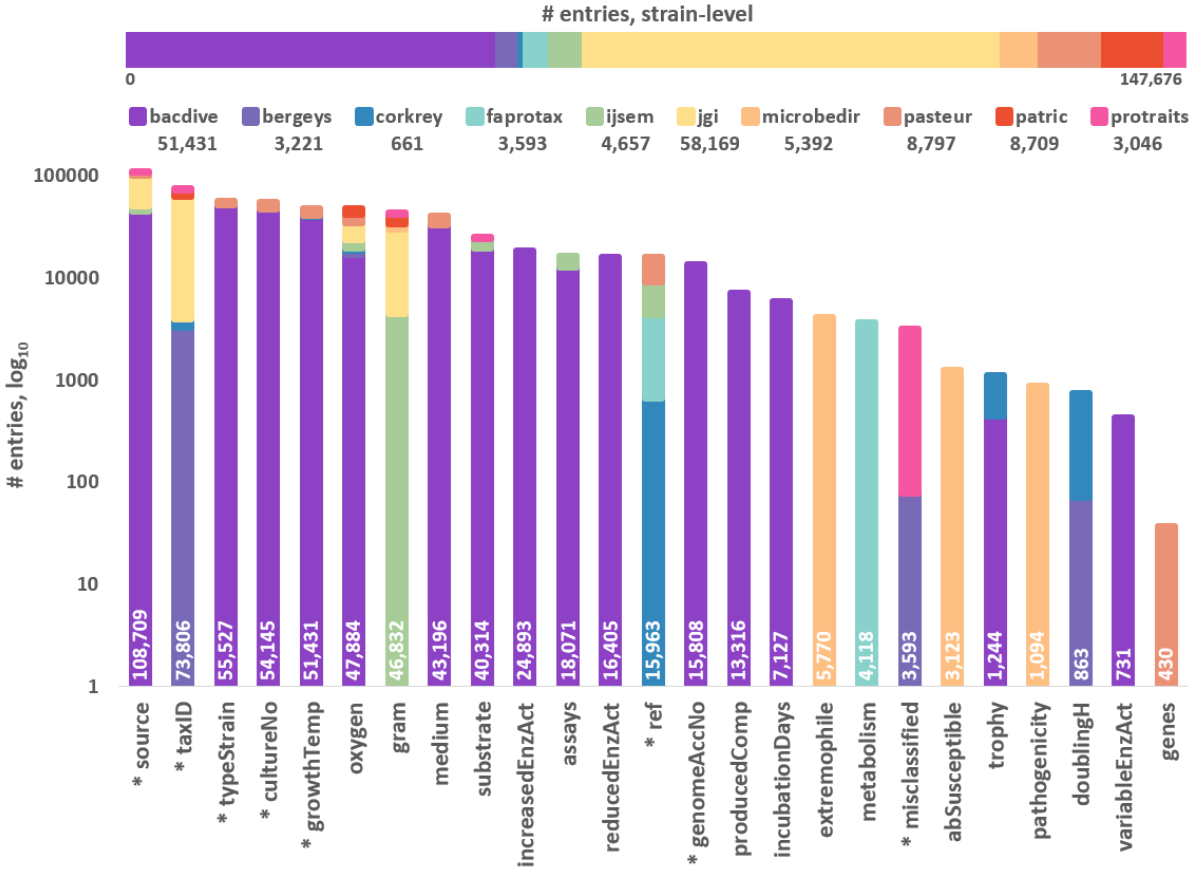


Figure 4.2: Data category overview of the assembled dataset *comprising of ten sources of microbial trait information, indicated by colourations. The top bar represents the number of entries from each data source. The main histogram shows the number of non-NaN data entries within the included data categories, represented on a log_{10}-transformed scale. The exact sizes are noted as white data labels on each bar. Category names marked with asterisks are considered as identifiers or supporting information, rather than strict "traits" as per the defined terminology.*

As to the data categories represented in the bottom section of Fig. 4.2 (p. 34), BacDive asserts itself as a major source of trait information in most categories. Notably, it is the only source reporting on enzyme activities, produced compounds, and incubation time measured in days. For this latter trait, other data sources report similar categories but utilise doubling time in hours. Additionally, the non-trait field genome accession number is solely included in BacDive entries. The accession numbers are central for the later functional annotation. This possible limitation will be regarded in Chapter 5 (p. 53).

The charts in Fig. 4.3 (p. 36) show the attribute variety found in the assembled dataset. It is based on the overview provided in "assembledDataset_statistics.csv" (Supplementary information 4, App. A, p. 85), where all possible attributes and their frequency within the dataset is provided for each field. Fig. 4.3 panel A (p. 36) shows that the eight most common genera comprise just over a third of the dataset, which contains 565 unique genera in total. These genera are all common organisms that have prominent applications and roles: *Staphylococcus*, *Streptococcus*, and *Mycobacterium* are all common pathogens [118-120]; *Escherichia* is a well-known constituent of the human microbiome [121]; and *Bacillus, Pseudomonas*, and *Streptomyces* are often utilised for industrial purposes [122-124]. These genera are studied extensively due to their relevance for human activities, hence the many reports on their expressed features.

When combining the information provided by Fig. 4.3 panels A and B (p. 36), it can be derived that for five of the eight most frequent genera in the dataset, many of the entries are of the same species. For the 8,882 entries in the *Streptococcus* genus, over half are of *S. pneumoniae*. Filtering the assembled dataset for this species returns 5,506 different strains reported by eight different data sources. Thus there are duplicate rows in which different sources report on the same strain, which likely due to their particular relevance in various applications. This observation will be discussed further in Chapter 5 (p. 56). Still, most entries are identified as separate strains, corroborating the great microbial diversity.

Panels D-V in Fig. 4.3 (p. 36) provide insights into the most common attributes within each trait. Most entries with known gram stain (panel D) are gram-negative. Indole testing is the most frequent metabolic assay (panel E) with positive outcome, an observation which is complimented by the high frequency of indole production (panel L). About three fourths of the known oxygen requirements (panel F) indicate oxygen tolerance (comprising the attributes aerobic, facultative, microaerophilic, or aerotolerant). For the registered entries, panel H reveals that carbon is most commonly acquired through organic sources (heterotrophy); and energy is more often obtained by electron donor oxidation (chemotrophy) than by solar energy (phototrophy, within "other").

The incubation times (Fig. 4.3 panel T, p. 36) range from 0 to 56 days, with most entries requiring less than two weeks of incubation. For doubling time (panel U), most reports are of growth rates below 10 hours per cell division. There are instances of significantly higher doubling times, with the maximum of 4566.1 hours (approximately 190 days) for *Methylococcus capsulatus*. This seemingly low growth rate is explained by referring to the named reference of this particular entry: a study of growth at low temperatures [125]. Panel V shows that most of the registered growth temperatures fall under the mesophilic range (20-45°C), with the most frequent reports at 28°C, 30°C, and 37°C. Of note in this panel is the possible bias towards round and "common" values, such as the average human body temperature. Similar biases are likely present within other trait categories and thus may challenge the representativeness of the attribute distributions indicated in Fig. 4.3 (p. 36). This topic will be explored further in Chapter 5 (p. 56).

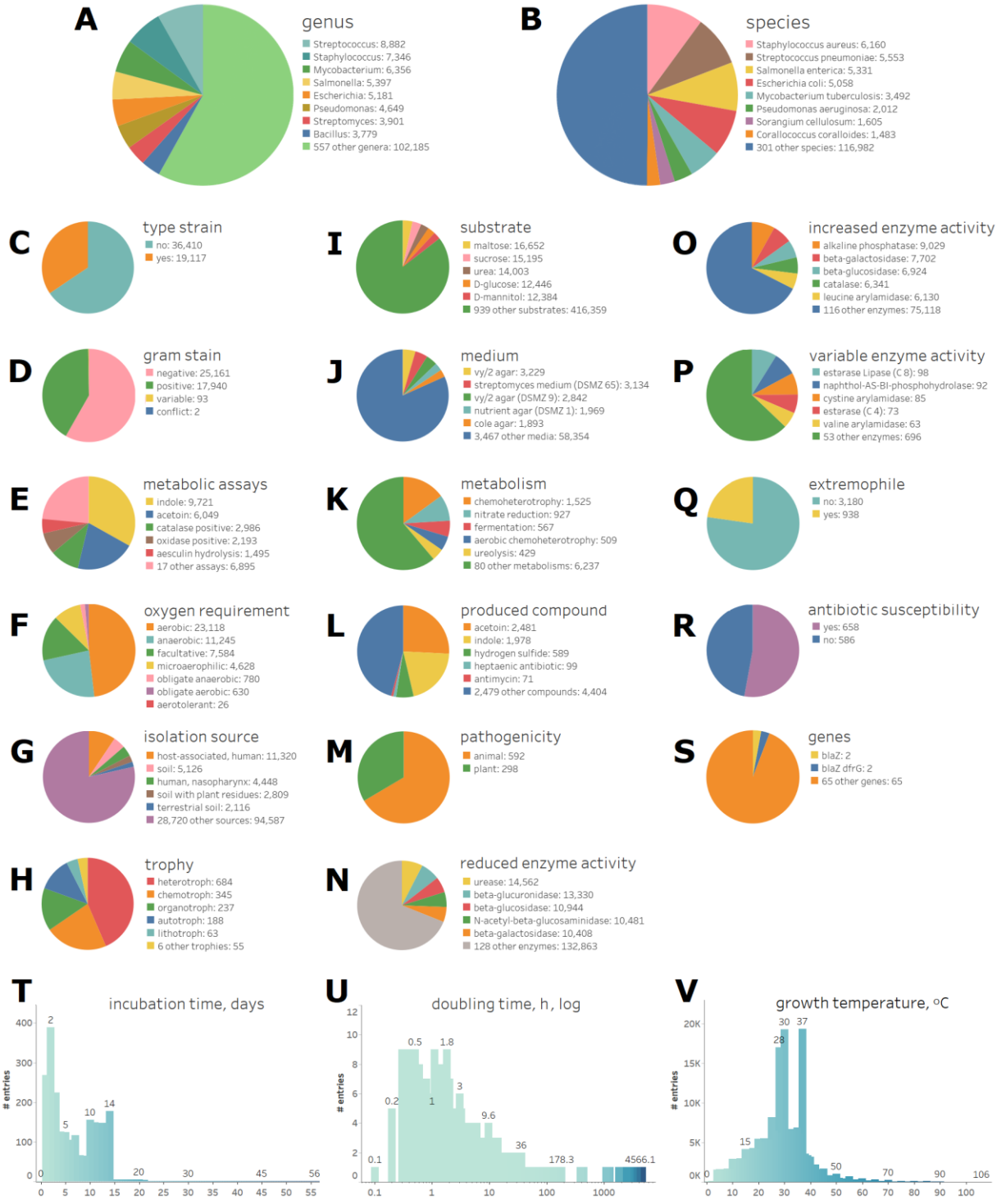


Figure 4.3: Trait content overview of the assembled dataset*. Panels* ***A-C*** *cover taxonomic levels and identifiers, while panels* ***D-V*** *represent traits included in the dataset. For panels* ***A-S****, the most common attributes are displayed as slices with sizes relative to their frequency, whose numerical value is found in the legend. The attributes in panel* ***H*** *have been reduced to consituents of combined terms, so that "chemolitoautotroph" contributes to frequencies of three attributes. In panels* ***T-V****, numerical traits have been presented as categorical bar charts. Their colours follow a linear gradient scaled to each value range. The horizontal axis of panel* ***U*** *is* log_{10}*-transformed to increase readability of the most frequent doubling times.*

https://public.tableau.com/app/profile/jenny.merkesvik/viz/dataset_comparison/mergedDataset_statistics_1

With the standardised organism names excluding taxonomic level indicators, capital letters, and species characters, the entries across datasets could be compared more readily. For instance, *Acidovorax* sp. NO-1 (from JGI) and *Acidovorax* sp. NO 1 (from TMD) would be considered the same entry; as would *Vibrio cholerae* O1 str. EC-0009 (PATRIC) and *Vibrio cholerae* O1 EC-0009 (JGI). Based on these standardised names, the number of unique organisms in the dataset is 126,763. The panels in Fig. 4.4 show the overlap between data sources, using these standardised organism names as keys.

Each panel presents the organismal coverage of the ten datasets across the taxonomic levels of genus, species, and strain. They indicate the coverage by the vertical bars in left part of each panel, while the coloured dots below the histograms convey which datasets overlap, indicating which sources may contain information on the same organisms. For example, Fig. 4.4 panel A tells that BacDive contains a total of 3,037 genera, of which 844 are found only in BacDive, as indicated by the vertical bar and the filled-in dot beneath it. It thus follows that 2,193 genera in BacDive must be found reported by other data sources as well. For instance, the third column in panel A shows that BacDive and JGI have 207 genera in common which are not found in any other datasets.

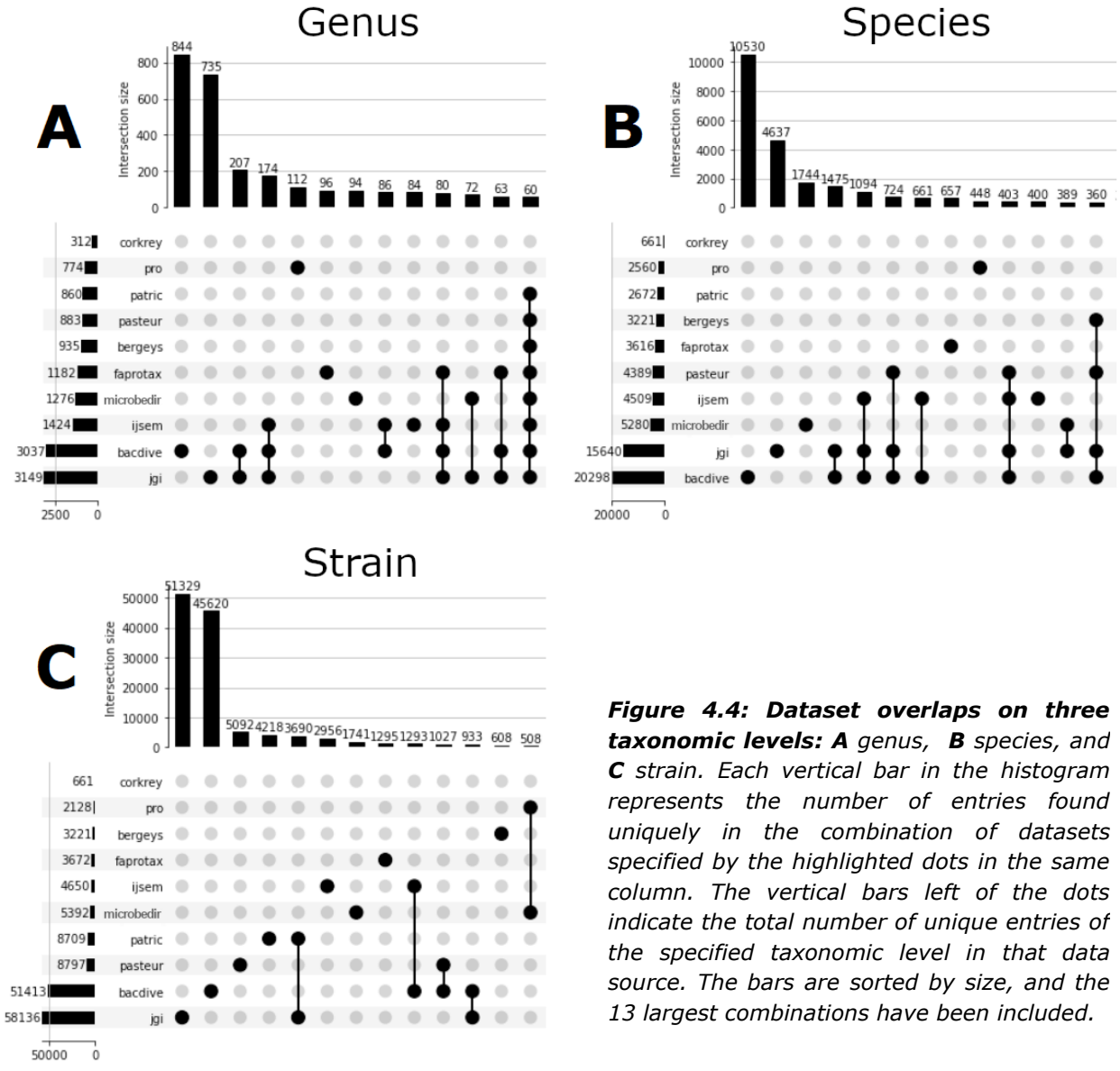


Figure 4.4: Dataset overlaps on three taxonomic levels: A *genus,* ***B*** *species, and* ***C*** *strain. Each vertical bar in the histogram represents the number of entries found uniquely in the combination of datasets specified by the highlighted dots in the same column. The vertical bars left of the dots indicate the total number of unique entries of the specified taxonomic level in that data source. The bars are sorted by size, and the 13 largest combinations have been included.*

In accordance with the interpretations from Figs. 4.1 (p. 33) and 4.2 (p. 34), JGI and BacDive appear in Fig. 4.4 (p. 37) as the largest data sources within the assembled dataset. The relative heights of their bars do however indicate some notable differences between them. Fig. 4.4 panel B (p. 37) shows that BacDive contains 20,298 species, which exceeds JGI's count of 15,640. Still, JGI reports on 7,000 organisms more than BacDive when regarding the strain-level overlap in panel C. This indicates that BacDive has a broader range, while JGI provides more total entries on a narrower selection of microbial species. Given that BacDive collects microbial data from a variety of databases and scientific publications [38], it is expected that its range is broader than that of JGI, whose entries originates from genome projects registered to the repository [47].

There is relatively little overlap between the datasets. The level of uniqueness increases with taxonomic specificity, which is evident due to the top one category for genus; top three for species; and top four for strain, all being single datasets. This observation is believed to be due to the organisms actually being different, however missed overlaps is also a possible explanation of the high level of uniqueness across datasets. The matter of false negatives within organism name matching is discussed further in Chapter 5 (p. 55).

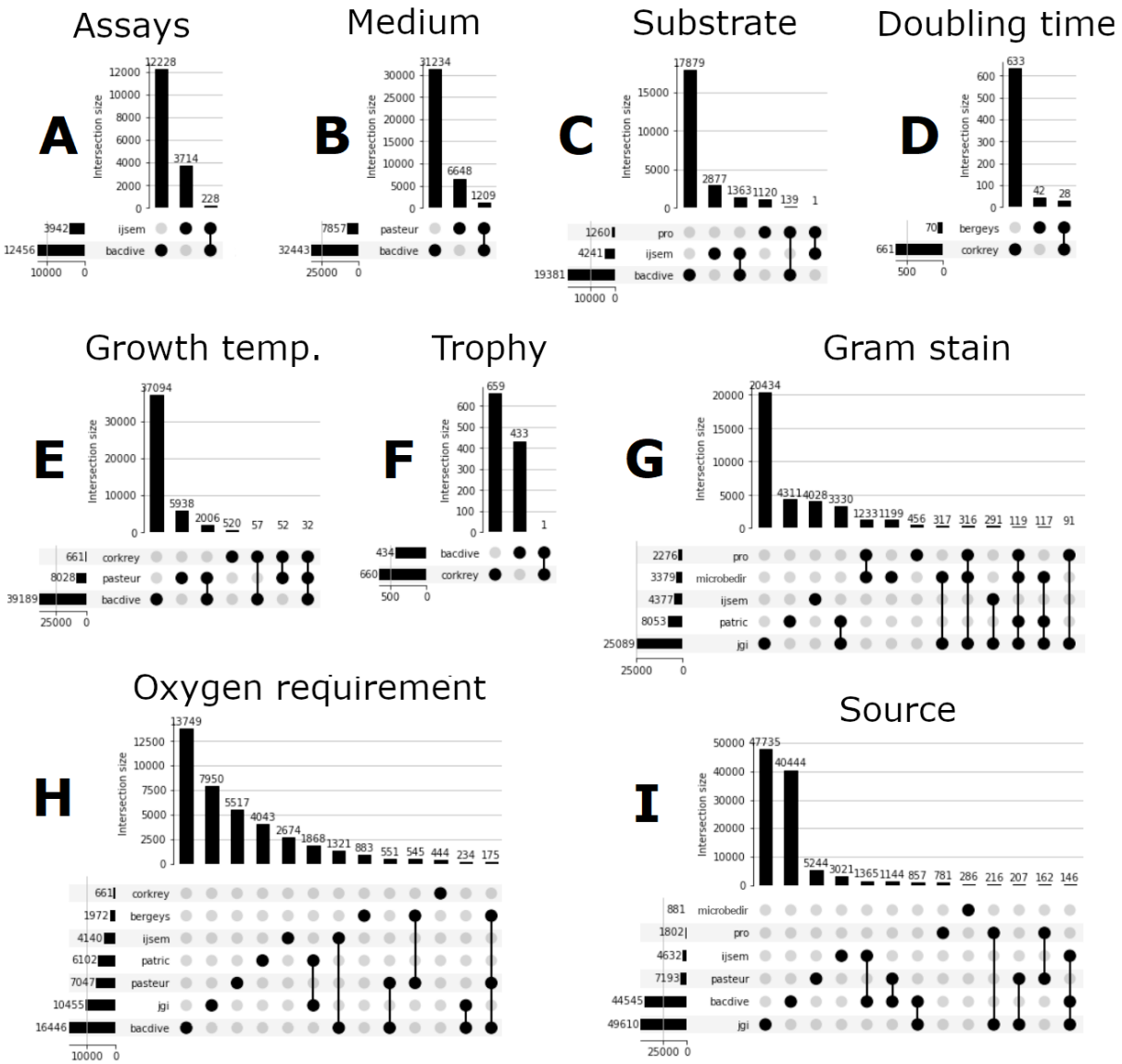


***Figure 4.5: Dataset overlaps for nine data categories**, measured on strain-level. Each panel **A-I** represent one data category. For panels **G-I**, only the largest 13 combinations have been included. See explanation of plots in Fig. 4.4 (p. 35).*

Panels A-I in Fig. 4.5 (p. 38) indicate the overlaps found within the assembled dataset's data categories on eight traits (panels A-H), and the supporting information isolation source (panel I). All are measured on strain-level organisation with the standardised organism names as key. Notably, only the traits present within at least two datasets have been included. Thus enzyme activities, produced compounds, and incubation time from BacDive; extremophile, pathogenicity, and antibiotic susceptibility from TMD; metabolism from FAPROTAX; and genes from Pasteur are not included, as no overlaps with other data sources would occur.

Using the plots in the panels of Fig. 4.5, it is possible to infer the biggest contributors to any particular data category within the assembled dataset. For instance, panel G shows that JGI is the main contributor to gram stain with 25,089 reports, of which 20,434 are not covered by any other data source. Additionally, the plots indicate which data sources report information on the same data category for identical entries in the assembled dataset. In the instance of JGI and gram stain, almost 5,000 of its reports are covered by other data sources (approximately 25,000 total and 20,000 unique reports). Referring to the fourth column in panel G, it becomes apparent that 3,330 of these are organisms for which PATRIC also reports gram stain. If the assembled dataset is merged to yield only one entry per unique full organism name, ensuring that these reports agree should be a priority. Row merging and conflict resolutions between trait reports is regarded in the next Subsection.

