HIGH-THROUGHPUT DISENTANGLEMENT OF ENVIRONMENTAL MICROBIOMES

The Writing of Microflora Danica – The Microbiome of Denmark

Jensen, Thomas BN

DOI (link to publication from Publisher):
10.54337/aau749596475

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

HIGH-THROUGHPUT DISENTANGLEMENT OF ENVIRONMENTAL MICROBIOMES

THE WRITING OF MICROFLORA DANICA
– THE MICROBIOME OF DENMARK

BY
THOMAS BYGH NYMANN JENSEN

PhD Thesis 2024

AALBORG UNIVERSITY
DENMARK

HIGH-THROUGHPUT DISENTANGLEMENT OF ENVIRONMENTAL MICROBIOMES

THE WRITING OF MICROFLORA DANICA
– THE MICROBIOME OF DENMARK

BY
THOMAS BYGH NYMANN JENSEN

AALBORG UNIVERSITY
DENMARK

PhD Thesis 2024

Submitted: August 2024

Main Supervisor: Professor Mads Albertsen
Aalborg University

Co-supervisor: Professor Per Halkjær Nielsen
Aalborg University

Assessment: Professor Torsten Nygaard Kristensen (chair)
Aalborg University, Denmark

Associate Professor Mette Haubjerg Nicolaisen
University of Copenhagen, Denmark

Professor David Berry
University of Vienna, Austria

PhD Series: Faculty of Engineering and Science, Aalborg University

Department: Department of Chemistry and Bioscience

ISSN: 2446-1636
ISBN: 978-87-94563-84-0

Published by:
Aalborg University Open Publishing
Kroghstræde 1-3
DK – 9220 Aalborg Øst
aauopen@aau.dk

PREFACE

This dissertation was submitted to fulfil the requirements for obtaining the degree of Doctor of Philosophy (Ph.D.) from the Faculty of Engineering and Science at Aalborg University.

The research for this thesis was conducted from January 2020 to July 2024 at the Center for Microbial Communities at Aalborg University under the supervision of Prof. Mads Albertsen and was funded by the Poul Due Jensen Foundation. The PhD project included a 3.5-month research visit at the Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA under supervision by Prof. Noah Fierer. Throughout the duration of the Ph.D. project, I was affiliated with the Department of Chemistry and Bioscience within the Faculty of Engineering and Science at Aalborg University.

Chapter 1 provides an introduction to microbiology. It offers historical context, explains how ecological theory can be applied to assemblages of microbial communities, and explores the major differences between microorganisms and macroorganisms. Chapter 2 presents the specific objectives of the PhD project. Chapter 3 serves as a walkthrough of the processes involved in the workflow from environmental sample to a DNA sequence as well as the subsequent analysis of microbial community data, thereby setting the stage for the topics of the presented papers. Chapter 4 provides a summary of the most important findings of the presented papers, placing them in the context of the current challenges in characterising microbial communities and highlighting how these findings serve to extend the knowledge within the field of microbial ecology. Chapter 5 till Chapter 12 presents one paper and its associated supplementary material at a time.

Thomas Bygh Nymann Jensen,

July 2024

“You could also ask who’s in charge. Lots of people think, well, we’re humans; we’re the most intelligent and accomplished species; we’re in charge. Bacteria may have a different outlook: more bacteria live and work in one linear centimeter of your lower colon than all the humans who have ever lived. That’s what’s going on in your digestive tract right now. Are we in charge, or are we simply hosts for bacteria? It all depends on your outlook.”
— Neil deGrasse Tyson

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Professor Mads Albertsen for allowing me to pursue a Ph.D. degree with this dissertation. Thank you for your patience, advice and helpful discussions. Thank you for allowing me to chase a few hundred ideas (while shooting down a few thousands), leaving time and space for me to learn from my own mistakes. The time invested in this project (significantly more than originally anticipated), has not only greatly improved my academic skills but has also let me learn a few things about myself.

Besides my advisor, I would like to thank all my colleagues at the Albertsen Lab and the Center for Microbial Communities for making a pleasant and encouraging working environment and making unforgettable moments at conferences and social gatherings. A special thanks to Mantas Sereika, Emil Aare Sørenen, Sebastian Mølvang Dall, Simon Knutsson, Vibeke Rudkjøbing Jørgensen, Caitlin Margaret Singleton, Thomas Yssing Michaelsen, Søren Michael Karst, Kalinka Sand Knudsen and Fransesco Delogu, without you the Microflora Danica project and this PhD dissertation would not have been possible.

I also want to express my gratitude to Professor Noah Fierer for inviting me to visit his group at University of Colorado and supervising a deep dive into shotgun metagenomics and microbial ecology. I thank everyone at the Fierer Lab for making the stay very enjoyable and memorable. A very special thanks to my friends Josep, Elias and Josh for immediately making me feel included and at home in a foreign country.

My appreciation goes to my friends and family for their encouragement and support all through my studies, as well for assisting me in putting my work aside and just making enjoyable memories. I would like to express my utmost thankfulness to my partner Sara for her unwavering love and support. I am truly grateful for her efforts in trying to understand my endeavours. Thank you for supporting me and bearing with me as I complained when things got tough.

Last but not least I send a thought towards my late grandfather, who did not live to see me graduate. I am certain that my fascination with the natural sciences was sparked by your old chemistry model