In summary, 19 data sources were considered for inclusion in an assembled dataset of microbial traits. After data cleaning and standardisation, the ten largest sources (Fig. 4.1, 33) were joined to form a dataset with 147,676 strain-level entries. The dataset includes 17 trait fields and 14 columns of Supplementary information (such as taxonomy, identification, isolation source, and genome accessions). JGI and BacDive are its biggest contributors to organism coverage, and the latter is also prominent in most included trait fields (Fig. 4.2, p. 34). An overview of attribute frequencies for selected fields in the dataset (Fig. 4.3, p. 36) displays the great diversity of the recorded microbes, both in terms of taxonomy and phenotype. There are less overlaps between the sources than expected (Fig. 4.4, p. 37), with as many as 51,329 unique strains found in BacDive alone. The diversity of fields included in different sources is also accentuated (Fig. 4.5, p. 38): many data sources report on gram stain, oxygen requirement, and isolation source, while traits like trophy and doubling time have a smaller coverage, even within large data repositories.

4.1.3. Reduced dataset for genotype—phenotype association

To delimit the scope of the GPA methodology utilising the collected trait data, a copy of the assembled dataset was reduced to organism entries which met certain requirements. Additionally, only the most complete data categories were considered for inclusion in the reduced dataset: isolation source, growth temperature, oxygen requirement, gram stain, medium, and substrate. Both isolation source and growth medium contain several thousand different attributes; and the former is not strictly a microbial trait. Growth temperatures are reported as a mix of temperature ranges and discrete points. Thus the categorical values of oxygen requirement (with seven different attributes), gram stain (four attributes), and substrate (3,472 attributes) were selected for the reduced dataset and thus were candidate traits for the following GPA.

Of the 147,676 entries in the assembled dataset, 32,137 contain reports for at least two of the three selected traits. At this stage however, the dataset still contains multiple rows for many organisms. These originate from different data sources and thus may only report sufficient data on the organism when joined. For example, *Anaerobaculum mobile* (tax ID

97477) has four entries from different data sources, and none report on more than one of the traits of interest: two report gram stain, and a third on oxygen tolerance. Joined, the entries will however meet the requirement of at least two trait reports. Similar observations were made for 20,912 additional entries. By merging the rows reporting on the same organism, 3,402 additional entries met the trait coverage criteria. This effectively increased the sample size of organisms with sufficient coverage for inclusion in the reduced trait dataset that would be used for the GPA.

For substrate, varying attribute reports from different sources could be merged without further action, seeing as the ability to utilise one specific compound does not necessarily exclude the utilisation of other substrates. For gram stain and oxygen tolerance however, conflicts between reports may occur, and did for 178 and 2,971 organisms, respectively. With the conflict resolution based on majority and generalisation described in Chapter 3 (p. 27), 145 conflicts remained for gram stain (equal number of negative and positive reports), and 589 for oxygen tolerance (equal number of reports indicating oxygen tolerance and -intolerance). The gram stains were resolved by manual literature searches, while the oxygen requirement conflicts were marked as "conflict". This solution was adapted due to gram stain information being readily available, which was true to a lesser extent for oxygen requirements.

Of the entries in the dataset with merged rows and with at least two trait reports, 4,491 organisms had genome accession numbers with corresponding assemblies registered in the NCBI overview of sequenced microbes. Among these, 3,307 assemblies contained a FAA file. Mapping the assembly links and genome keys (1-3307) back into the merged dataset filtered for trait completeness, narrowed the selection from 4,491 to 3,409 organisms. However, given that there were only 3,307 unique accession numbers, there are organisms which are regarded as separate by standardised strain name that have identical accession numbers. This is true for 37 accessions in total, of which most are repeated once. The effect of this overlap will be further regarded in Chapter 5 (p. 58).

4.2. Genome sequence annotation

The previous section detailed the outcome of the assembled trait dataset and its reduced counterpart which were prepared for being utilised in the association of genomic contents with phenotypic traits. This Section considers the acquisition of this genomic content. Firstly, the choice of genome data and sources is presented, followed by the immediate impressions from the conducted functional annotation.

4.2.1. Genome sequence comparisons

Most genome assemblies have genome entries both in the GenBank (GCA) and the RefSeq (GCF) repositories of NCBI. The entries might be identical, but some vary greatly in terms of quality and completion. Therefore, a procedure was required to decide which assembly to use as the genome sequence source for each of the entries in the reduced dataset.

Comparing the FAA files downloaded from GenBank and RefSeq for each of the 3,307 genomes, yielded the results included in "comparison_GCA_GCF.csv" (Supplementary information 8, App. A, p. 85). Summarised, the genomes from GenBank contained more proteins than its equivalent from RefSeq in 67% of the assemblies. They were the same in 3% of instances, while the remaining 30% saw more proteins found within the RefSeq-acquired genome. The genomes for which the RefSeq protein count exceeded that of GenBank by at least 15%, have been included in Table 4.1 (p. 41).

Table 4.1: Comparison of GenBank and RefSeq genome assemblies. *Excerpt from "comparison_GCA_GCF.csv" with increases of at least 15% in favour of RefSeq.*

Key	Organism	Protein counts		Increase	
		GenBank	RefSeq	#	%
2882	*Altererythrobacter insulae* BPTF-M16	13	2,598	2,585	19,885
1302	*Corynebacterium aquilae* S-613	356	4,990	4,634	1,302
2530	*Corynebacterium frankenforstense* ST18	244	2,561	2,317	950
1565	*Corynebacterium sphenisci* 38	1,164	1,532	368	32
901	*Dietzia maris* DSM 44904	1,827	2,283	456	25
880	*Elstera litoralis* Dia-1, Nil	1,801	2,159	358	20
268	*Haloechinothrix halophila*	2,968	3,517	549	18
872	*Hankyongella ginsenosidimutans* W1-2-3	2,013	2,381	368	18
1027	*Hydrogenibacillus schlegelii* MA-48	3,346	3,934	588	18
3131	*Lawsonella clevelandensis* CCF-01, X1036	1,453	1,705	252	17
1080	*Nocardioides daphniae* D287	3,310	3,883	573	17
1388	*Paracoccus sphaerophysae* Zy-3	2,082	2,415	333	16
2317	*Prevotella oryzae* KB3	2,822	3,244	422	15
1354	*Sphingomonas jaspsi* TDMA-16	1,993	2,291	298	15
2124	*Thermodesulfobacterium commune* YSRA-1	2,853	3,276	423	15

For the instances included in Table 4.1, the number of proteins found within RefSeq-registered FAA genomes were reasonably higher than its GenBank equivalent. Thus for these 15 genomes, RefSeq was seen as the preferred source of genome sequence data. In all, 3,292 GenBank accessions (GCA) and 15 RefSeq accessions (GCF) were utilised to access and download microbial genomes submitted to the following functional annotation.

4.2.2. Functional annotation

After the selection of which assemblies to utilise as the genome sequence source for each entry of the reduced dataset, the genomes were submitted to eggNOG for functional annotation. Across the 3,307 annotated genomes, the terms within the classes COG, GO, KO were extracted. These were chosen due to their high coverage across the annotation files, and their facilitation of comparison of genome contents across species and genera. Fig. 4.6 shows the frequencies of terms within the annotation classes, in addition to the counts of unique and total annotations found across all genomes for each term class.

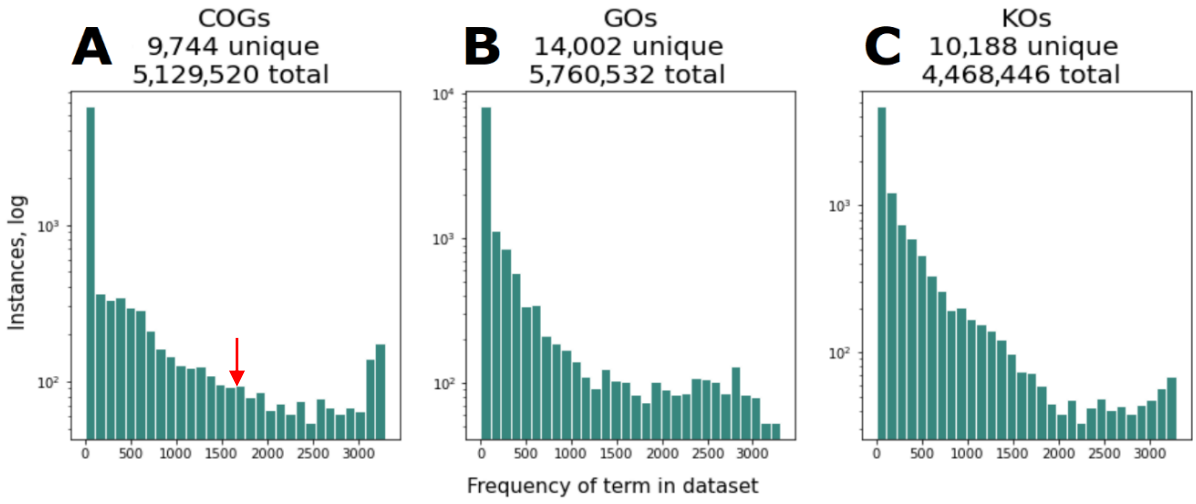


Figure 4.6: Term frequencies across annotated genomes*, from the three classes **A** COG, **B** GO, and **C** KO. Bar heights indicate how many terms belong in the frequency categories on the horizontal axis. These denote how many genomes a term is associated with out of the 3,307 possible. Example in panel **A:** about 100 COG terms have been associated with about 1,500 genomes.*

All bars within each panel in Fig. 4.6 (p. 41) constitute the total number of unique terms of that class, indicated above the panel. It can be derived from these data that COG was the least diverse annotation class with 9,744 unique terms distributed across 5,129,520 total annotations. KO terms were slightly more diverse but saw the fewest number of total annotations within the genomes. Lastly, GO terms were the most numerous class, both with respect to unique (14,002) and total (5,750,532) number of annotations found for the submitted genomes. Chapter 5 (p. 60) will see a more thorough investigation into the relative coverage and diversity of the three annotation classes.

The frequency distributions of all three annotation classes display similar patterns. The most numerous term categories (the highest bars) are those with low frequency across the 3,307 annotated genomes. This indicates that most of the terms within all three classes represent highly particular functions. As the frequency increases, the number of instances generally decreases. This pattern suggest that fewer terms are commonly found within several genomes, while still attesting the presence of more conserved functions. The negative correlation between frequency category and number of instances does not continue across the full frequency range, however. For COG and KO in particular, their final frequency categories see an increased number of instances. These could be representations of highly conserved functions, which would be expected to be found in many of the genomes. If they in fact are particular and thus rarer functions however, their high frequencies could indicate over-enrichment of specific features in the reduced dataset. As a consequence, the current selection of entries used for GPA might not be sufficiently representative to infer potential patterns between annotated genome contents and observed phenotypes. This matter will be discussed in Chapter 5 (p. 60).

Overall, the annotation term distributions in Fig. 4.6 (p. 41) suggest that the functional annotation of the 3,307 genomes yielded a variety of terms in all three annotation classes. Most terms are highly uncommon, as indicated by the prominent bars for low frequencies in all three panels of Fig. 4.6. The more common terms are at this point assumed to represent conserved features, which are expected to be found within many of the genomes. The functional annotation is thus believed to have yielded sufficiently specific genomic contents for the reduced dataset. The following Section sees the association of the two levels of microbial information acquired thus far: phenotypic trait attributes, and genomic contents in the form of genome annotations.

4.3. Genotype—phenotype association

This Section regards the association of the collected phenotypic features of gram staining with the annotated genomic contents of the 3,409 organisms in the reduced dataset. Only the annotations present in between 5% and 95% of the genomes were regarded for this purpose. This disregarded the rarest and the most conserved annotations, which are not expected to be of significant relevance. Thus 3,653 COGs, 4,718 KOs, and 5,100 GOs were associated to the known phenotypes of the organisms. The contingency tables within the result files "fisher_COG.csv", "fisher_GO.csv", and "fisher_KO.csv" summarise the distribution of organisms with respect to gram stain attribute (positive p or negative n) and whether they have (h) or lack (l) each annotation term. The resulting $\widetilde{OR}$ and p-values from the conducted Fisher's exact tests were used to determine which associations were statistically significant. Fig. 4.7 (p. 43) presents an overview of the number of significant associations found for each annotation class and the two gram stain attributes.

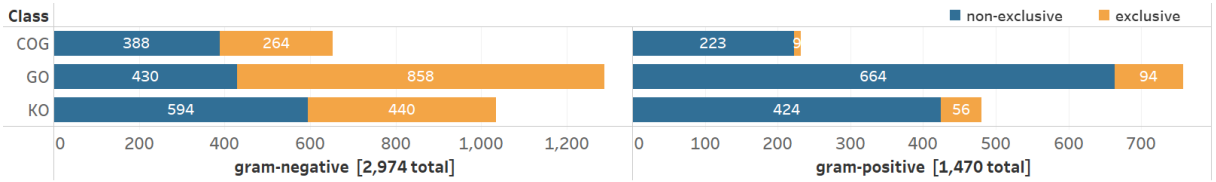


Figure 4.7: Distribution of significantly associated annotation terms *with the two attributes of gram stain "negative" and "positive", across the three annotation classes COG, GO, and KO. A term may have been found only for organisms with a particular gram stain attribute: these are called "exclusive" terms for that gram attribute.*

https://public.tableau.com/app/profile/jenny.merkesvik/viz/dataset_comparison/significant_count

In total, 4,444 terms were found significantly associated with either gram stain attribute. All annotation classes see significantly more annotations associated with gram-negativity, an observation which is regarded further in Chapter 5 (p. 63). Most of these terms are exclusives, meaning they have not been annotated in any genome from a gram-positive organism. It may thus be a candidate for unambiguously determination of gram attributes for other organisms. Possible reasons for this distribution are considered in Chapter 5 (p. 63) but nevertheless, the subsequent interpretations will not differentiate strictly between exclusive and non-exclusive associations. Another observations inferred from Fig. 4.7 is that for both gram attributes, COG sees the fewest significant terms, while GO is the most prominent. This is concurrent with the observations from Subsection 4.2.2 (p. 41), where GO was asserted as the most numerous annotation class.

The next Subsections regard the annotation classes one by one in order to discover possible patterns inferred from the terms found significantly associated with gram stain, and their immediate biological interpretations. The fourth and final Subsection will combine the impressions from each annotation class in order to examine the overall connotations of the conducted GPA. Lastly, an attempt is made to determine the gram stain attributes of a few organisms, based on their annotated genomic contents and the patterns inferred from the organisms with known gram attributes.

4.3.1. Clusters of orthologous genes

The volcano plot in Fig. 4.8 (p. 44) shows the association between annotated COG terms and the phenotypic feature of gram staining. Each non-grey mark represents a term which is found significantly associated with gram-negativity (red) or gram-positivity (blue). For instance, COG3599 has an $\widetilde{OR}$ of around 2.2, equivalent to an OR of 173.1. Based on the observations in the reduced trait dataset, the odds of an organism being associated with this term is 173.1 times higher when the trait of gram-positivity is given. This COG represents the cell division septum initiation protein DivIVA, which indeed is highly confined to gram-positives [126, 127]. DivIVA is involved in homeostasis of the peptidoglycan layer found in the gram-positive cell wall [127]: hence a gram-negative bacteria lacking this prominent layer of peptidoglycan may not depend on DivIVA to the same extent.

Another example is COG4105, representing the outer membrane protein assembly factor BamD, located in the upper left corner of Fig. 4.8. It has an $\widetilde{OR}$ of -2.7 and an OR of 0.002. Hence the odds of an organism being associated with BamD is 500 times higher when the organism is known to be gram-negative. A literature search reveals this protein's function in incorporation of β-barrel membrane proteins in the cell wall [128, 129], which is essential for the viability of the double membrane-enclosed gram-negative cell [130]. Thus the COG's association with gram-negativity is coherent.

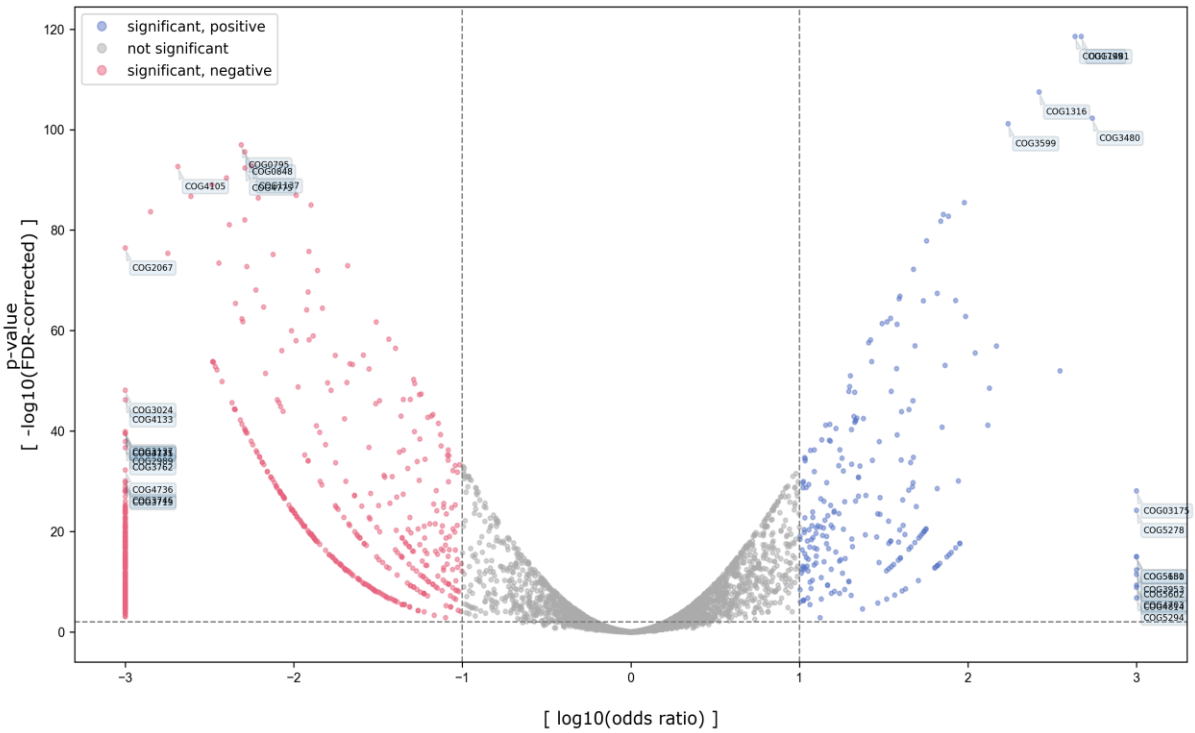


***Figure 4.8: COG terms significantly associated with gram stain**: negative (red) and positive (blue). Significance thresholds are set to 1.0 for $\widetilde{OR}$ [log_{10}(OR)], and 0.01 for FDR-corrected p-values. Any $\widetilde{OR}$ with magnitude 3 is manually truncated from infinite magnitude, representing annotation terms found uniquely for one gram stain attribute.*

In addition to COG terms strongly associated with a particular gram stain attribute such as the two previous examples, there are 273 COG terms that are exclusively found in organisms expressing a specific gram stain attribute. These terms are placed in the columns of $\widetilde{OR}$ with magnitude 3, which were truncated from infinite magnitudes to be included in the plot. Among these are seven terms exclusively found associated with sequences from gram-positive organisms, and 264 found exclusively in gram-negative organisms. The importance of these exclusive terms relative to the other data points is discussed in Chapter 5 (p. 63).

To facilitate the interpretation of the significantly gram-associated COGs, the terms were grouped into their parent categories for the visualisation in Fig. 4.9 (p. 45). Through Panels A+B and C+D, COG categories which differentiate gram-negative from gram-positive organisms may be inferred. Notably, the category "Function unknown" is the most prominent COG parent in all four panels. This may attest to the relatively limited vocabulary of this annotation class, a topic which is discussed in Chapter 5 (p. 60).

Overall, the parent terms suggest candidate functions which differ between gram-negative and -positive types of microorganisms. The cell wall structures used to differentiate between the gram attributes impose varying requirements for the cell, implying that cell wall biogenesis is an obvious feature which will deviate between the two types. Appropriately, "cell wall, membrane, and envelope biogenesis" ("Cell" in Fig. 4.9, panel A) is the biggest defined category within gram negative-associated COGs, containing 73 terms with an average $\widetilde{OR}$ of -2.091. Its children terms include the outer membrane lipoprotein LolB (COG3017), the lipoprotein subunit MlaA (COG2853), and the periplasmic

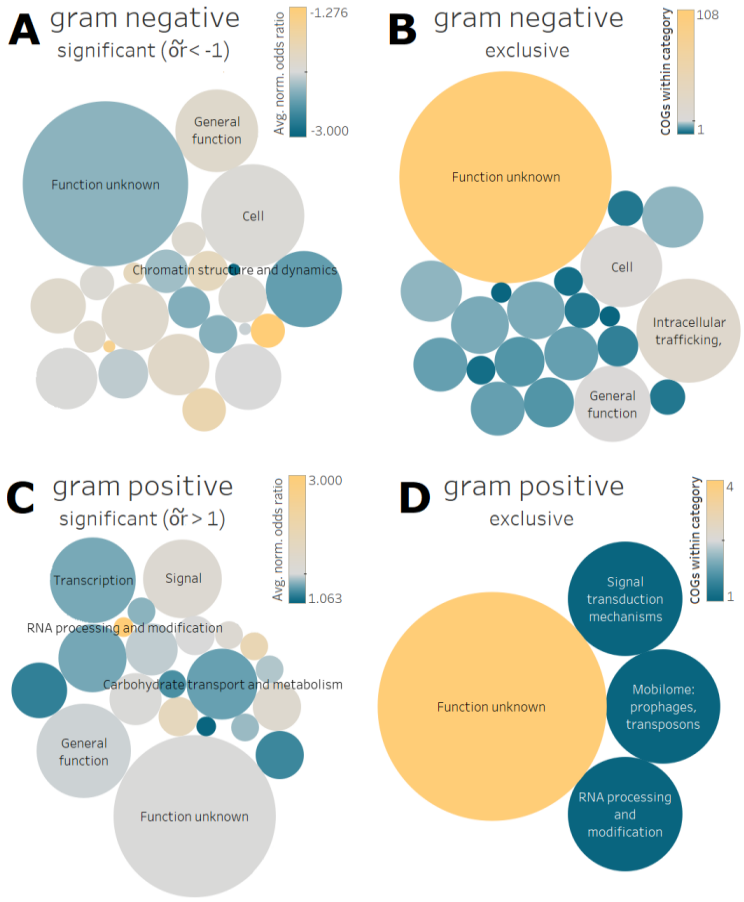


Figure 4.9: COG term categories significantly associated with gram stain.

A: *significantly associated with gram-negative organisms, comprises all reds in Fig. 4.8.*

B: *exclusively associated with gram-negative organisms, left-most reds in Fig. 4.8.*

C: *significantly associated with gram-positive organisms, comprises all blues in Fig. 4.8.*

D: *exclusively associated with gram-positive organisms, right-most blues in Fig. 4.8.*

*Panels **A** and **C:** circle sizes represent the number of COG terms within the category; colour indicates average $\widetilde{OR}$ of COG terms in the category, from blue (low) to yellow (high).*

*Panels **B** and **D**: both circle size and colour represent number of COG terms within the category, increasing from blue to yellow.*

https://public.tableau.com/app/profile/jenny.merkesvik/viz/dataset_comparison/COG_overview

subunit MlaC (COG2854); the latter two are both members of the ABC-type intermembrane phospholipid transporter Mla. For gram-positive organisms, the same parent category is present as a grey circle directly north-east of "General Function" in panel C. It contains only seven COG representatives however, with an average $\widetilde{OR}$ of 1.535. Its children terms include the spore coat protein CotF (COG5577), the poly-D-alanine transfer protein DltD (COG3966), and the anionic cell wall polymer biosynthesis enzymes TagV/TagU (COG1316).

Several other parent terms are found significant in both gram-negative and gram-positive associations. Examples are "intracellular trafficking, secretion, and vesicular transport" (right-most blue circle in panel A, and beige circle north-east of "Carbohydrate transport" in panel C), and "signal transduction mechanisms" (left-most beige in panel A, "Signal" in panel C). Their presence is immediately interpreted as appropriate, seeing as the differing cell wall structures in gram-negative and gram-positive cells necessarily will require different mechanisms for metabolite transport in- and outside the cell [131], for instance for vesicles to be created from or fuse with the cell membrane.

The category most strongly associated with gram-negativity (lowest $\widetilde{OR}$, thus coloured blue in panel A) is "chromatin structure and dynamics", representing COG5531: the DNA-binding SWIB/MDM2 domain. A literature search supports the observation of this domain only being found in gram-negative representatives of bacteria [132, 133]. For gram-positive organisms, the strongest associated COG (highest $\widetilde{OR}$, thus coloured yellow in panel C) is "RNA processing and modification" with COG5180: PAB1-binding protein. This

COG is however also found registered for gram negative-staining bacteria, such as the Bacteriodetes species *Tenacibaculum dicentrarchi* [134]. Despite *T. dicentrarchi* being present in the trait dataset as a gram-negative bacterium, its exclusion from the reduced dataset used for the annotation erroneously indicates that the COG is unique to gram-positive bacteria. A similar issue is seen for COG3953, representing the SLT domain protein and comprising the category "Mobilome" in panel D (p. 45). It is found annotated uniquely for gram-positive organisms in the reduced dataset, however it is also found present within the genomic content of five gram-negative bacterial species [135].

Overall, there are many COGs found associated with either gram stain attribute. Some associations appear reasonable given biological context, while others have been revealed as erroneous. Thus although some of the COG terms highlighted in the present work may represent features that could be used as indicators of particular gram attributes, close consideration of the association is needed to ensure its genuineness.

4.3.2. Gene Ontology

A volcano plot over the annotated GO terms is presented in Fig. 4.10. It displays 2,046 significant terms: 758 for gram-negative organisms, and 1,288 for gram-positives. The plot may seem to suggest that there are few GO terms significantly associated with gram-negative organisms relative to for gram-positives. This impression is adjusted by the consideration of the ratios vidualised in Fig. 4.7 (p. 43): most terms associated with gram-negativity do so exclusively and are thus populating the column of data points with $\widetilde{OR} = -3$. Exclusivity is seen for 67% of the gram negtive-associated terms, contrasting the corresponding 12% rate seen for gram positive-associated terms. The rate of 67% far exceeds those seen for any other attribute and term class.

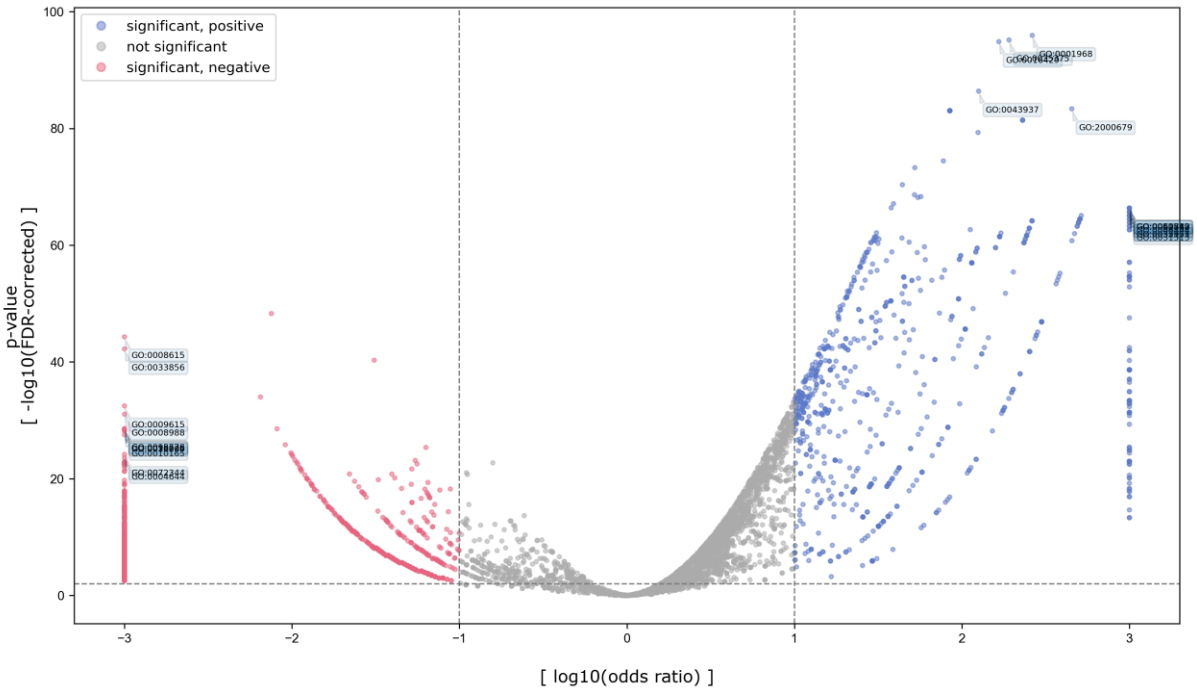


***Figure 4.10: GO terms significantly associated with gram stain**: negative (red) and positive (blue). Significance thresholds are set to 1.0 for $\widetilde{OR}$ [log_{10}(OR)], and 0.01 for p-values. Any $\widetilde{OR}$ with magnitude 3 is manually corrected from infinite magnitude, representing annotation terms found uniquely for one gram stain attribute.*

A comprehensive overview of the significant and exclusive terms for both gram stain attributes were produced by submitting the four lists of GO terms to REVIGO. Figs. B1-6 (App. B, p. 87) show the signifiant terms associated with gram-negativity and gram-positivity, respectively, in each of the three main GO hierarchies: BP, CC, and MF. Similar to the COG term parent categories in Fig. 4.9 (p. 45), the REVIGO plots cluster related terms and use circle sizes and colours to indicate the number of terms included within a category and the associated $\widetilde{OR}$.

An example of a GO category found significantly associated with sequences from gram-negative organisms is the BP "gram-negative-bacterium-type cell wall biogenesis" (Fig. B1, panel A, App. B, p. 87), along with related terms "membrane biogenesis" and "membrane organisation". Similarly, "gram-negative-bacterium-type cell wall" is a significant CC term as well (Fig. B2, panel A, p. 88). Thus there are several categories within cell wall structure and mainentance present in the significant GO terms, which aligns with similarly significant COG terms seen previously. Additionally, terms related to chromatin and DNA packacking is found exclusively for gram-negatives (Fig. B2, panel B, p. 88). These GOs may be related to the DNA-binding domains SWIB/MDM2, whose COG was exclusive to gram-negatives.

For gram-positive organisms, terms like "mycolate cell wall layer assembly" (Fig. B4, panel A, p. 90), "cell division site" and "spore wall" (Fig. B5, panel A, p. 91) all represent term categories which are likely to differ between cells with different gram attributes. Some terms indicating exclusivity however seem to do so erroneously. "Protein—pyridoxal-5-phosphate linkage" (bottom left, Fig. B4, panel A, p. 90) designates binding of the cofactor pyridoxal-5-phosphate (PLP) to various proteins. However, this function is not exclusive to gram-positive bacteria, but is in fact seen in all kingdoms of life [136]. Thus it may be inferred that despite the small size of this circle, it is the child term that is exclusively associated with gram-positivity, and not the more general function of PLP binding.

Overall, The REVIGO plots are more difficult to use for comparison of term categories across gram stain attributes. Most include a signinificantly higher number of circles even after related terms have been condensed into their parent term. This is likely due to the high number of GO terms and its loose hierarchy structure. This will be explored further in Chapter 5 (p. 60). Furthermore, the plots do not have an interctive counterpart like the Tableau-generated plots. Thus disables simple identifiction of which GO terms each circle represent. Further, this might lead to the issue of erroneously assuming that a narrow parent category is a single term which is significantly associted with a gram attribute. In addition to the protein—PLP interaction for gram-positives, a prominent example of this is "response to virus", found exclusively in gram-negative organisms in Fig. B1, panel B (p. 87). Its presence as a circle in this plot does not mean that gram-positive organisms do not respond to virus infections, but rather that there are significant differences between the processes involved in the viral response of the two bacterial types. Chapter 5 (p. 65) will further explore this discussion on the use of generalised parent terms for interpretation of characteristic features.

4.3.3. KEGG Orthology

The volcano plot of the association between KO terms and gram stain attributes is provided in Fig. 4.11 (p. 48). There are 1,514 terms significantly associated with either gram stain attribute: 480 for negative and 1,034 for positive. Examples of prominent gram positive-associated KOs are K09762 (top right, Fig. 4.11) for the cell division protein whiA, which is found conserved in most gram-positives [137]; and K09772 for the cell division inhibitor SepF, which indeed lacks a known homolog in gram-negatives [138]. On the other end of

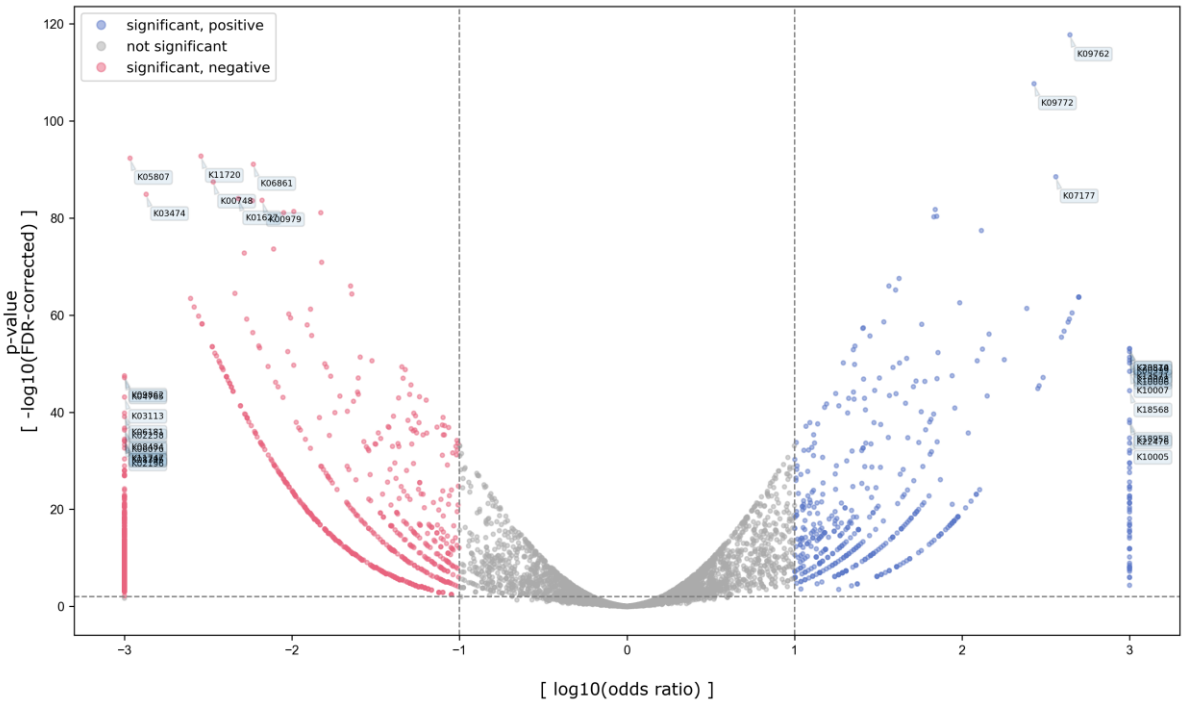


Figure 4.11: KO terms significantly associated with gram stain*: negative (red) and positive (blue). Significance thresholds are set to 1.0 for $\widetilde{OR}$, and 0.01 for p-values. Any $\widetilde{OR}$ with magnitude 3 is manually corrected from infinite magnitude, representing annotation terms found uniquely for one gram stain attribute.*

the spectrum is K05807 (top left, Fig. 4.11), representing BamD which was previously discovered as essential for gram-negative cell viability (Subsection 4.3.1, p. 43). K03113 is returned as exclusive to gram-negative organisms: it represents the translation initiation factor 1 EIF1/SUI1, which is also found in gram-positive organisms such as *Streptococcus mutans* [139]. Despite being in the reduced dataset as a gram-positive bacterium, *S. mutans* did not yield an observation for gram-positives associated with K03113. Similar cases are seen for other terms across annotation classes, and possible reasons for this are discussed in Chapter 5 (p. 63).

The significant and exclusive KO terms for both gram stain attributes were submitted to the KO Database mapper, organising the KOs into KEGG modules. The results of this mapping, found in its entirety in "ko_mapper.csv" (Supplementary information 14, App. A, p. 85), have been plotted with Tableau Desktop in Fig. 4.12 (p. 47). A total of 194 modules are represented by the KO terms significantly associated with gram-negativity (panel A), while its gram-positive counterpart (panel C) sees 72 modules. Of these are 73 and 7 exclusive to gram-negative and gram-positive entries, respectively, and have been included in separate panels (B and D, respectively).

One of the modules found exclusively for gram-negatives is "arginine succinyltransferase pathway" (marked "Arginine" in Fig. 4.12, panel B). The KO terms included in this module represent five enzymes encoded by genes *astA, astB, astC, astD,* and *astE* (K00673, K01484, K00840, K06447, and K05526, respectively). They are part of the *aru* operon, which indeed has no known homologues in gram-positive bacteria [140, 141]. For gram-positives, the modules "fatty acid biosynthesis, initiation" and "fatty acid biosynthesis, elongation" are both returned as exclusives (panel D). Their only participating KO is "fatty acid synthase, bacteria type" (K11533), which incorrectly suggest that only gram-positive bacteria express the enzyme [142]. Since it is the categories' only term, its rejection

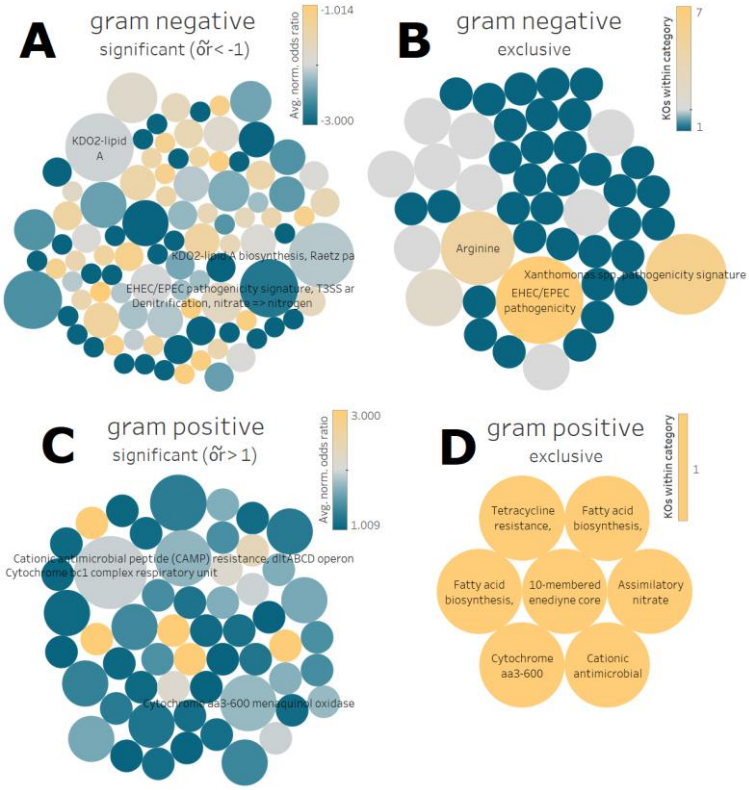


Figure 4.12: KO modules significantly associated with gram stain.

A: *significantly associated with gram-negative organisms, comprises all reds in Fig. 4.11.*

B: *exclusively associated with gram-negative organisms, left-most reds in Fig. 4.11.*

C: *significantly associated with gram-positive organisms, comprises all blues in Fig. 4.11.*

D: *exclusively associated with gram-positive organism, right-most blues in Fig. 4.11.*

*Panels **A** and **C**: circle sizes represent the number of KOs within the module; colour indicates avg. $\widetilde{OR}$ of KOs in the module.*

*Panels **B** and **D**: both circle size and colour represent the number of KOs within the module.*

https://public.tableau.com/app/profile/jenny.merkesvik/viz/dataset_comparison/KO_overview

eliminates both entire categories. Nevertheless, fatty acid biosynthesis is known to differ between gram-negative and -positive organisms [143, 144 p. 221]. How issues like this may still be resolved is regarded in Chapter 5 (p. 65).

4.3.4. Inferring gram stain attributes for new organisms

The previous Subsections saw the use of Fisher's exact tests to determine whether the presence of particular genomic contents was significantly associated with a given gram stain attribute. Across all annotation terms and both trait attributes, over 4,000 significant associations were found. Additional tools were used to group the significant terms to ease their interpretation. For several investigated instances, studies supporting the biological connotations of the terms' association with gram stain were found. This section seeks to gather the impressions across all three annotation classes (COG, GO, and KO) to see if the inferred patterns can be used to indicate the gram stain attribute of organisms with similar genomic contents.

Many significant terms found for both gram attributes seem to correctly suggest features which differs between the two cell types. As described in Chapter 2 (p. 9), both gram-negative and gram-positive cells have a cell wall consisting of peptidoglycan. However, this layer is generally much thicker in gram-positive cells. Gram-negatives are instead enclosed by an additional outer membrane. The significantly higher number of terms associated with gram-negativity may thus be due to this additional structure; a notion which is supported by the prominent presence of terms directly described as associated with the outer membrane and the larger inter-membrane space seen in these types of microbes. On the

other hand, many of the terms associated with gram positivity are related to surface proteins and enzymes involved in the biogenesis and maintenance of its prominent cell wall and peptidoglycan layer. With these general indications in mind, the annotated genomic contents of three random organisms in the assembled dataset can be generated in order to attempt to infer their gram status.

Gleimia coleocanis strain M343/98/2 (GCA_000159015) is registered without a gram attribute in the assembled dataset. When annotated by eggNOG, it is found to contain several of the terms mentioned previously in this Chapter, such as K09772 (cell-division protein SepF), and K09762 (cell-division inhibitor whiA). These terms have previously been found significantly associated with gram-positive organisms from the reduced dataset. *G. coleocanis* is therefore indicated to be gram-positive, which is true [145].

Another organism found in the dataset without registered gram stain is *Komagataeibacter xylinus* (GCA_004006375). Within its annotated genomic contents, the terms COG4105 and K05807 (outer membrane protein assembly factor BamD), and COG2853 and COG2854 (intermembrane phospholipid subunits MlaA and MlaC, respectively) are found. These are all indicative of the presence of an outer membrane, suggesting *K. xylinus* as a gram-negative organism: this is correct [146].

A final example is *Kyrpidia tusciae* strain T2 (GCA_000092905). Its annotations include COG3599 (cell division septum initiation protein DivIVA) and COG4105 (BamD), which have previously been significantly associated with gram-negative organisms. This is indeed the gram attribute of *K. tusciae* [147].

Thus with the GPA patterns derived in the conducted methodology, several annotation terms were recognised within the genomic content of three randomly selected organisms from the assembled dataset that did not previously have gram stain registered. Based on the known attributes of the organisms for which the same annotations were found, correct suggestions of the gram attributes of the three organisms could be made. Despite this demonstration exemplifying the use of the gathered data for inferring phenotypic features based on genomic content, particular care was taken when choosing which of the annotation terms to pursue. Among the 4,444 annotation terms found significantly associated with either gram stain attribute, several has been found to be misleading or directly erroneous in the introductory investigations conducted over the past three Subsections. The next Chapter will regard the patterns seen for these instances and attempt to discover the factors that may be causing them, and how they may be remedied.

5. Discussion

In this chapter, the conducted methodology will be reviewed in light of the results they have provided. The earliest interpretations of the analysis outcome will be expanded upon and considered along with other relevant areas of the present work. These matters will be presented and discussed corresponding to the sectioning utilised in previous chapters.

5.1. Microbial trait dataset

This section will regard the earliest efforts of the present work. It includes the assessment of microbial trait data sources with respect to the great variety between them; the selection of sources and traits to include in the assembled dataset; and whether the dataset fulfils its purpose of being a comprehensive and homogenous source of microbial trait data.

5.1.1. Data sources are highly variable

One of the main takeaways from the conducted work is the impression that current sources of biological data vary greatly with respect to many characteristics. Most prominent are differences in coverage, completeness, focus, format, and accessibility. These have all affected the approach needed for the utilisation of the datasets' contents.

Overall, datasets' organism coverage and category completeness seem to be negatively correlated measures. Moore, Nielsen, Mason, PhyMet[2], Kremer, and Campedelli are the smallest data sources considered in this thesis, and all but two have perfect completeness for all trait fields they report on. In contrast, the largest two data sources have at most 85% and 70% completeness, which drop to 43% and 56% respectively, when considering their second most complete categories. This difference is likely explained by the fact that the smallest data sources often are products of independent research studies, rather than being intended as repositories of microbial data. These studies either depend on or produce these microbial data and usually have a specific focus. Hence their high completeness. For instance, Kremer regards the temperature- and size-scaling of growth rates for phytoplankton and thus reports complete data on growth rates and temperatures. On the other hand, databases like BacDive and JGI collect data from such studies. They therefore contain observations from many sources on a vast number of different organisms and trait categories. Their large coverage and low completeness may therefore be said to symbolise the immense diversity of microbes, and how little is known about most of them.

Overall, the plots of Fig. 4.5 (p. 38) further accentuate the variation between the utilised data sources. Traits like gram stain, isolation source, and oxygen requirement are reported by at least half of the data sources. Others are less frequent and thus usually suffers a lower completeness within the assembled dataset. Doubling time and trophy are among the traits most scarcely reported on. Corkrey is the biggest contributor to both of these fields, which is the main reason for its inclusion in the dataset. Other sources have also reported doubling times [46, 86, 101], although these are all journal publications with scopes that cannot compete with the coverage found in databases such as BacDive, JGI, and IJSEM. Hence their exclusion from the assembled dataset. BacDive does report on microbial trophy, although contains fewer reports than Corkrey does for the same field: BacDive only overlaps with one of Corkrey's 660 trophy reports. This observation attests to the issue of microbial trait data being scattered across sources. With the ever-increasing number of scientific results being made available, the discovery and integration of data is a challenge repositories such as BacDive, JGI, and the present work must overcome in order to fulfil their purpose.

It should also be considered that even repositories like BacDive, which promotes itself as the largest database for bacterial information, is likely unable to gather all existing data on any particular organism. The sheer number of participants to microbial research alone poses a great challenge, and the additional factor of data availability further complicates this effort. Even within the relatively small number of data sources regarded in the present work, a great variety of approaches were needed to access them. Some were publicly unavailable (Campedelli, JGI, Moore, Nielsen) or in unfavourable formats (Bergey's, Corkrey, Mason, RefSeq), requiring them to be obtained through a secondary source (Madin *et al.* [44]). Some were available through command-line or programmatical approaches and required a lower (MediaDB, PATRIC) or higher (BacDive, FAPROTAX) degree of processing to assume an appropriate structure. Several were readily available for download, either through web interfaces (TMD, Pasteur, PhyMet^2, ProTraits) or as part of scientific publications (IJSEM, Kremer, Vieira-Silva).

Furthermore, no two datasets had the same structure or standards for reporting their microbial information, for instance by the use of trait ontologies such described in Chapter 2 (p. 15). An example is isolation source, which was reported using categorical tags (in IJSEM and ProTraits), non-categorical strings (BacDive, JGI, and Pasteur), or binary values (TMD). Hence some reformatting was required for all utilised data sources. Additionally, several data sources lack sufficient metadata. For instance, missing explanations of included traits resulted in the fields of *incubation time [days]* and *doubling time [hours]* being kept separate in the present work (see Fig. 4.3, panels T-U, p. 36). The connotation of "incubation time" is uncertain, as it for instance could denote both the time required for cell division, and the time required for the formation of a visible colony. If the former is true, then this field could have been combined with the field *doubling time*, increasing the completeness of this trait while removing the duplicate field. To avoid a possible erroneous merging however, no such reformatting was conducted.

In summary, the discovery and integration of various sources of microbial data is not a simple task. To facilitate the collection of biological data, effort is required both from the individual researchers producing the data, and the repositories seeking to gather and store the information. Examples for the former agent include making sure the data is accessible; follows set standards, for instance with respect to taxonomy and ontologies; and have sufficient metadata. With high quality data and metadata available, the repositories may exert their purpose of including new data sources by curation or submissions; maintain the database and its existing information; and further enhance the data's accessibility.

5.1.2. Creating a new trait dataset was beneficial

Having discussed the issues faced when accessing and standardising microbial trait data, a decision which should be mentioned was to not utilise any one existing source of microbial information for the GPA. As is evident from the data visualisation in Fig. 4.2 (p. 34), BacDive has asserted itself as a prominent data source both with respect to coverage and completeness of most trait categories. However, no data on gram stain was obtained from this data source. Despite gram stain being featured on its websites, BacDive does not readily list it as a checkable trait in its download section. Gram staining was regarded as a promising candidate for GPA at an early time due to its highly unambiguous nature and limited number of possible attributes (positive, negative, or variable). Hence BacDive not including it was unfortunate.

The work of Madin *et al.* [44] was of great advantage to the present work: they had gained access to datasets which were not publicly available; received corrections from original

authors; and made improvements to select datasets, for instance by updating organism names to current NCBI standards. However, neither the prepared trait dataset nor the automated pipelines for data preparation and assembly of Madin *et al.* were utilised in the present work. The main reason for this is the seemingly strict requirements imposed on data entries for inclusion in their dataset. An example is JGI: the raw data file utilised by Madin *et al*. consists of 280,750 strain-level entries, with just 12,083 entries in the resulting dataset ("condensed_species_NCBI.csv" in Madin *et al.* [44]) listing JGI as one of its contributing data sources. The present work utilised the same raw data and could extract 58,169 strain-level entries with data for least one of the relevant trait fields.

For data sources that are maintained and updated, such as BacDive, Pasteur, and TMD, the continued growth of the repositories is another factor discouraging the use of previously prepared dataset like Madin *et al*. No raw data from BacDive was provided in Madin *et al.,* however their final dataset counts 1,184 entries with BacDive as an origin. 82,892 entries were available at BacDive at the time of download for the present work (16 months after the publication of Madin *et al.)*. Of these, 51,431 entries had trait data of relevance for this thesis. BacDive continues to grow, and as of time of writing (May 5th, 2022), its repository has increased by an additional 6,653 strain-level entries since October 11th, 2021. This growth further demonstrates the ever-expanding knowledge gained within the field of microbiology, attesting the versatility of and need for effective tools and methodologies to gather, store, and maintain biological data.

5.1.3. Select data sources and traits were utilised

Of the 19 data sources first regarded, only ten were included in the assembled dataset of the present work. The selection was based the dataset coverage, and only the nine largest datasets were used. In addition, Corkrey was included despite being the eleventh largest source. It provides over 600 reports of growth rates exclusively for strains also found in other datasets: thus it improves completeness for this field without compromising that of others. The potential inclusion of the remaining nine datasets would have increased the total number of entries of 147,676 only by 2,654. RefSeq alone accounts for 1,727 of these. It only reports on the non-trait field isolation source, and for several hundred unique entries not found in any other database. Since the contributions of the remaining nine datasets would have been rather inconspicuous, and would have required additional efforts for standardisation, their inclusion was decided against. Still, they represent sources with high specificity and quality within their respective themes, such as growth conditions for methanogens in PhyMet^2; growth and antibiotic resistance of *Lactobacillus* in Campedelli; and genome projects in RefSeq. The latter resource was for instance used for accessing genome sequences in the second part of the present work.

Not all data categories or traits present in the source datasets were included in the assembled dataset. This was to prioritise efforts on traits which could be of particular interest for GPA. For instance, being able to suggest substrates for growth was regarded as of higher value than to suggest cell size. Thus limiting the number of traits to consider aided in setting the scope of the present work. Future projects may choose to include more data categories, for which the regarded data sources may still be good candidates.

5.1.4. Comparison of datasets relies on flawed identification methods

Being able to correctly identify entries on the same organism is of importance for the quality of any database. With erroneous merging of entries, the dataset might indicate incorrect traits for organisms, and lack appropriate records for others. Conversely, failing

to identify overlaps limits the accumulation of all known data for an organism, ultimately compromising the purpose of data repositories in their entirety.

To decide whether two entries within one or between datasets were of the same organism, most instances in the present work used string comparisons as the qualifier. This method was adopted due to many entries lacking an unambiguous identifier. An example could be NCBI taxonomy ID. In particular, the raw data files from neither BacDive, FAPROTAX, IJSEM, TMD, nor Pasteur contained taxonomy IDs, leaving 73,870 entries with the organism names as their only identifiers. Being the only field all dataset entries contained; the organism names were the only available identifier.

To facilitate the use of organism names for evaluating overlapping entries, they were standardised to omit imprecisions in naming conventions. These differences would not change the semantic of the organism names but would trigger a simple string comparison function into differentiating between the entries. Spaces, taxonomic level specifiers, and special characters were the main focus in this effort and eliminated the most evident variations in the name formatting. Still, the method is rather unforgiving, seeing as the presence of rarer taxonomic specifiers (e.g. serovar or biovar), special characters, or misspellings in the registered names would cause two entries in fact representing the same organism, to be regarded as separate. The reported overlaps between data sources indicate that the standardisation of organism names was successful to an extent. Still, the overlap is likely bigger than indicated in these figures (Fig. 4.4 p. 37, and Fig. 4.5 p. 38).

Other standards of organism identifiers which do not rely on exact string comparisons may be preferred. NCBI taxonomy ID might be a candidate, however not all entries within this identification scheme differentiates between strain-level entries of the same species. For instance, Bergey's, ProTraits, and JGI report four entries total with tax ID 24, none with identical trait reports. The first two report one entry each with the same registered organism name (*Shewanella putrefaciens*), but with no strain designations. On the other hand, JGI reports two entries including strain names, however these are not the same (ML-S2 and W3-6-1). Thus judging overlap by the relatively unforgiving string comparison method may be considered better than by tax ID in this instance. Ultimately, the present work highlights the benefit of adopting one particular convention for identifying microbial species. This would limit ambiguity and misunderstanding between those involved in microbial research, and any agent seeking to utilise the results of its efforts.

5.1.5. The dataset may be representative of the *known* microbial diversity

The present work has seen the gathering of data for select traits from ten prominent sources of microbial information, thereby assembling more data entries in total than its constituents. With this in mind, the assembled dataset might be thought of as representative for the diversity of microbes. For instance, of all data entries with reports on gram stain, just over 58% are reported as gram-negative (Fig. 4.3, panel D, p. 36). Could this suggest that approximately three out of every five microbial species are gram-negatives? Similarly, will 54% of microbes yield positive indole tests (Fig. 4.3, panel I, p. 36)? No such conclusions encompassing the entire microbial diversity will be drawn for any trait attribute distributions in the present work. This is due to the issue of sample sizes and observed bias in the collected data.

The number of trait reports for the different data categories in the dataset vary greatly. Fig. 4.2 (p. 34) illustrates that the features most commonly reported on are isolation source, growth temperature, oxygen requirement, and gram staining. Thus the statement that "about 60% of microbes are gram-negative" would be based on almost 47,000 data

records. In contrast, stating that "21% of organisms are lithotrophic rather than organotrophic" (Fig. 4.3, panel H, p. 36) is only based on 300 reports. Thus if any generalisations were to be made based on the collected data, the traits with the most reports would provide have the better foundation for such generalisations.

Despite a trait having sufficient sample size, it does not account for potential bias in its records. The isolation source is a prime example of this conflict. Being the data category with the most reports, it might be expected that it should be the best representative for any generalisations made using the assembled dataset. Inspecting the attribute distribution within the category accentuates the issue of bias in the dataset. 34% of the reports on isolation source contain the word "human". Despite the human microbiome being a vast community of species, it is unlikely that this high a proportion of all microbial species would be found in its constituents. This observation demonstrates the fact that many of these records are made based on their relations to humans, for instance by having been found in tissues or samples of patients during diagnosis and study of illnesses. Because these organisms seem to have a direct impact on humans, they are of particular scientific interest. Hence they have been studied and reported on more than other species.

Similar indications of bias can be seen in other data categories. Fig. 4.3, panel V (p. 36) indicates the most commonly reported growth temperatures for the recorded entries. There are slight indications of a bell curve shifted towards the lower end of the range, and with relatively few reports in the range from 50°C to over 100°C. There are however three prominent bars at 28°C, 30°C, and 37°C. These have at least double the frequency of surrounding temperature points. There is likely to be more microbes that can grow at 35°C than the data indicates, but the round number 30°C and the often referred to average human body temperature of 37°C is more commonly utilised in the reports of the data sources. Another testament to this bias is seen in the temperatures range 50-90°C. Here, any temperature with an increment of 5 are all more frequent than the temperature attribute right before it, such as 60°C versus 59°C. These observations are due to the fact that most data sources report specific growth temperature points, for example the exact growth temperature used in the reference work, rather than the full growth temperature range of the organism.

The assembled dataset is observed biased towards organisms and data reports of particular interest to humans. Thus it might be justified to state that the current dataset represents the diversity of the *currently known* microbes. As our perspective of microbial diversity expands and develops, additional information will be known and made available through data repositories. Distributions indicating the most commonly produced products or cell wall structures will develop along with these discoveries, and continuously change our perception of the microbial diversity surrounding us.

5.1.6. Strict criteria were set for inclusion in reduced trait dataset

Following the assembly of the microbial trait dataset, a selection of the entries was extracted into a reduced dataset for use in the GPA. To be included in the reduced dataset, an entry had meet two requirements regarding their reported data.

Firstly, the entry must have had non-NaN data entries for at least two of the traits of interest: gram stain, oxygen requirement, and substrate. As defined in Subsection 4.1.3 (p. 39), these were the traits with highest completeness of suitable formats. This criterium was based on the ambition of generating predictions based on genome annotations across all three traits. Additionally, only using entries which had at least two trait reports served as a way of reducing the number of genome annotations to consider to an amount more

appropriate for the scope of the project. However, gram stain was ultimately the only trait for which a GPA was conducted. With this in mind, the methodology could have included any organism with a gram stain report in the assembled into the reduced dataset. This would have increased the reduced dataset coverage by almost 2,000 entries. Such an increased sample size for the GPA could potentially have seen the elimination of erroneously indicated term significances or exclusivities, as mentioned in Subsections 4.3.1 p. 43, and 4.3.3. p. 47). Thus for future efforts, the trade-off between data management simplicity and sample sizes should be considered before subsets of data is disregarded.

In addition to requirements on trait reports, any entry in the reduced dataset must also have had an accession number which lead to a genome assembly within the NCBI servers GenBank or RefSeq. This approach is a direct consequence of the observation that data sources do not always report taxonomy IDs or unambiguous organism names. Had this been the case, the genomes would have been readily available by querying the NCBI database for their genomes. Instead, the reported genome accession numbers were the only readily available way of unambiguously connecting the dataset entries to their genome assemblies. The chosen approach is most limiting when considering that BacDive was the only source from which genome accession numbers were included. All entries within the reduced dataset are thus ultimately organisms that can be found within BacDive. Simultaneously, BacDive entries did not include gram stain reports. The entries must therefore be part of the overlap between BacDive and at least one of the sources reporting gram stain: IJSEM, JGI, TMD, PATRIC, and ProTraits. Firstly, this observation demonstrates the necessity for assembling a new trait dataset for use in the present work, instead of using existing datasets. Secondly, it tells that for at least 3,409 organisms (the coverage of the reduced dataset), the standardised names yielded merging between at least two similar reports from different data sources. These are assumed correct row merges due to the precision of the full organism names. Any of these entries thus contain more data on the included traits than its corresponding entries in any of the utilised datasets.

A consequence of having but one source report genome accession numbers is the immediate exclusion of over 25,000 entries which fulfil the trait count requirement, simply because they are not covered by this one source. Additionally, there is no way of corroborating the genome accessions reported by BacDive. For instance, one accession (GCA_900455645, key 2619) is found for 21 organisms in the reduced dataset. These are all entries of *Pseudomonas putida* with specified and unique strain names. However, a literature search for the accession number only returns the type strain of the species [148]. In this instance, there is no conflicting information in the trait reports for the entries with the same accession number. If this is the case for any other repeat accession numbers in the reduced dataset however, it could result in contradicting association reports.

Overall, lacking accession numbers is a hindrance to studying genomes and traits as two sides of the same coin, as there is no simple way to tell which two sides belong together. To facilitate GPA studies, efforts must be made to allow identification of organisms and linking of information across databases, seeing as repositories for trait and genomic information has historically been kept separate.

5.2. Genome sequence annotation

Having regarded the first main task of the thesis, namely assembling a dataset on microbial traits, this Section concerns the acquisition of their genotypic counterparts. Of relevance for this part of the project was to determine the source for genome sequences and how their contents could be annotated. Further, the immediate analysis of the returned

annotation terms yielded insight into the organisms within the reduced dataset, in addition to setting the stage for the subsequent GPA.

5.2.1. Both GenBank and RefSeq were utilised as genome sources

As discussed in Subsection 5.1.6 (p. 57), genome accession numbers included in the reduced dataset served as the bridge between the assembled trait data, and the organisms' sequenced genomes. Most of these accessions were directed towards GenBank, thus by mapping the accession numbers to the NCBI overview listing each accession number and its last updated assembly ("ncbi_prokaryotes.txt", Supplementary information 5, App. A, p. 85), direct genome download links to GenBank were retrieved. Some of the genomes acquired from this initial query yielded several particularly short genomes. Upon closer inspection, it was discovered that some contained as few as 11 protein sequences. According to the organisms' Nucleotide Database entries, the utilised GenBank accession was older than its RefSeq equivalent, while the latter reported a genome containing significantly more proteins. For another particularly small file however, the GenBank genome was newer, but the RefSeq genome was the utilised standard and contained more sequences.

This discovery initiated an investigation that saw the comparison of protein counts within corresponding genomes acquired from GenBank and RefSeq. To follow the reported genome accession numbers as often as possible, only genomes which saw an increase in proteins by at least 15% were updated to their RefSeq assembly counterpart. This way, the data source for the genomes would remain as consistent as possible, while still omitting the use of highly incomplete genomes. Future efforts requiring a mapping from accession numbers to FTP server links should seek out or query an updated overview file. If provided with multiple genome options, that containing the most proteins should be prioritised to yield annotations both of better quality and quantity.

5.2.2. EggNOG was the preferred tool for functional annotation

To highlight the genomic contents present within the collected genome sequences, three sources of functional annotations were considered. Firstly, most assemblies in NCBI's GenBank and RefSeq servers contain an annotation of the assembly genome ("*_genomic.gff.gz" or "*_genomic.gtf.gz" files). These would have provided quick and easy access to genomic characteristics through utilising a modified version of the pipeline made for downloading genomes from the same source. However, the NCBI-derived annotations do not currently contain information on orthology, which was required in the present work to enable the relation of features across taxa. Hence the need for a separate tool for functional annotation of the acquired genomes.

Two tools were assessed for the task of functional annotation. The first was Prokka (ver. 1.14.5) [149], utilising Orione [150], and accessed through the Galaxy server (ver. 21.09) [151]. The second was eggNOG, which for the purpose of tool testing was accessed through its web interface [152]. Both tools report COG, however they did not yield the same annotations when given the same input genomes. The full tests and outputs are included in the notebook "annotation_tool_comparisons.ipynb" (Supplementary information 9, App. A, p. 85), while Appendix C (p. 95) contains an excerpt of the test results. Overall, eggNOG took slightly longer to run but yielded significantly more COG terms than Prokka for all four test species. At most, Prokka had eight COGs not reported by eggNOG, while eggNOG had 218 unique entries at the least (see Fig. C1, App. C, p. 85). Thus eggNOG was chosen as the tool for functional annotation. Instead of utilising the web interface however, the annotations were conducted with the same settings using a command-line approach.

A notable observation with the eggNOG annotations is the high number of COG terms used by the tool (included in the eggNOG raw data (ver. 5) [153]). In the present work, eggNOG annotated 9,744 unique COG terms across all submitted genomes. However, only 4,877 unique terms are listed within the official COG database [67]. Thus any additional orthology terms in eggNOG that does not follow the COG standard was disregarded in the analysis, due to no COG category being mapped to this genomic content.

5.2.3. Annotation classes vary in coverage and diversity

EggNOG returns several annotation classes and for the present work, three were regarded: COG, KO and GO. The former two are orthologies, thus annotation genes with the same function originating from a common ancestor in different microorganisms. Additionally, GO is a well-established resource for computational representations of biological concepts and can also be used across taxa. It was thus included despite not being a true orthology.

The overview of term diversity presented in Subsection 4.2.3 (Fig. 4.6, p. 41) reveal that COG is the least diverse annotation type; KO has the fewest total annotations; and GO is prominent in terms of both total and unique annotations. A probable explanation for this observation is explored in Subsection 5.2.3 (p. 60). Concerning the number of total annotations, these observations are expected due to the annotation output only including the lowest possible level within the annotation hierarchies for both COG and KO. This yields the most particular data for the association of traits with genomic contents. For GO terms however, each row in the annotation results could contain the full path from the lowest possible hierarchy level back to a top node (BP, CC, or MF). This explains the high number of total and unique GO annotations.

To avoid this formatting difference from affecting the frequency distribution, only the most specific GO terms have been extracted from the annotation results. However, not all entries followed one path in exact order back to the origin node. For instance, one annotation for the test species *Cellulomonas fimi* NRS133 listed over a hundred terms, starting with "GO:0000166, GO:0003674, GO:0003676, (…)". Assessing the placement of these three terms in the hierarchy returns the relative positions "2, 1, 3", and the middle term is in fact an origin node, MF. Thus the returned terms cannot be assumed to be in order for the simple extraction of the most specific term. To extract only the most specific term for each annotation, an additional tool would be needed. This matter is partially solved by the generalisation of GO terms seen in Figs. B1-6 (App. B, p. 87) and will be discussed further in Subsection 5.3.5 (p. 65).41

Overall, the frequency distributions of the annotation classes (Fig. 4.6, p. 41) showed a decreasing number of instances for increasing frequencies, however not across the entire frequency range. This demonstrates the presence of larger groups of terms that are found in almost all of the annotated genomes. If these represent conserved characteristics, such as housekeeping functionalities required by all cells for survival, they are expected to be found in most genomes. If they represent some rarer characteristic and are still found among most of the 3,307 annotated genomes, it will demonstrate that the reduced dataset is over-enriched for this function. Thus the dataset may not be considered representative, and the quality of any results provided by its use will be compromised. Hence it is of relevance that these terms found in high frequency within the annotations, represent common microbial characteristics.

The most common COGs are COG0454 and COG0142, with respective frequencies 3,290 and 3,288: these represent enzyme families which indeed are found within all kingdoms of life [154-156]. Their high frequencies are therefore coherent with their lack of specificity.

Similar observations are made for the most frequent GO terms: GO:0003674 (frequency of 3,306) represents "molecular function", one of the three top nodes in the GO hierarchy [157]. GO:0044237 (frequency of 3,305) is "cellular metabolic process", a term only two edges away from the top node BP [158]. For KO, a similar pattern is seen. K02988 and K02892 are the most common terms with frequencies of 3,282, and both represent ribosomal subunits [159, 160] which are highly conserved functions needed for survival by all living cells.

In order to circumvent the consideration of annotation terms which where unlikely to yield any significant associations when related to gram staining, any term present in less than 5% or more than 95% of the genomes were removed. Since the reduced dataset contained about a 2/3 ratio of gram attributes in the favour of gram-negativity, a term present in very few (less than 166) or very many (more than 3,141) genomes were considered less relevant for the GPA. Thus following this exclusion, the genomic contents used for the GPA are believed to represent a sufficient diversity of features. The terms for the most particular and most conserved features have been disregarded, leaving three sets of annotation terms which may further be attempted associated with observed phenotypic attributes.

5.3. Genotype—phenotype association

Following the acquisition of microbial features on both the phenotype and genotype level, this Section will regard the analysis of the association between the functional annotations and the collected trait attributes. Themes include the choice of gram staining as the trait and Fisher's exact tests for the association; the biological interpretations of inferred patterns, and their use for inferring phenotypic attributes for new organisms.

5.3.1. Gram staining was chosen for genotype—phenotype association

The reduced dataset for which functional annotations had been generated, contained three traits in total: gram stain, oxygen requirement, and substrate. Gram stain was prioritised for the GPA for three reasons. The first being its high completeness relative to other trait fields in the assembled dataset. It was reported by several data sources and thus would provide a better sample size of organisms for the GPA. Secondly, because it is a common way of distinguishing between two types of microorganisms. The main features separating the two are well studied, thus any patterns indicated by the analysis results could be substantiated or discouraged by literature searches. Lastly, the trait of gram staining was easily divisible into just two attributes: positive or negative. Only having two possible attributes was beneficial because it would only require the creation of one contingency table for each annotation term. The third possible attribute for gram stain was "variable", however it was present in less than 1.3% of instances. It could therefore be disregarded without compromising the sample size. Dividing the reduced dataset based on this one trait thus yielded two groups of organisms of sufficient size.

Still, other trait fields can be chosen for a similar GPA approaches given slight modifications to the conducted methodology. For instance, the trait of oxygen requirement has seven possible attributes in the present dataset when excluding the "conflict" attribute. In its potential GPA procedure, a contingency table would be created for each attribute to divide the dataset into groups based on its presence in the reported phenotypes. For instance, the reduced dataset would see groups comprising of 116 *with* and 3,239 *without* the attribute "obligate aerobe"; 266 with and 3,089 without the attribute "microaerophile"; and so on. A Fisher's exact test would be conducted per attribute per annotation term, yielding output OR and p-values for each association. Ultimately, the analysis and biological

interpretation would consider seven times the number of data points compared to the current gram stain analysis, due to the increased number of attributes. The interpretation effort could be aided by similar term generalisation approaches as utilised in the present work. Alternatively, the attributes themselves could be generalised to yield fewer contingency tables and parameters to regard. This would however yield a trade-off between overview and particularity, as is regarded in Subsection 5.3.5 (p. 65).

Nevertheless, the trait of oxygen requirement is still considered a future candidate for similar GPA for the discovery of patterns between genomic contents and known phenotypes. The same is true for other trait fields, such as substrate. This trait sees an even higher number of possible attributes (944 in the reduced dataset) however, which would yield a prominent number of contingency tables and parameters to regard in the subsequent analyses. The generalisation scheme for these traits could therefore greatly benefit from the generalisation of the utilised attributes. This could be achieved by a trait or compound ontology, providing a standardised framework for reporting trait attributes.

5.3.2. Fisher' exact tests facilitated biological interpretation

The Fisher's exact tests used for the current GPA investigated the null hypotheses stating no dependence between the presence of any particular annotation term and any particular attribute of gram staining. Each test yielded an OR and a p-value. The OR denotes the odds of a particular outcome (having an annotation term "x", denoted $\mathtt{h}$) given a prior exposure (gram attribute "positive", denoted $\mathtt{p}$) compared to the same outcome in the absence of the exposure (gram attribute "negative", denoted $\mathtt{n}$) (see Section 2.3.3, p. 13). For each test, the OR was normalised by $\log_{10}$-transformations, after which they were denoted $\widetilde{OR}$. The absolute value of $\widetilde{OR}$ represents the magnitude of the association, while the sign indicates its direction: negative values are associated with the absence of exposure $\mathtt{n}$, and positive values with the presence of exposure $\mathtt{p}$. Thus $\widetilde{OR}$ is a more intuitive measurement of the association and was therefore used for the remaining analyses.

An additional change introduced to $\widetilde{OR}$ was the truncation of any values of infinite magnitude to that of 3. The calculation of OR (Eq. 1, p. 13) requires divisions involving the number of occurrences of the four possible instances ($\mathtt{ph}$, $\mathtt{nh}$, $\mathtt{pl}$, and $\mathtt{nl}$). If an annotation term was present only within entries with one particular gram stain, two of these four values would equal zero and yield divisions equalling infinite magnitudes. To prevent these terms from being excluded in the Volcano plots (Figs. 4.8 p. 44, 4.10 p. 46, 4.11 p. 48), they were instead set to a particular threshold value outside the range of any non-infinite $\widetilde{OR}$. This way, they would still be included in the plot but be kept separate from the remaining terms. They all have magnitudes of 3 but may have different p-values. Therefore, they form two vertical columns in each end of the x-axis in the Figures. For p-values, FDR correction was utilised. It seeks to limit the number of erroneous rejections of the null hypotheses during multiple testing [61]. This would apply to any p-value $\neq 1$.

The significance level set for $\widetilde{OR}$ was 1.0. This implicates that for any association between a term and a particular gram stain, the magnitude of the odds ratio must be at least 10. For p-values, the significance level was set to 0.01. Overall, this means that a GPA between a term and a gram stain attribute is considered significant if the odds of an organism being annotated with the term in question is 10 times higher for one particular gram stain attribute than another; and that there is at most a probability of 1% of obtaining a similarly extreme distribution of observations if no association was present (i.e. the null hypothesis is true). This was regarded as a sufficiently high indicator of significance for the results. However, some observations and investigations of particular instances does reveal

erroneous associations within the results. These will be explained and discussed in the next Subsections.

5.3.3. Thousands of annotation terms have significant gram associations

Across the all three annotation classes, 4,444 annotation terms were found significantly associated with either gram-negativity or gram-positivity. Most terms were associated with gram-negativity, and over half of these were found exclusively for this gram attribute. This Subsection regards the observed diversity and extent of each annotation class, aided by the overviews presented in Fig. 4.7 (p. 43) and the Volcano plots in Figs. 4.8 (p. 44), 4.10 (p. 46), and 4.11 (p. 48).

Asserting itself as the most numerous and diverse annotation class before the association analysis, GO unsurprisingly appears as the most prominent annotation class among the terms with significant association to either gram stain. It could be attributed to the fact that it is a well-established representation of biological concepts with over 40,000 valid terms. In comparison, the NCBI COG database spans just under 5,000 terms [67]: hence its limited coverage. KO however consists of over 58,000 unique terms at its lowest level, while still seeing fewer total and unique annotations than GO. The notion that GO is more commonly utilised than COG and KO may result in these terms having relatively more established methods for annotation. Still, the prominence of GO is most likely due to many general terms being included in this class, as was discussed in Subsection 5.2.3 (p. 60). The effect of structural differences between the hierarchies will be regarded further in Subsection 5.3.5 (p. 65).

Nevertheless, all three annotation classes provided several hundred annotation terms of interest for investigation. A selection of terms was found exclusively in genomes from organisms with a specific attribute of gram staining. These exclusive terms were most prominent for gram-negativity, surpassing even the number of non-exclusive terms for the same attribute. Generally, there seems to be more annotations known for genomes of gram-negative organisms. The total number of significantly associated terms are about twice as high for gram-negativity than gram-positivity, corroborating this assumption. This might be seen in light of the fact that there are more gram-negative organisms included in the dataset overall. As discussed in Subsection 5.1.5 (p. 56), any data repository is biased towards the organisms and data categories of interest in the scientific community. If more entries on gram-negative organisms are present, they might represent more popular subjects for study; hence the prominent number of genomic contents identified within their genomes. Inspecting the exclusive annotation terms however, reveals another possible explanation: many regard the biogenesis and structure of the gram-negative cell wall. Considering that gram-negatives contain a second phospholipid bilayer with several additional compounds (see Fig. 2.1, p. 9), these many extra terms relative to gram-positives might be due to this additional cellular feature.

5.3.4. Claims of attribute exclusivity are most uncertain

With the described use of OR to denote the magnitude of association between annotation terms and gram attributes, the group of terms were effectively divided in three: the exclusive, the significant, and the non-significant terms. The exclusive terms, only found annotated in organisms with a particular gram attribute, might appear as the most promising candidates for relating genomic contents to that attribute.

If these terms are true exclusives, this might be true. They could represent genomic content that directly contribute to the particular phenotype, hence their absence from

genomes of organisms expressing the opposite attribute. The present analysis could thereby have unambiguously identified genomic contents which could be used to predict the gram stain of an organism displaying the same genomic contents, without requiring cultivation and an actual gram staining procedure. Furthermore, any exclusive term may be of interest even if their implied causality for the current attribute of interest is refuted. Because of the low significance probability used in the tests, it is unlikely that a similarly extreme distribution would occur by coincidence. Hence the sequence represented by this term is likely of importance; but perhaps for another phenotypic feature present within the same group of organisms.

An example of this is the term K11533, representing the bacterial type fatty acid synthase. It was found exclusively associated with gram-positive organisms (Fig. 4.12, panel D, p. 49), but this association was refuted as in Chapter 4 (p. 46). Still, it is known that fatty acid biogenesis differs between gram-negative and gram-positive entries [143, 144 p. 221]. This means that there must be other terms present in the overview that account for the phenotypic difference, or that these annotations have been missed. A search for fatty acid-related KOs in the overview returns three additional terms: an enzyme needed in biosynthesis for the gram-negative outer membrane (K16363) [161], and two specialised proteins involved in unsaturated fatty acid synthesis in gram-negative α- and γ-proteobacteria (K01716, K00647) [162]. Hence there are still annotations present in the overview that may account for the differences known to exist between gram-negatives and -positives for this cellular function. Continued GPA studies could thus be attempted in order to identify the features which explain the observed extreme distribution.

A term's status as exclusive is highly susceptible to change, however. If the term is observed but once for the other gram attribute, it would be demoted to a non-exclusive term. It is likely that it would still be significantly associated with the same gram stain for which it was assumed exclusive. Hence the biological interpretation could remain similar. Still, this issue underlines that the presence of attribute-exclusive terms for any particular annotation class does not justify the disregard of its non-exclusive terms. After all, the exclusives are but one observation away from joining this group. As a consequence, significant terms should be included in the interpretations of the GPA whether or not they are found as exclusive.

In the present work, this mindset let to the creation of two sets of plots for each annotation class and gram attribute. Thus even if a term's status as exclusive is refuted, it would still be included as a non-exclusive, significant term in a similar term generalisation scheme and be available for interpretation. By only using these visualisations for inferring association patterns, interpretations can be made without the risk of false attribute exclusivity. This however assumes that the inclusion of potentially overlooked observations does not compromise the significance of the association. For instance, COG3722 is a term significantly associated with gram-negativity. It represents a DNA-binding transcriptional regulator in the MltR family and is thus generalised into the "Transcription" parent term (bottom beige circle in Fig. 4.9, panel A, p. 45). Its contingency table reveals the small sample size this result is based on: $\big[[1,19],[221,334]\big]$, meaning one gram-positive with the term; 19 gram-negative with the term; 221 gram-positive without the term, and 334 gram-negative without the term. Due the first observation in the table, the COG is not returned as exclusive to gram-negative organisms, but it is still significantly in favour of this attribute. If an additional observation of this COG was found for a gram-positive organism (yielding the table $[[2,19],[221,334]]$) however, the outcome of the Fisher's exact test changes. The association $\widetilde{OR} = -1.099 < -1.0$ would change to $\widetilde{OR} = -0.798 \not< -1.0$, which is

not considered significant. For instance, COG3722 is also found in the gram-positive bacterium *Sphaerobacter thermophilus* , which is included in the reduced dataset but lacks its gram attribute. Thus due to the small sample size for this term, the addition of just one observation changes an indicated GPA from exclusive to non-exclusive, and possibly erases its significancy entirely.

The term K06181 represents the 23S rRNA pseudouridine synthase *rluE* and was found exclusively for gram-negative entries in its association (Fig. 4.11, p. 48). Kim *et al*. [139] however reported the presence of *rluE* within the *relQ* operon in the gram-positive bacterium *Streptococcus mutans* in 2012. Surprisingly, this species is included in the reduced dataset as gram-positive, facultative, with 15 registered growth substrates, and with the accession number GCA_009738105 (key 2941); but the KO term for *rluE* is not annotated within its genome. One explanation could be possible errors within the utilised genome sequence, preventing its correct annotation. It is however reported as the complete representative genome [163]. Hence substantial errors within the genome sequence are unlikely. Alternatively, eggNOG could have failed to annotate parts of the submitted genome. Kim *et al*. state that inconsistencies are often found for the *relQ* operon in various databases: most importantly that *rluE* of *S. mutans* often is incorrectly annotated as *rluD* [139]. Reviewing the raw annotation file for this genome yields the discovery of three *rlu* elements, which does include *rluD*. It is termed by eggNOG as K06180, a KO which is not significantly associated with any gram attribute. Thus the KO term originally returned as exclusive to gram-negative organisms seems to be an incorrect annotation of a common feature seen for organisms of both gram attributes.

This latter example demonstrates the presence of inconsistencies within the field, and that these may persist for long periods of time without correction. Any attempt at interpretation and biological connotation utilising the presented results or similar methodologies must be done with care. Thorough investigations should be conducted to ensure that the indicated association may be genuine, and not a product of small sample sizes, ambiguous annotations, or other inconveniences.

5.3.5. Term generalisation yields trade-off between overview and specificity

To facilitate the interpretation of the GPA between genome annotations and gram stain attributes, terms within each annotation class were generalised by assembling related annotations into their parent terms or broader categories. The illustrations of these categories (Figs. 4.9 p. 45 for COGs; B1-6 p. 87 for GOs; and 4.12 p. 49 for KOs) convey both the number of terms included within each parent, and the average $\widetilde{OR}$ of their constituents. By continuing the measure of significance into the generalisations, the parent terms may still be used to infer what genomic content appears associated with either gram attribute. For instance, they successfully represented several features involved in the biogenesis of the outer membrane of gram-negative cells (see Subsection 4.3.1, p. 43). However, the generalisations also revealed potential issues with the produced results and may themselves also contribute to erroneous interpretations.

The level of generalisation should be considered when they are used to summarise biological information. For all hierarchies, a concern may be the loss of detail and thus potential intuitive GPA patterns. Other hierarchies, namely those where a child can have several parent terms, could additionally face the challenges of duplicating terms into several parents, effectively increasing the number of terms to consider rather than achieving the intended generalisation. In the instance of K11533, one round of generalisation resulted its duplication into two parent modules (M00082 and M00083). This

could be remedied by further generalising these modules into *their* common parent, thus eliminating the redundancy of K11533. Simultaneously, this solution will see the loss of even more detail. Thus the right level of generalisation may be difficult to determine. It should nevertheless be sought in order to achieve an appropriate balance between overview and detail.

Similar to the hierarchy of KEGG Modules, GO terms may also have several parent terms. Additionally, the GO hierarchical structure is described as "loose", for instance meaning that term A can be a direct parent of terms B and C, while term B also is a parent of C. This structure further complicates the clustering of related terms, which is likely the reason for the low level of generalisation seen for the GO plot overviews in Figs. B1-6 (p. 87). Most terms seem to have been grouped into parent terms. For instance, the highly general term "chromosome" is found in Fig. B2, panel B (p. 88). Conversely, terms like "bacterial type flagellum stator complex" (small, red circle in Fig. B2, panel B) is presented as more specific terms both by name and the visual cues of the circles. The REVIGO plots suffer particularly from this varying degree of generalisation due to the results not indicating which GO terms are contributing to the circle sizes. Hence no closer inspection of the terms, like that seen for COG and KEGG modules, are possible. There are alternative tools that may yield better generalisations of GO terms, with examples being GO Slim Term Mapper [164] and Generic GO Term Mapper [165]). For the sake of exemplifying the utilisation of exiting microbial data in this thesis however, the current generalisation provided sufficient overview and contributed to the general indication of candidate genomic contents for suggesting gram attributes.

The unambiguous categorisation seen for COG contrasts that of both GO and KO. All COGs were divided into just one of 26 possible categories, hence their visualisation in Fig. 4.9 (p. 45) yielded the most concise overview of all three annotation classes. A caveat of this structure is however a high level of generalisation, so much so that there is a difference in just three categories between significant gram-negatives (Fig. 4.9, panel A) and significant gram-positives (Fig. 4.9, panel C): "cell motility", "chromatin structure and dynamics", and "extracellular structures". Since these categories are highly general, investigation of their constituent COGs is necessary to understand the differences indicated between gram-negatives and -positives. Since there are relatively few COG terms in general, this task is not as laborious as it would have been for GO and KO. Some groups of terms may be relatively intuitive. For instance, six of eight COGs within the "cell motility" category contain "flagellar biosynthesis" or "flagellar basal body" in their definitions, making their function easily inferable. Such a clear sub-categorisation might not be present to the same extent within "extracellular structures", whose contributing COGs are "chaperone PapD" (COG3121), "fimbrial subunit ScuA/B" (COG5430), and "pilus biogenesis protein CpaD/CtpE" (COG5461).

Another notable observation from Fig. 4.9 (p. 45) is that the COG category "Function unknown" is the most prominent in all four panels. It may further accentuate the COG system as less established than its KO and GO counterparts. Additionally, considering the descriptions of the COGs found within the "unknown function" category reveals that the interpretation could have benefitted from a less extensive generalisation than the 26 COG categories can offer. For the 188 gram negative-associated COGs with unknown function, 33 are defined as membrane proteins. Another five terms have definitions containing "periplasm". Thus with additional distinctions within the COG categories, several results would have been more readily available for interpretation of their biological connotations.

For the present work, the highly general categorisation of COG terms did not infer significantly with the utility of the generalised plot of Fig. 4.9 (p. 45). It could still be used to indicate functional areas of interest for the consideration of features differing between gram-positive and gram-negative organisms. However, if other analyses see the inclusion of significantly higher number of COG annotations, inferring a general overview of functions they represent will likely be more laborious.

6. Conclusions and outlook

The final chapter of this thesis seeks to summarise the essential takeaways from the present work. It will firstly address the creation of a dataset of microbial traits and the acquisition of functionally annotated genome sequences. This is followed by the association of genomic contents with observed attributes of gram stain, whose patterns were ultimately utilised to predict gram attributes for other organisms present in the dataset. In light of the results of these Sections and the points raised in the discussion conducted in the previous Chapter, tentative conclusions are formed with the intention of addressing the thesis aim of exemplifying uses of existing microbial data and annotated genomic contents to suggest phenotypic traits. Lastly, potential improvements and further areas of interest will be suggested for the hypothetical continuation of the project or its aim.

The conducted methodology saw the consideration of 19 sources for microbial data. They were highly diverse in terms of coverage, theme, format, and completeness. As a consequence, particular procedures were needed to access and manage the data from each source. By enforcing existing terminology and custom standardisations of data categories, data such as organism names, gram stain, trophy, and growth temperatures was homogenised. After assuming similar structures and formats, comparisons of the datasets' content could be used in the decision of which data sources to move forward with.

Ten sources were selected based on their coverage and completeness, yielding an assembled dataset of 147,676 entries, spanning 126,763 unique strain-level organisms and 25 data categories. The variation seen in the individual data sources is apparent in the assembled dataset as well, manifesting itself in terms of great diversity of data fields and their varying completeness. For instance, seven sources report on oxygen requirements, while only two contribute to the field of doubling time. The completeness may be greater than its current impression indicates, a limitation attributed to the chosen structuring method. In the assembled dataset, entries from different data sources have not been merged. Some organisms are therefore found represented by several rows in the dataset. With this structure however, it is straightforward to infer the data sources for each reported data point, and no potential erroneous merging of organisms are made. This latter point is of concern due to the lack of a universal, unambiguous identification methods, which should be a priority in any future efforts of producing and gathering microbial data.

By considering the variety of attributes within each data field and their relative frequencies, an image of the currently reported microbial diversity may be obtained. Examples of the most common attributes of the recorded traits are negative gram staining, positive indole metabolic assay, aerobic respiration, and growth in the mesophilic temperature range. Whether these observations are representative of the entire microbial diversity is less likely, considering bias in data reports and conducted research. As more is learned about known species and additional organisms are discovered and reported on, our perception of microbial diversity will continue to evolve.

A subset of the collected organisms and their phenotypic data was extracted into a reduced dataset to provide trait information for organisms with readily available genome data. To condense the dataset, reports from different sources on the same organisms were merged. This way, all genomes would be paired with all the information available in the dataset for the selected fields. By only using a subset of organisms and traits, the relative completeness of each data category was increased and only the entries with the most reports were included in the establishment of genotype—phenotype associations. Alternatively, the procedure could have imposed more permissive criteria for the

considered species, and thus increased the sample size of the reduced dataset. The trade-off between quality and quantity should be regarded in any further attempts at utilising existing data for association studies.

The genome sequences of each organism in the reduced dataset were downloaded and annotated, yielding ontology and orthology terms describing the genomic contents of the collected genomes. Of the utilised term hierarchies, gene ontology was the most diverse and numerous. Clusters of orthologous genes were initially more numerous than KEGG orthology, however less than half of the terms were later mapped to their official database, compromising its relative coverage. Upon reviewing the annotated terms, most were highly particular and thus only present in a few genomes. The terms found more commonly among the genomes generally represented conserved cellular features, with examples such as universal enzyme superfamilies and ribosomal subunits. Hence it was deemed sufficiently representative for its continued use.

The association of phenotypic features and genomic contents was conducted with respect to the trait of gram stain. This choice was based on the well-established understanding of gram-specific cellular features, its prevalence in the dataset, and its limited number of possible attributes: positive or negative. Thus to represent the distribution of any particular term across different gram attributes, only one contingency table was required per annotation term. Thus for each term, the organisms were divided into four categories: gram-positive with the term; gram-negative with the term; gram-positive without the term; and gram-negative without the term. The distributions were noted in contingency tables which were assessed with Fisher's exact tests. The resulting associations were represented by two values. Normalised odds ratios assigned the direction and magnitude of the term' association with either gram attribute, while p-values denoted the probability of erroneously rejecting the null hypothesis given a similarly extreme distribution. With the set significance levels ($||\widetilde{OR}|| > 1, \;\; pVal < 0.01$), 4,444 annotations were found associated with either gram stain attribute. Furthermore, some terms were found exclusively in genomes of organisms with a particular gram stain attribute. These may represent the most promising candidate genomic contents for predicting gram stain. However, many are determined using small sample sizes, which compromises their significance relative to the terms found associated with, but not exclusive to, particular gram stain attributes. Hence the exclusive terms have not been favoured in the present analysis.

Nevertheless, the biological connotations of the significantly associated annotation terms were inferred as appropriate in several of the investigated instances. Prominent examples include the association of terms related to outer membrane biogenesis with gram negativity, and anionic acid polymers with gram-positivity. Having identified terms with significant association to a gram stain attribute among the known data based on the dataset, the same patterns were applied to three random organisms without a registered gram stain attribute in the dataset. Different patterns were recognised within their annotated genomes and resulted in the successful determination of their gram attributes.

Thus it is believed that based on the present work, gram stain attributes for other organisms may be inferred by searching for similar patterns in their genomic content. Furthermore, other traits may be investigated using the same concept of inferring relations between known genotypic content and phenotypic features. Potential subjects for association from the present dataset include oxygen requirements and substrate. If similarly successful patterns of genomic content are discovered for these two traits, they can be used in the determination of growth capabilities of any organism with the same

genomic contents. This may be of use for instance in wet-lab methodologies seeking to facilitate growth of particular organisms, or to help overcome the issue of the great plate count anomaly. Further, similar studies could be conducted to help detect genomic targets involved in the production of desired metabolites, aiding the identification of new microbial species to utilise for industrial and medicinal purposes.

The present work highlights many areas of major potential for further investigation. Given an additional few months, the present work could have continued its effort of generating and maintaining a large repository for microbial data. Such repositories have already demonstrated their versatility by facilitating scientific efforts and discoveries, and by promoting the accessibility of its findings. Automated processes for fetching and standardising data from the utilised sources could ease this effort. Furthermore, additional sources and trait fields may be incorporated to incentivise the utilisation of more of the already existing microbial data for genotype—phenotype association. Additional tools that may simplify the standardisation efforts is the establishment of new or utilisation of existing ontologies, for instance for isolation sources and growth media. Given their acceptance and utilisation in the scientific community, such ontologies provide beneficial frameworks for organising biological information and help streamline the flow of information between researchers, repositories, and computers.

With the availability of additional data points, the current methodology of genotype—phenotype analysis could be expanded upon. Including additional data points for the already explored association between genotypic content and gram stain is but one example. Through more sophisticated analyses and machine learning, models for phenotype prediction could be developed to consider the presence of multiple features in an organism’s genomic content. This would limit a central issue with the concept of the present work, namely that the presence of a gene or other genotypic marker does not guarantee its expression in the organism’s phenotype. In concept, development, growth, and refinement of such models could result in powerful tools which when provided with the genomic content of any microbe, could give detailed predictions of their expressed phenotype. This could be indicating of its optimal growth requirements, facilitating subsequent web-lab procedures for its cultivation and study of whether the traits predicted by the model accurately manifest themselves in the organism. In addition to providing well-founded predictions of the organism’s capabilities before cultivation, such models may present themselves as the most prominent tools for understanding biological systems, enabling simulation of complex interactions and the interplay of the many components and levels of organisation occurring within a cell.

References

1. Gest H. the discovery of microorganisms by Robert Hooke and Antoni van Leeuwenhoek, Fellows of the Royal Society. Notes Rec R Soc Lond. 2004:58(2);187-201. DOI: 10.1098/rsnr.2004.0055.
2. Hooke R. Micrographia: or some physiological descriptions of minute bodies made by magnifying glasses, with observations and inquiries thereupon. London, England: The Royal Society; 1665.
3. National Cancer Institute. Tumor markers [Internet]. National Institutes of Health: Bethesda, USA. Updated 11.05.21. Retrieved 10.05.22 from https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/tumor-markers-fact-sheet.
4. Jiang C, Lv G, Tu Y, Cheng X, Duan Y, Zeng B, He B. Applications of CRISPR/Cas9 in the synthesis of secondary metabolites in filamentous fungi. Front Microbiol. 2021:12;638096. DOI: 10.3389/fmicb.2021.638096.
5. Armitage H. Fastest DNA sequencing technique helps undiagnosed patients find answers in mere hours [Internet]. Stanford Medicine: California, USA. Published 12.01.22. Retrieved 10.05.22 from https://med.stanford.edu/news/all-news/2022/01/dna-sequencing-technique.html.
6. Costello MJ, May RM, Stork NE. Can we name Earth's species before they go extinct? Science. 2013:339(6118);413-6. DOI: 10.1126/science.1230318.
7. Dykhuizen D. Species numbers in bacteria. Proc Calif Acad Sci. 2005:56(6, Suppl 1);62-71. PMID: 21874075.
8. Editorial. Microbiology by numbers. Nat Rev Microbiol. 2011:9;628. DOI: 10.1038/nrmicro2644.
9. Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. Proc Natl Acad Sci. 2016:113(21);5970-5. DOI: 10.1073/pnas.1521291113.
10. Tang YW, Ellis NM, Hopkins MK, Smith DH, Dodge DE, Persing DH. Comparison of phenotypic and genotypic techniques for identification of unusual aerobic pathogenic gram-negative Bacilli. J Clin Microbiol. 1998:36(12);3674-9. DOI: 10.1128/jcm.36.12.3674-3679.1998.
11. Donelli G, Vuotto C, Mastromarino P. Phenotyping and genotyping are both essential to identify and classify a probiotic microorganism. Microb Ecol Health Dis. 2013:24. DOI: 10.3402/mehd.v24i0.20105.
12. Connon SA, Giovannoni SJ. High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. Appl Environ Microbiol. 2002:68(8);3878-85. DOI: 10.1128/AEM.68.8.3878-3885.2002.
13. Clark DP, Pazdernik NJ, McGehee MR. Molecular biology. 3rd ed. Academic Press: Cambridge, USA; 2018.
14. Weissenback J. The rise of genomics. C R Biologies. 2016:339(7-8);231-9. DOI: 10.1016/j.crvi.2016.05.002.
15. Keeney JB. Microorganisms: applications in molecular biology. In John Wiley & Sons, Ltd (ed). eLS. 2007. DOI: 10.1002/9780470015902.a0000971.pub2.
16. Madigan MT, Bender KS, Buckley DH, Sattley WM, Stahl DA. Brock Biology of microorganisms. 15th global ed. New York, USA: Pearson Education; 2017.
17. Cockell CS. Microbial rights? EMBO Rep. 2011;12(3):181. DOI: 10.1038/embor.2011.13.
18. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, *et al*. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015:15(2);141-61. DOI: 10.1007/s10142-015-0433-4.

19. Krakashev D, Galabova D, Simeonov I. A simple and rapid test for differentiation of aerobic from anaerobic bacteria. World J Microbiol Biotechnol. 2003:19;233-8. DOI: 10.1023/A:1023674315047.
20. Braissant O, Astasov-Frauenhoffer M, Waltimo T, Bonkat G. A review of methods to determine viability, vitality, and metabolic rates in microbiology. Front Microbiol. 2020:11;547458. DOI: 10.3389/fmicb.2020.547458.
21. Pei L, Schmidt M. Fast-growing engineered microbes: new concerns for gain-of-function research? Front Genet. 2018:9;207 DOI: 10.3389/fgene.2018.00207.
22. ALS Environmental Ltd. Limit of detection: microbiology [Internet]. ALS: Coventry, UK. Updated 03.11.17. Retrieved 11.05.22 from https://www.alsenvironmental.co.uk/media-uk/pdf/datasheets/micro-lp/als_micro_limit-of-detection---microbiology_uk_2017.pdf.
23. Hull SC. Trait [Internet]. National Human Genome Research Institute: Bethesda, USA. Updated 11.05.22. Retrieved 11.05.22 from https://www.genome.gov/genetics-glossary/Trait.
24. Violle C, Navas ML, Vile D, Kazakou E, Fortunel C, Hummel I, Garnier E. Let the concept of trait be functional! Oikos. 2007:116(5);882-92. DOI: 10.1111/j.0030-1299.2007.15559.x.
25. Petchey OL, Hector A, Gaston KJ. How do different measures of functional diversity perform? Ecology. 2004:85(3);847-57. DOI: 10.1890/03-0226.
26. Eviner VT. Plant traits that influence ecosystem processes vary independently among species. Ecology. 2004:85(8);2215-29. DOI: 10.1890/03-0405.
27. Darwin C. On the origin of species my means of natural selection, or the preservation of favoured races in the struggle for life. 1st ed. John Murray: London, UK. 1859.
28. Lajoie G, Kembel SW. Making the most of trait-based approaches for microbial ecology. Trends Microbiol. 2019;27(10):814-23. DOI: 10.1016/j.tim.2019.06.003.
29. Nichols D. Cultivation gives context to the microbial ecologist. FEMS Microbiol Ecol. 2007:60(3);351-7. DOI: 10.1111/j.1574-6941.2007.00332.x.
30. Abby SS, Tannier E, Gouy M, Daubin V. Lateral gene transfer as support for the tree of life. PNAS. 2012:198(13);4962-7. DOI: 10.1073/pnas.1116871109.
31. Lanoil BD, Carlsom CA, Giovannoni SJ. Bacterial chromosomal painting for *in situ* monitoring of clustered marine bacteria. Environ Microbiol. 2000:2(6);654-65. DOI: 10.1046/j.1462-2920.2000.00148.x.
32. Suzuki MT, Rappé MS, Haimberger ZW, Winfield H, Adair N, Ströbel J, Giovannoni SJ. Bacterial diversity among small-subunit rRNA gene clones and cellular isolates from the same seawater sample. Appl Environ Microbiol. 1997:63(3);983-9. DOI: 10.1128/aem.63.3.983-989.1997.
33. Ramamurthy T, Ghosh A, Pazhani GP, Shinoda S. Current perspectives on viable but non-culturable (VBNC) pathogenic bacteria. Front Public Health. 2014:2;103. DOI: 10.3389/fpubh.2014.00103.
34. Zhulin IB. Databases for microbiologists. J Bacteriol. 2015:197(15);2458-67. DOI: 10.1128/JB.00330-15.
35. Campedelli I, Mathur H, Salvetti E, Clarke S, Rea MC, Torriani S *et al*. Genus-wide assessment of antibiotic resistance in *Lactobacillus* spp. Appl Environ Microbiol. 2018:85(1):e01738-18. DOI: 10.1128/AEM.01738-18.
36. Moore LR, Coe A, Zinser ER, Saito MA, Sullivan MB, Lindell D *et al*. Culturing the marine cyanobacterium *Prochlorococcus*. Limnol Oceanogr Methods. 2007:5(10);353-62. DOI: 10.4319/lom.2007.5.353.
37. Michał B, Gagat P, Jabłoński S, Chilimoniuk J, Gaworski M, Mackiewicz P, Marcin Ł. PhyMet [database, ver. 2]. Faculty of Biotechnology, University of Wroclaw: Poland.

Published 20.01.15, updated 06.04.19. Retrieved 07.11.21 from http://phymet2.biotech.uni.wroc.pl.
38. Reimer LC, Carbasse JS, Koblitz J, Ebeling C, Podstawka A, Overmann J. BacDive in 2022: the knowledge base for standardized bacterial and archaeal data. Nucleic Acids Res. 2021:50(D1);D741-6. DOI: 10.1093/nar/gkab961.
39. Sierra MA, Bhattacharya C, Ryon K, Meierovich S, Shaaban H, Westfall D *et al*. Microbe Directory [dataset, ver. 2]. Mason Lab, Weill Cornell Medicine, Cornell University: New York, USA. Published 05.01.18, updated 31.01.20. Retrieved 07.11.21 from https://github.com/dcdanko/MD2/blob/master/datasets/all/microbe-directory.csv.
40. Biological Resource Center of Institut Pasteur. Catalogue of Microorganisms [database]. Institut Pasteur: Paris, France. Retrieved 07.11.21 from https://catalogue-crbip.pasteur.fr/recherche_catalogue.xhtml.
41. PATRIC. *Veillonella parvula DSM 2008* [Internet]. Bioinformatics Resource Center: Illinois, USA. Retrieved 13.05.22 from https://www.patricbrc.org/view/Genome/479436.6.
42. GOLD. *Veillonella parvula Te3, DSM 2008 [Internet]. Joint Genome Institute: California, USA. Retrieved 13.05.22 from https://gold.jgi.doe.gov/analysis_project?id=Ga0031405.*
43. Nucleic Acids Research. NAR database summary paper alpha list [Internet]. Oxford Academic: Oxford, UK. Retrieved 13.05.22 from http://www.oxfordjournals.org/our_journals/nar/database/a/.
44. Madin JS, Nielsen DA, Brbic M, Corkrey R, Danko D, Edwards K, *et al*. A synthesis of bacterial and archaeal phenotypic trait data. Sci Data. 2020;7:170. DOI: 10.1038/s41597-020-0497-4.
45. Bergey DH, Krieg NR, Holt JG. Bergey's manual of systematic bacteriology. Baltimore, MD: Williams & Wilkins; 1984.
46. Hackman TJ, Zhang B. Using neural networks to mine text and predict metabolic traits for thousands of microbes. PLoS Comput Biol. 2021:17(3);e1008757. DOI: 10.1371/journal.pcbi.1008757.
47. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova, Sundaramurthi JC, Lee J *et al*. Genomes OnLine Database (GOLD) v.8: overview and updates. Nucleic Acids Res. 2021:49(D1);D723-33. DOI: 10.1093/nar/gkaa983.
48. Sierra MA, Bhattacharya C, Ryon K, Meierovich S, Shaaban H, Westfall D *et al*. The Microbe Directory v2.0: an expanded database for ecological and phenotypical features of microbes. BioRxiv. 2019. DOI: 10.1101/2019.12.20.860569.
49. Ørstavik KH. Rosalind Franklin [Internet] Store Norske Leksikon. Published 14.02.09, updated 10.05.22. Retrieved 13.05.22 from https://snl.no/Rosalind_Franklin.
50. Illumina. DNA sequencing [Internet]. San Diego, USA. Retrieved 12.05.22 from https://www.illumina.com/techniques/sequencing/dna-sequencing.html.
51. ThermoFisher Scientific. Ion Torrent [Internet]. Waltham, USA. Retrieved 12.05.22 from https://www.thermofisher.com/no/en/home/brands/ion-torrent.html.
52. Oxford Nanopore Technologies. Nanopore DNA sequencing [Internet]. Oxford, UK. Retrieved 12.05.22 from https://nanoporetech.com/applications/dna-nanopore-sequencing.
53. PacBio. Delivering highly accurate long reads to drive discovery in life science [Internet]. California, USA. Retrieved 12.05.2022 from https://www.pacb.com/technology/hifi-sequencing/how-it-works/.
54. Trewavas A. A brief history of systems biology: "Every object that biology studies is a system of systems." Francois Jacob (1974). Plant Cell. 2006:18(10);2420-30. DOI: 10.1105/tpc.106.042267.

55. Simeonidis V. Is systems biology doomed to fail? (Hint: NO!) [Internet]. Web of V: Seattle, USA. Published 07.04.11. Retrieved 12.05.22 from http://vangelissimeonidis.com/?p=21.
56. Brenner S. Sequences and consequences. Philos Trans R Soc Lond B Biol Sci. 2010:365(1537);207-12. DOI: 10.1098/rstb.2009.0221.
57. Hoffman JIE. Biostatistics for medical and biomedical practitioners. Academic Press: California, USA; 2015. Chapter 13: Hypergeometric distribution, pp. 179-82. DOI: 10.1016/C2014-0-02732-3.
58. Centers for Disease Control and Prevention. Interpreting results of case-control studies [Internet]. Atlanta, USA. Published 2014. Retrieved 13.05.22 from https://www.cdc.gov/training/SIC_CaseStudy/Interpreting_Odds_ptversion.pdf.
59. SciPy. scipy.stats.fisher_exact [Internet]. Retrieved 13.05.22 from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.fisher_exact.html.
60. Szumilas M. Explaining odds ratios. J Can Acad Child Adolesc Psychiatry. 2010:19(3);227-9. PMID: 20842279.
61. Noble WS. How does multiple testing correction work? Nat Biotechnol. 2009:27(12);1135-7. DOI: 10.1038/nbt1209-1135.
62. QuickGO. GO:0042546 cell wall biogenesis [Figure]. EMBL-EBI: Cambridgeshire, UK. Retrieved 13.05.22 from https://www.ebi.ac.uk/QuickGO/term/GO:0042546.
63. Gene Ontology Consortium, Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, *et al*. Gene Ontology: tool for the unification of biology. Nat Genet. 2000:25(1);25-9. DOI: 10.1038/75556.
64. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016:44(D1);D457-62. DOI: 10.1093/nar/gkv1070.
65. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J Mol Biol. 2016:428(4);726-31. DOI: 10.1016/j.jmb.2015.11.006.
66. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997:278(5338);631-7. DOI: 10.1126/science.278.5338.631.
67. Galperin MY, Wolf YI, Makarova KS, Alvarez RV, Landsman D, Koonin EV. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. 2021:49(D1);D274-81. DOI: 10.1093/nar/gkaa1018.
68. BacDive. *Hymenobacter defluvii* POA9 [Internet]. BacDive. Retrieved 13.05.22 from https://bacdive.dsmz.de/strain/158592.
69. BacDive. *Lutibacter profundi* LP1 [Internet]. BacDive. Retrieved 13.05.22 from https://bacdive.dsmz.de/strain/132984.
70. Mungall C. Ontology of Biological Attributes [Internet]. BioPortal. Updated 19.01.22. Retrieved 13.05.22 from https://bioportal.bioontology.org/ontologies/OBA.
71. Buttigieg PL, Mungall C, Blumberg K, Wilkie I, Duncan B, Meyer R, *et al*. EnvironmentOntology/envo: 2021-05-14 release [Internet]. Updated 19.05.21. Retrieved 13.05.22 from https://zenodo.org/record/4772492#.Yn7O3U5By5c.
72. Chibucos MC, Zweifel AE, Herrera JC, Meza W, Eslamfam S, Uetz P, *et al*. An ontology for microbial phenotypes. BMC Microbiol. 2014:14;294. DOI: 10.1186/s12866-014-0294-3.
73. Microsoft Corporation. Microsoft Office Excel [software, ver. 2203]. Published 1987, updated 12.04.22. Available from https://www.microsoft.com/nb-no/microsoft-365/excel.

74. Corkrey R, McMeekin TA, Bowman JP, Ratkowsky DA, Olley J, Ross T. The Biokinetic Spectrum for Temperature. PLoS One. 2016:11(4):e0153343. DOI: 10.1371/journal.pone.0153343.
75. Berg EG. A new spin on the old gram stain [Figure]. Chemical & Engineering News. Published 28.04.15. Retrieved 12.05.22 from https://cen.acs.org/articles/93/web/2015/04/New-Spin-Old-Gram-Stain.html.
76. Steward K. Gram positive vs. gram negative [Figure]. Technology Networks, Immunology & Microbiology. Published 21.08.19, updated 31.03.22. Retrieved 12.05.22 from https://www.technologynetworks.com/immunology/articles/gram-positive-vs-gram-negative-323007.
77. Louca S, Parfrey LW, Doebeli M. Decoupling function and taxonomy in the global ocean microbiome. Science. 2016:353(5305):1272-7. DOI: 10.1126/science.aaf4507.
78. Louca S. Functional Annotation of Prokaryotic Taxa [dataset, ver. 1.2.4]. Louca Lab, University of Oregon: Eugene, USA. Published 16.09.16, retrieved 07.11.21 from http://www.loucalab.com/archive/FAPROTAX/lib/php/index.php?section=Download.
79. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, *et al*. Jupyter Notebooks – a publishing format for reproducible computational workflows. In Loizides F, Schmidt B (eds.), Positioning and Power in Academic Publishing: Players, Agents and Agendas. 2016: IOS Press, pp. 87–90. DOI: 10.3233/978-1-61499-649-1-87.
80. Barberan A. International Journal of Systematic and Evolutionary Microbiology (IJSEM) phenotypic database [dataset, ver. 3]. Figshare. Published 01.12.16, updated 06.12.16. DOI: https://doi.org/10.6084/m9.figshare.4272392.v3.
81. Ramírez-Bahena MH, Peix A, Rivas R, Camacho M, Rodríguez-Navarro DN, Mateos PF, *et al. Bradyrhizobium pachyrhizi* sp. nov. isolated from effective nodules of *Pachyrhizus erosus*. Int J Syst Evol Microbiol. 2009:59(8);1929-34. DOI: 10.1099/ijs.0.006320-0.
82. Kim BJ, Hong SH, Kook YH, Kim BJ. *Mycobacterium pragordonae* sp. nov., a slowly growing, scotochromogenic species closely related to *Mycobacterium gordonae*. Int J Syst Evol Microbiol. 2014:64(1);39-45. DOI: 10.1099/ijs.0.051540-0.
83. NCBI Taxonomy. Pseudomonas aeruginosa [Internet]. National Centre for Biotechnology Information: Bethesda, USA. Retrieved 17.11.21 from https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=287&lvl=3&lin=f&keep=1&srchmode=1&unlock.
84. NCBI Taxonomy. *Stenotrophomonas maltophila* [Internet]. National Centre for Biotechnology Information: Bethesda, USA. Retrieved 17.11.21 from https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=40324.
85. Kremer CT, Thomas MK, Litchman E. Temperature- and size-scaling of phytoplankton population growth rates: Reconciling the Eppley curve and the metabolic theory of ecology. Limnol Oceanogr. 2017:62(4);1658-70. DOI: 10.1002/lno.10523.
86. Mason MM. A comparison of the maximal growth rates of various bacteria under optimal conditions. J Bacteriol. 1935:29(2);103-10. DOI: 10.1128/jb.29.2.103-110.1935.
87. Richards MA, Cassen V, Heavner BD, Ajami NE, Herrmann A, Simeonidis E, Price ND. MediaDB: a database of microbial growth conditions in defined media. PLoS One. 2014:9(8);e103548. DOI: 10.1371/journal.pone.0103548.
88. Richards MA, Cassen V, Heavner BD, Ajami NE, Herrmann A, Simeonidis E, Price ND. MediaDB [database]. Price Lab, Institute for Systems Biology: Seattle, USA. Published 06.08.14, updated 07.10.15. Retrieved 07.11.21 from https://mediadb.systemsbiology.net/defined_media/downloads/.
89. Nielsen SL. Size-dependent growth rates in eukaryotic and prokaryotic algae exemplified by green algae and cyanobacteria: comparisons between unicells and

colonial growth forms. J Plankton Res. 2006:28(5);489-98. DOI: 10.1093/plankt/fbi134.
90. Davis JJ, Wattam AR, Aziz RK, Brettin T, Butler R, Chlenski P *et al*. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. Nucleic Acids Res. 2020:48(D1):D606-12. DOI: 10.1093/nar/gkz943.
91. Davis JJ, Wattam AR, Aziz RK, Brettin T, Butler R, Chlenski P *et al*. genome_metadata [dataset]. Bioinformatics Resource Center, University of Chicago: Illinois, USA. Published 31.10.19, updated 08.01.20. Retrieved 07.11.21 from ftp://ftp.patricbrc.org/RELEASE_NOTES/.
92. NCBI Taxonomy. *Clostridioles mangenotii* TR [Internet]. National Centre for Biotechnology Information: Bethesda, USA. Retrieved 20.11.21 from https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1408823.
93. NCBI Taxonomy. *Faecalitalea cylindroides* ATCC 278303 [Internet]. National Centre for Biotechnology Information: Bethesda, USA. Retrieved 20.11.21 from https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=64975560-70p.
94. NCBI Taxonomy. [*Scytonema hofmanni*] UTEX 2349 [Internet]. National Centre for Biotechnology Information: Bethesda, USA. Retrieved 20.11.21 from https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1469607.
95. Michał B, Gagat P, Jabłoński S, Chilimoniuk J, Gaworski M, Mackiewicz P, Marcin Ł. PhyMet2: a database and toolkit for phylogenetic and metabolic analyses of methanogens. Environ Microbiol Rep. 2018:10(3);378-82. DOI: 10.1111/1758-2229.12648.
96. Brbić M, Piškorec M, Vidulin V, Kriško A, Šmuc T, Supek F. The landscape of microbial phenotypic traits and associated genes. Nucleic Acids Res. 2016:44(21); 10074–90. DOI: 10.1093/nar/gkw964.
97. Brbić M, Piškorec M, Vidulin V, Kriško A, Šmuc T, Supek F. ProTraits binaryIntergatedPr0.95.txt [dataset]. Rudjer Boskovic Institute: Zagreb, Croatia. Published 25.10.16. Retrieved 07.11.21 from http://protraits.irb.hr/data.html.
98. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Syers EW. GenBank. Nucleic Acids Res. 2016:44(D);D67-72. DOI: 10.1093/nar/gkv1276.
99. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016:44(D1);D733-45. DOI: 10.1093/nar/gkv1189.
100. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R *et al*. RefSeq [database]. National Center for Biotechnology Information: Bethesda, USA. Published 01.01.05. Available at https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/.
101. Vieira-Silva S, Rocha EPC. The systemic imprint of growth and its uses in ecological (meta)genomics. PLoS Genet. 2010:6(1):e1000808. DOI: 10.1371/journal.pgen.1000808.
102. Tableau Software. Tableau Desktop Public Edition [software, ver. 2021.3.1]. Seattle, USA. Released 2004, updated 22.09.21. Available from https://www.tableau.com/support/releases/desktop/2021.3.1.
103. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: visualization of intersection sets. IEEE Trans Vis Comput Graph. 2014:20(12);1983-92. DOI: 10.1109/TVCG.2014.2346248.
104. Nothman J. UpSetPlot [software, ver. 0.6.0]. Released 21.02.19, updated 10.08.21. Available from https://pypi.org/project/UpSetPlot/.

105. NCBI. Prokaryotes.txt [dataset, release 248]. National Centre for Biotechnology Information: Bethesda, USA. Retrieved 15.02.22 from https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/.
106. Cock PA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009:25(11);1422-3. DOI: 10.1093/bioinformatics/btp163.
107. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol. 2021:38(12);5825–9. DOI: 10.1093/molbev/msab293.
108. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019:47(D1);D309–14. DOI: 10.1093/nar/gky1085.
109. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021:18;366–8. DOI: 10.1038/s41592-021-01101-x.
110. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007:9(3);90-5. DOI: 10.1109/MCSE.2007.55.
111. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020:17(3);261-72. DOI: 10.1038/s41592-019-0686-2.
112. Seabold S, Perktold J. Statsmodels: economic and statistical modeling with Python. Proc Python Sci Conf. 2010;18. DOI: 10.25080/Majora-92bf1922-012.
113. Bedre R. reneshbedre/bioinfokit: Bioinformatics data analysis and visualisation toolkit [software, ver. 2.0.8]. Zenodo. Published 05.03.20, updated 06.01.21. DOI: 10.5281/zenodo.3698145.
114. NCBI. cog-20.def.tab [dataset]. National Centre for Biotechnology Information: Bethesda, USA. Updated 11.09.20. Retrieved 01.05.02.22 from https://ftp.ncbi.nih.gov/pub/COG/COG2020/data/.
115. NCBI. fun-20.tab [dataset]. National Centre for Biotechnology Information: Bethesda, USA. Updated 22.08.20. Retrieved 01.05.02.22 from https://ftp.ncbi.nih.gov/pub/COG/COG2020/data/.
116. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of Gene Ontology terms. PLoS ONE. 2011:6(7);e21800. DOI: 10.1371/journal.pone.0021800.
117. Kanehisa M, Sato Y, Kawashima M. KEGG mapping tools for uncovering hidden features in biological data. Protein Sci. 2022:31(1);47-53. DOI: 10.1002/pro.4172.
118. Britannica editors. Staphylococcus [Internet]. Encyclopaedia Britannica. Published 20.11.17. Retrieved 26.04.22 from https://www.britannica.com/science/Staphylococcus.
119. Patterson MJ. Streptococcus. In Baron S, editor. Medicinal Microbiology, 4th ed. Galveston, USA: University of Texas, Medical Branch; 1996. Chapter 13. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK7611/.
120. Bergan, JV. Mycobacterium [Internet]. Store medisinske leksikon. Store Norske Leksikon. Published 14.02.2009, updated 16.11.21. Retrieved 26.04.22 from https://sml.snl.no/Mycobacterium.
121. Centres for Disease Control and Prevention. E. coli [Internet]. CDC. Updated 03.03.22. Retrieved 26.04.22 from https://www.cdc.gov/ecoli/index.html.

122. Danilova I, Sharipova M. The practical potential of Bacilli and their enzymes for industrial production. Front Microbiol. 2020:11;1782. DOI: 10.3389/fmicb.2020.01782.
123. Anayo OF, Scholastica EC, Peter OC, Nneji UG, Obinna A, Mistura LO. The beneficial roles of Pseudomonas in medicine, industries, and environment: a review. In Sriramulu D, editor. Pseudomonas aeruginosa – an armory within. London, UK: IntechOpen; 2019. DOI: 10.5772/intechopen.85996.
124. Ferraiuolo SB, Cammarota M, Schiraldi C, Restaino OF. Streptomyces as platform for biotechnological production processes of drugs. Appl Microbiol Bitechnol. 2021:105(2);551-68. DOI: 10.1007/s00253-020-11064-2.
125. Kevbrina MV, Okhapkina AA, Akhlynin DS, Kravchenko IK, Nozhevnikova AN, Gal'chenko VF. Rost mezofil'nykh metanotrofov pri nizkikh temperaturakh [Growth of mesophilic methanotrophs at low temperatures]. Mikrobiologiia. 2001;70(4):444-51. Available from https://pubmed.ncbi.nlm.nih.gov/11558268/.
126. Vicente M, García-Ovalle M. Making a point: the role of DivIVA in Streptococcal polar anatomy. J Bacteriol. 2007:189(4);1185-8. DOI: 10.1128/JB.01710-06.
127. Halbedel S, Lewis RJ. Structural basis for interaction of DivIVA/GpsB proteins with their ligands. Mol Microbiol. 2019:111(6);1404-15. DOI: 10.1111/mmi.14244.
128. Sandoval CM, Baker SL, Jansen K, Metzner SI, Sousa MC. Crystal structure of BamD, an essential component of the β-barrel assembly machinery of gram negative bacteria. J Mol Biol. 2011:409(3);348-57. DOI: 10.1016/j.jmb.2011.03.035.
129. Bergal HT, Hopkins AH, Metzner SI, Sousa MC. The structure of a BamA-BamD fusion illuminates the architecture of the β-barrel assembly machine core. Structure. 2016:24(2);243-51. DOI: 10.1016/j.str.2015.10.030.
130. Rossiter AE, Leyton DL, Tveen-Jensen K, Browning DF, Sevastsyanovich Y, Knowles TJ, *et al*. The essential β-barrel assembly machinery complex components BamD and BamA are required for autotransporter biogenesis. J Bacteriol. 2011:193(16);4250-3. DOI: 10.1128/JB.00192-11.
131. Briaud P, Carroll RK. Extracellular vesicle biogenesis and functions in gram-positive bacteria. Infect Immun. 2020:88(12);e00433.20. DOI: 10.1128/IAI.00433-20.
132. Bennett-Lovsey R, Hrt SE, Shirai H, Mizuguchi K. The SWIB and the MDM2 domains are homologous and share a common fold. Bioinformatics. 2002:18(4);626-30. DOI: 10.1093/bioinformatics/18.4.626.
133. Koonin EV, Makarova KS, Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification. In: Annual Reviews Collection [Internet]. National Centre for Biotechnology Information. Bethesda, USA; 2002. Retrieved 01.05.22 from: https://www.ncbi.nlm.nih.gov/books/NBK2228/.
134. National Library of Medicine. COG5180 [Internet]. National Centre of Biotechnology Information: Bethesda, USA. Retrieved 01.05.22 from https://www.ncbi.nlm.nih.gov/research/cog/cog/COG5180/.
135. National Library of Medicine. COG3953 [Internet]. National Centre of Biotechnology Information: Bethesda, USA. Retrieved 07.05.22 from https://www.ncbi.nlm.nih.gov/research/cog/cog/COG3953/.
136. Richts B, Rosenberg J, Commichau FM. A survey of pyridoxal 5'-phosphade-dependent proteins in the gram-positive model bacterium *Bacillus subtilis*. Front Mol Biosci. 2019:6;32. DOI: 10.3389/fmolb.2019.00032.
137. Aínsa JA, Ryding NJ, Hartley N, Findlay KC, Bruton CJ, Chater KF. WhiA, a protein of unknown function conserved among gram-positive bacteria, is essential for sporulation in *Streptomyces coelicolor* A3(2). J Bacteriol. 2000:182(19);5470-8. DOI: 10.1128/jb.182.19.5470-5478.2000.

138. Wenzel M, Gulsoy INC, Gao Y, Teng Z, Willemse J, Middelkamp, *et al.* Control of septum thickness by the curvature of SepF polymers. Proc Natl Acad Sci USA. 2021:118(2);e2002635118. DOI: 10.1073/pnas.2002635118.
139. Kim JN, Ahn SJ, Seaton K, Garrett S, Burne RA. Transcriptional organisation and physiological contributions of the relQ operon of Streptococcus mutans. J Bacteriol. 2012:194(8);1968-78. DOI: 10.1128/JB.00037-12.
140. Lu, CD. Pathways and regulation of bacterial arginine metabolism and perspectives for obtaining arginine overproducing strains. Appl Microbiol Biotechnol. 2006:70;261-72. DOI: 10.1007/s00253-005-0308-z.
141. Itoh Y, Matsumoto H. Mutations affecting regulation of the anabolic argF and the catabolic aru genes in Pseudomonas aeruginosa PAO. Mol Gen Genet. 1992;231(3):417-25. DOI: 10.1007/BF00292711.
142. Richie DL, Wang L, Chan H, De Pascale G, Six DA, Wei JR, Dean CR. A pathway directed positive growth restoration assay to facilitate the discovery of lipid A and fatty acid biosynthesis inhibitors in *Acinetobacter baumannii*. PLoS ONE. 2018:13(3);e0193851. DOI: 10.1371/journal.pone.0193851.
143. Lu YJ, Zhang YM, Rock CO. Product diversity and regulation of type II fatty acid synthases. Biochem Cell Biol. 2004:82(1);145-55. DOI: 10.1139/o03-076.
144. de Mendoza D, Schujam GE. Lipid biosynthesis. In Schaechter M, ed. Encyclopaedia of Microbiology, 3rd ed. Cambridge, USA: Academic Press; 2009, pp. 219-228.
145. Hoyles L, Falsen E, Foster G, Collins MD. *Actinomyces coleocnis* sp. noc., from the vagina of a dog. Inj J Syst Evol Microbiol. 2002:52(4);1201-3. DOI: 10.1099/00207713-52-4-1201.
146. Lavasani PS, Motevaseli E, Shirzad M, Modarressi MH. Isolation and identification of *Komagataeibacter xylinus* from Iranian traditional vinegars and molecular analyses. Iran J Microbiol. 2017:9(6);338-47. PMID: 29487732.
147. Klenk HP, Lapidus A, Chertkov O, Copeland A, Del Rio TG, Nolan M, *et al.* Complete genome sequence of the thermophilic, hydrogen-oxidizing *Bacillus tusciae* type strain (T2T) and reclassification in the new genus, *Kyrpidia* gen. nov. as *Kyrpidia tusciae* comb. nov. and emendation of the family Alicyclobacillaceae da Costa and Rainey, 2010. Stand Genomic Sci. 2011:5(1);121-34. DOI: 10.4056/sigs.2144922.
148. National Library of Medicine. *Pseudomonas putida* [Internet]. National Centre of Biotechnology Information: Bethesda, USA. Retrieved 04.05.22 from https://www.ncbi.nlm.nih.gov/genome/174?genome_assembly_id=392406.
149. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014:30(14);2068-9. DOI: 10.1093/bioinformatics/btu153.
150. Cuccuru G, Orsini M, Pinna A, Sbardellati A, Soranzo N, Travaglione A, *et al*. Orione, a web-based framework for NGS analysis in microbiology. Bioinformatics. 2014:30(13);1928-9. DOI: https://doi.org/10.1093/bioinformatics/btu135.
151. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, *et al*. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018:46(W1);W537–44. DOI: 10.1093/nar/gky379.
152. Cantalapiedra CP, Hernández-Plaza A, Coelho LP, Szklarczyk D, Forslund SK, Jensen LJ, *et al*. EggNOG-Mapper [Internet]. EggNOG. Retrieved 12.05.22 from http://eggnog-mapper.embl.de.
153. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, *et al*. e5.og_annotations.tsv [dataset, ver. 5]. EggNOG raw data. Updated 15.09.18. Retrieved 09.05.22 from http://eggnog5.embl.de/download/eggnog_5.0/.

154. National Library of Medicine. COG0454 [Internet]. National Centre of Biotechnology Information: Bethesda, USA. Retrieved 29.04.22 from https://www.ncbi.nlm.nih.gov/research/cog/cog/COG0454/.
155. Vetting MW, de Carvalho LPS, Yu M, Hedge SS, Magnet S, Roderick SL, Blanchard JS. Structure and functions of the GNAT superfamily of acetyltransferases. Arch Biochem Biophys. 2005:433(1);212-26. DOI: 10.1016/j.abb.2004.09.003.
156. National Library of Medicine. COG0142 [Internet]. National Centre of Biotechnology Information: Bethesda, USA. Retrieved 29.04.22 from https://www.ncbi.nlm.nih.gov/research/cog/cog/COG0142/.
157. EMBL-EBI QuickGO. GO:0003674 [Internet]. EMBL: Hinxton, UK. Published 30.03.01, updated 08.04.22. Retrieved 29.04.22 from https://www.ebi.ac.uk/QuickGO/term/GO:0003674.
158. EMBL-EBI QuickGO. GO:0044237 [Internet]. EMBL: Hinxton, UK. Published 16.12.04, updated 14.09.21. Retrieved 29.04.22 from https://www.ebi.ac.uk/QuickGO/term/GO:0044237.
159. KEGG Orthology. K02988 [Internet]. GenomeNet: Kyoto, Japan. Retrieved 29.04.22 from https://www.genome.jp/entry/K02988.
160. KEGG Orthology. K02892 [Internet]. GenomeNet: Kyoto, Japan. Retrieved 29.04.22 from https://www.genome.jp/entry/K02892.
161. Mostafavi M, Wang L, Xie L, Takeoka KT, Richie DL, Casey F, *et al*. Interplay of *Klebsiella pneumoniae fabZ* and *lpxC* mutations leads to LpxC inhibitor-dependent growth resulting from loss of membrane homeostasis. mSphere. 2018:31(5);e00508-18. DOI: 10.1128/mSphere.00508-18.
162. Wang H, Cronan JE. Functional replacement of the FabA and FabB proteins of Escherichia coli fatty acid synthesis by Enterococcus faecalis FabZ and FabF homologues. J Biol Chem. 2004:279(33);34489-95. DOI: 10.1074/jbc.M403874200.
163. National Library of Medicine. ASM973810v1 [Internet]. National Centre for Biotechnology Information. Bethesda, USA. Retrieved 06.05.22 from https://www.ncbi.nlm.nih.gov/assembly/GCA_009738105.1.
164. Sacchromyces Genome Database. Gene Ontology Slim Term Mapper [Internet]. Stanford University: Standord, USA. Retrieved 08.05.22 from https://www.yeastgenome.org/goSlimMapper.
165. Lewis-Singler Institute for Integrative Genomics. Generic Gene Ontology (GO) term mapper [Internet]. Princeton University: Princeton, USA. Retrieved 08.05.22 from https://go.princeton.edu/cgi-bin/GOTermMapper.
166. Själander M, Jahre M, Tufte G, Reissmann N. EPIC: an energy-efficient, high-performance GPGPU computing research infrastructure. arXiv preprint. 2019:arXiv:1912.05848. DOI: 10.48550/arXiv.1912.05848.

Appendix A: Supplementary information

This appendix presents an overview of the files made and utilised in the present work. All the files are found in the GitHub repository of the project, which can be accessed at https://github.com/jennymerkesvik/msc_thesis_supplementary_information. For simple navigation, use the "Go to file" option in the main directory and paste the path to the desired file using the shortcuts in Table A1.

Table A1: GitHub repository overview *detailing the thesis' supplementary information.*

No.	Description	Path
1	Raw trait dataset from various sources	database/raw_data
2	Scripts for preparing datasets	database/scripts
3	Prepared trait datasets from various sources	database/preprared_data
4	Output of the dataset assembly and analysis	database/output_files
5	NCBI overview of prokaryotes	sequences/ncbi_prokaryotes.txt
6	Protein sequence acquisition and verification	sequences/protein_checks
7	Reduced trait dataset used for association study	sequences/reducedDataset.csv
8	Protein sequence data comparison	sequences/protein_data
9	Functional annotation data and analysis	sequences/annotation
10	Fisher's tests for annotations and gram stain	analysis/fisher
11	Genotype—phenotype association	analysis/association
12	COG term generalisation and analysis	analysis/association/cog_terms
13	GO term generalisation and analysis	analysis/association/go_terms
14	KO term generalisation and analysis	analysis/association/ko_terms
15	Sequences and annotations for test species	analysis/suggest_gram_test

Appendix B: Gene Ontology term generalisation

This appendix contains the generalised GO term categories generated with REVIGO [116]. The interpretations of the figures are given in Subsection 4.3.2 (p. 46).

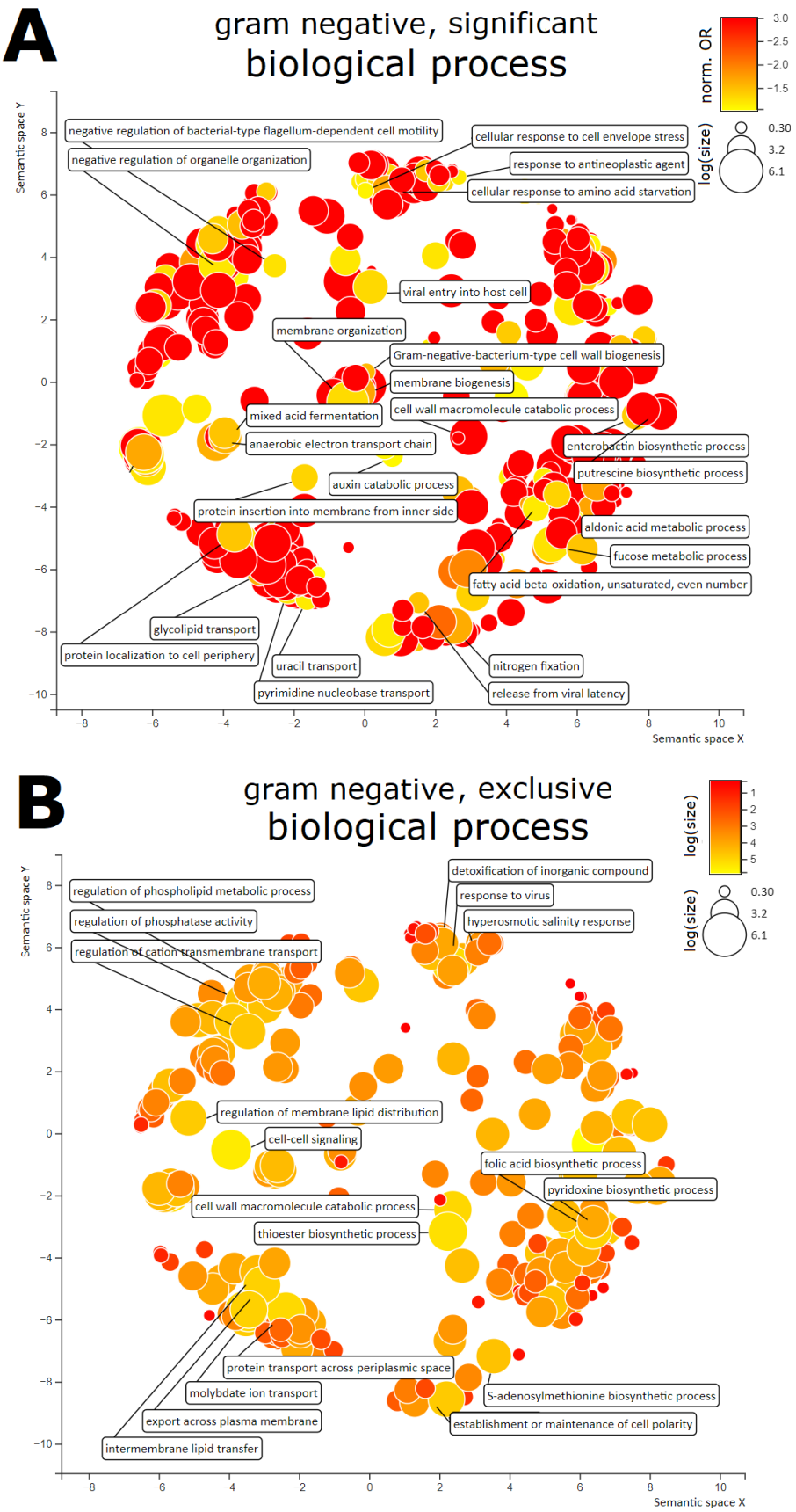


Figure B1: GO term biological process significantly associated with gram-negativity.

A: *significantly associated with gram-negativity. Circle size denotes the number of children terms within the category. Circle colour denotes $\widetilde{OR}$, from weak (yellow) to strong (red) association.*

B*: exclusively associated with gram-negativity. Both circle size and colour denote the number of children terms within the category.*

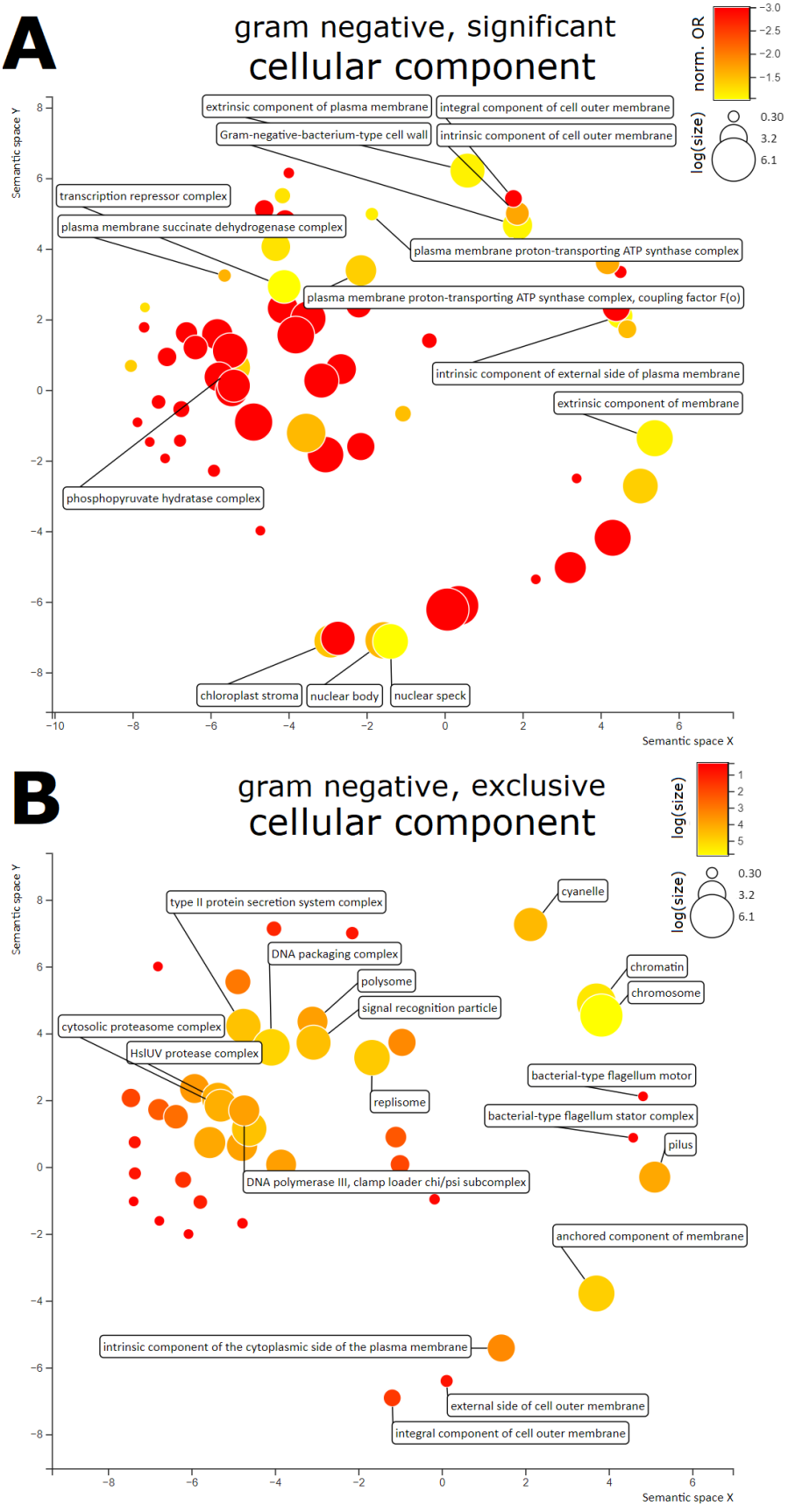


Figure B2: GO term cellular component significantly associated with gram-negativity.

A: *significantly associated with gram-negativity. Circle size denotes the number of children terms within the category. Circle colour denotes $\widetilde{OR}$, from weak (yellow) to strong (red) association.*

B*: exclusively associated with gram-negativity. Both circle size and colour denote the number of children terms within the category.*

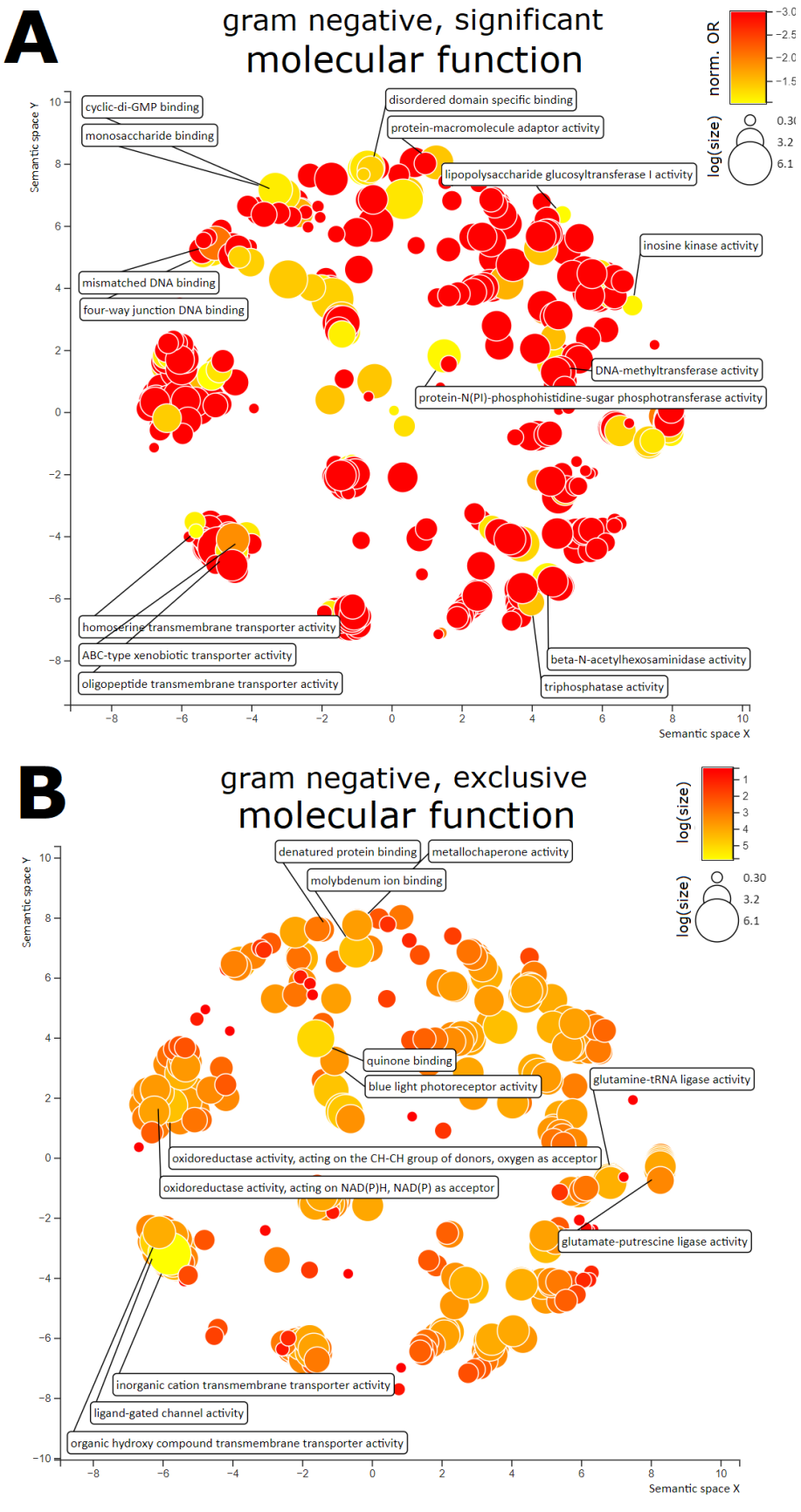
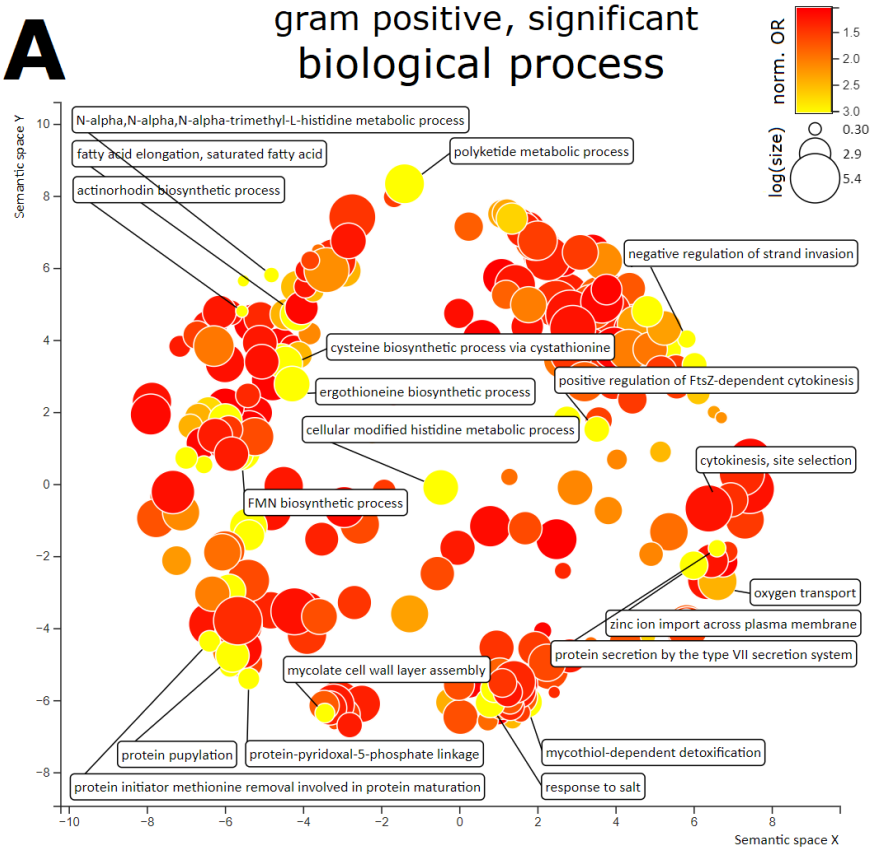


Figure B3: GO term molecular function significantly associated with gram-negativity.

A: *significantly associated with gram-negativity. Circle size denotes the number of children terms within the category. Circle colour denotes $\widetilde{OR}$, from weak (yellow) to strong (red) association.*

B*: exclusively associated with gram-negativity. Both circle size and colour denote the number of children terms within the category.*



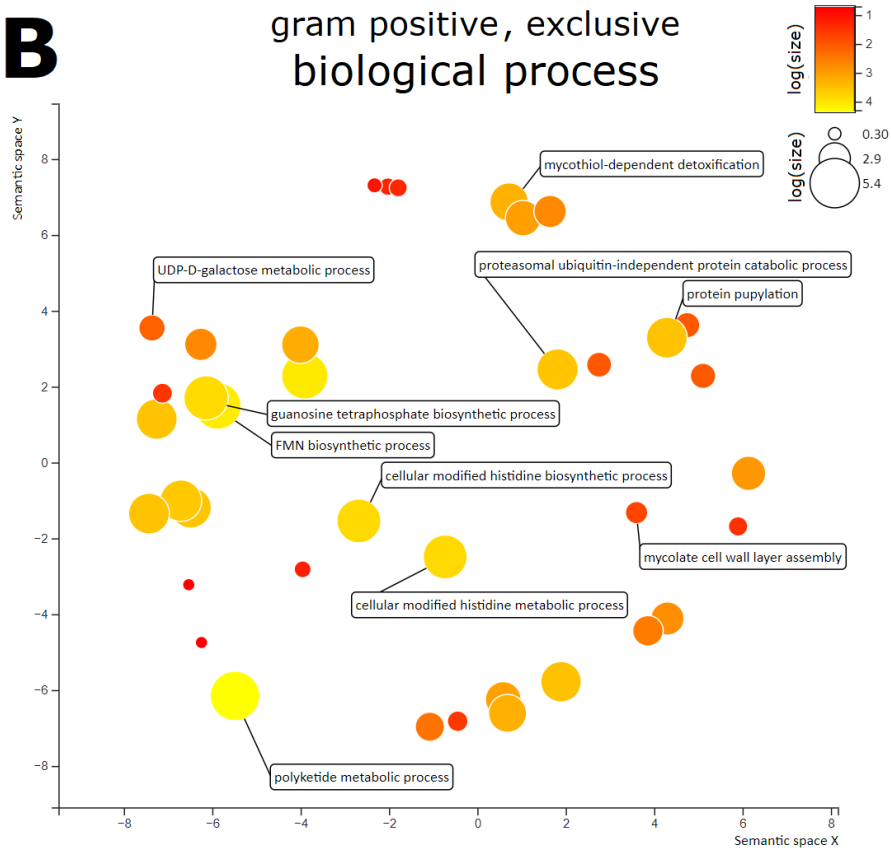
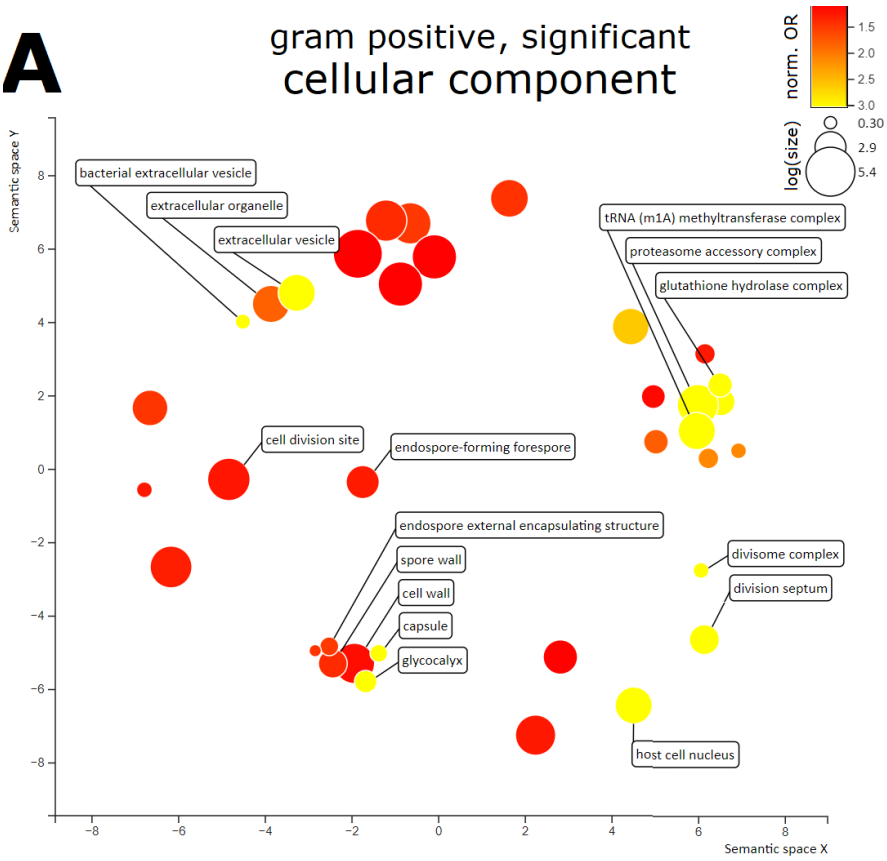


Figure B4: GO term biological process significantly associated with gram-positivity.

A: *significantly associated with gram-positivity. Circle size denotes the number of children terms within the category. Circle colour denotes $\widetilde{OR}$, from weak (red) to strong (yellow) association.*

B*: exclusively associated with gram-positivity. Both circle size and colour denote the number of children terms within the category.*



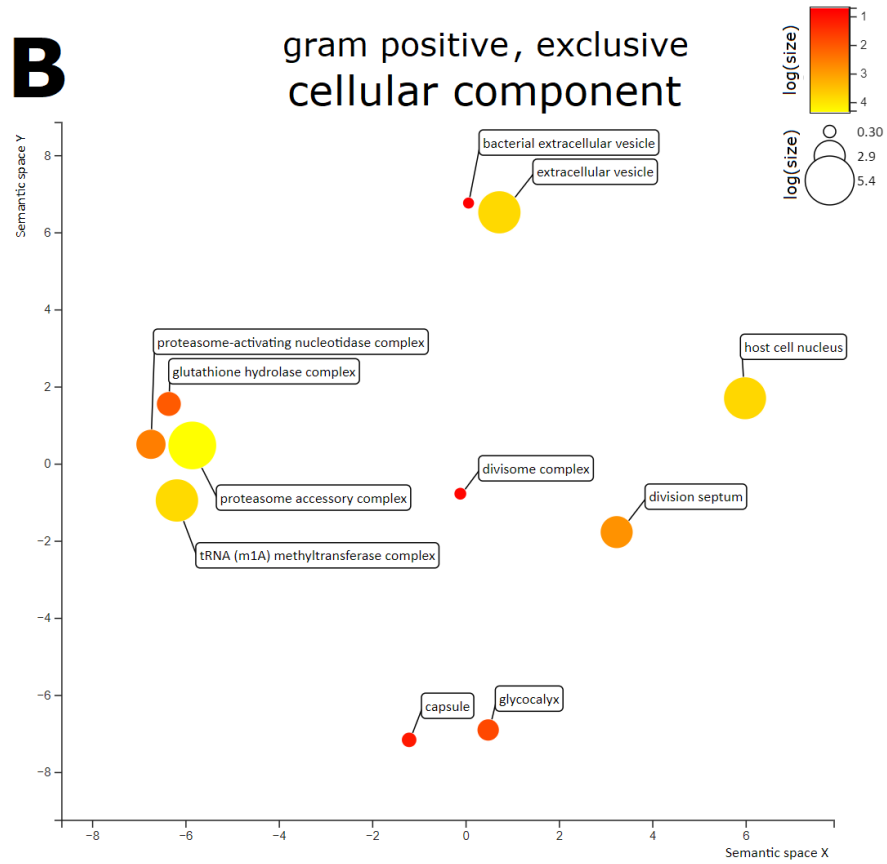


Figure B5: GO term cellular component significantly associated with gram-positivity.

A: *significantly associated with gram-positivity. Circle size denotes the number of children terms within the category. Circle colour denotes $\widetilde{OR}$, from weak (red) to strong (yellow) association.*

B*: exclusively associated with gram-positivity. Both circle size and colour denote the number of children terms within the category.*

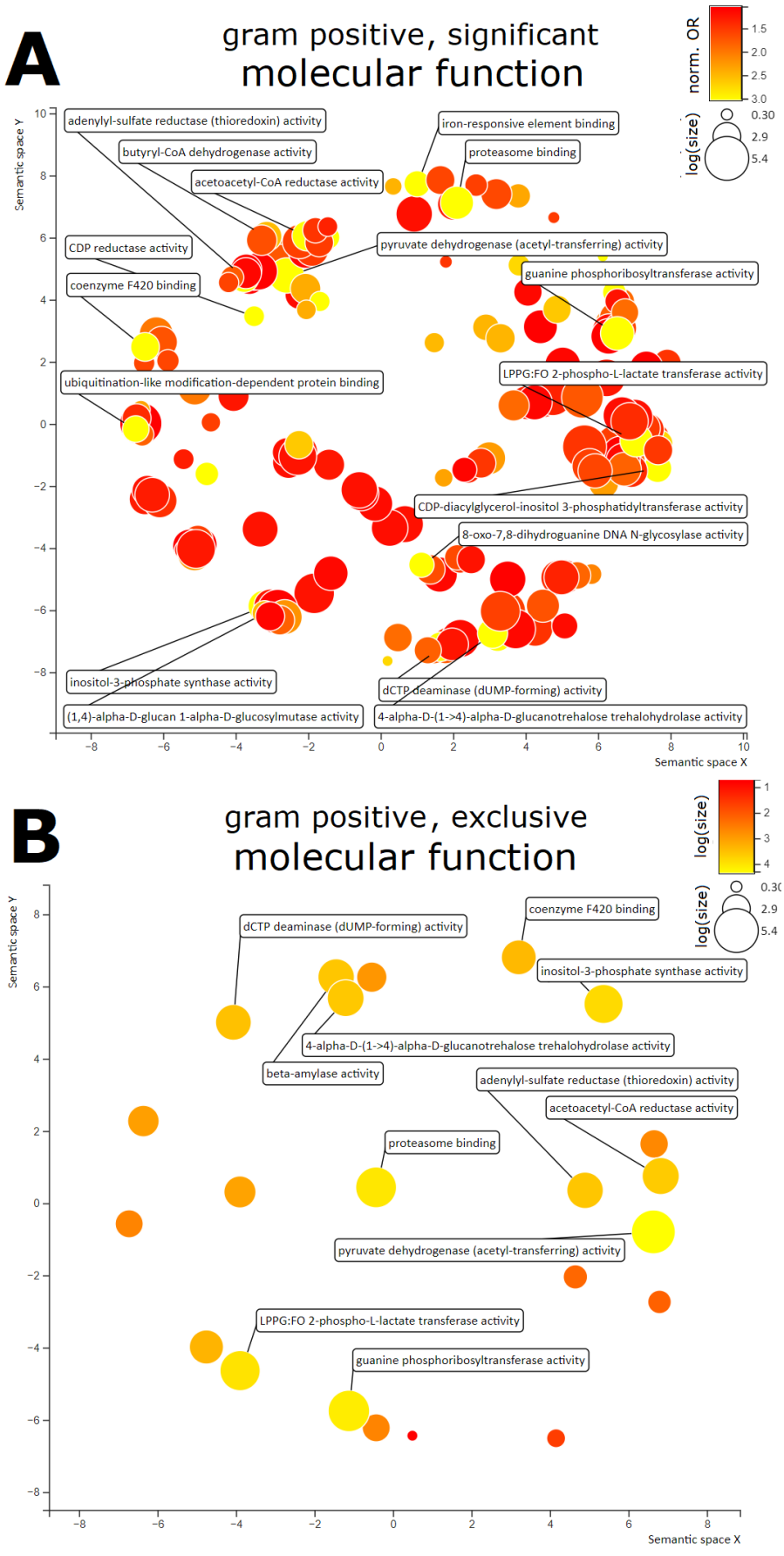


Figure B6: GO term molecular function significantly associated with gram-positivity.

A: *significantly associated with gram-positivity. Circle size denotes the number of children terms within the category. Circle colour denotes $\widetilde{OR}$, from weak (red) to strong (yellow) association.*

B*: exclusively associated with gram-positivity. Both circle size and colour denote the number of children terms within the category.*

Appendix C: Tests of annotation tools

This Appendix summarises the results of the conducted test of two tools for functional annotation of genome sequences: Prokka and eggNOG. The test saw the submission of identical genome sequences to the tools and the subsequent comparison of the Clusters of orthologous genes identified by the tools. The raw data, script, and output are available in Supplementary information 9 (App. A, p. 85).

Fig. C1 presents the term overlap produced by the two tools. For all four test species, the two tools overlap extensively. However, eggNOG saw the annotation of several terms not identified by Prokka in all four tests. Thus eggNOG was selected as the annotation tool for producing the genomic content used in the present genotype—phenotype association.

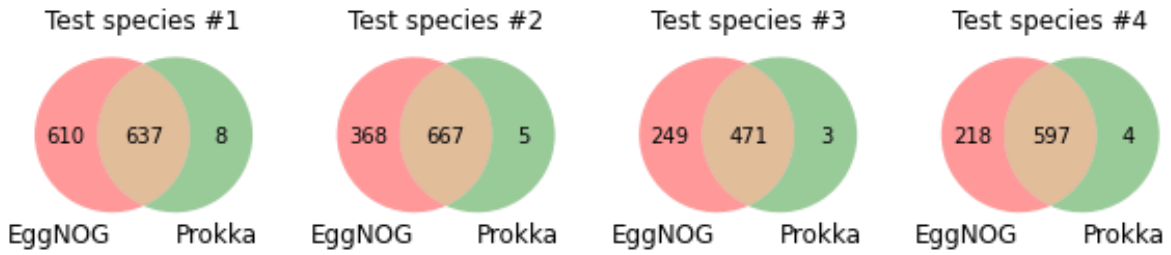


***Figure C1: Comparison of annotation tools**, indicating that eggNOG (red, left circles) was the tool with the more comprehensive output compared to Prokka (green, right circles) in all four test parallels (columns). The circles sizes are unweighted; thus the data labels are used to indicate the number of annotation term found within each source, and their intersection.*

Jenny Merkesvik

Towards genotype—phenotype association

NTNU
Norwegian University of
Science and Technology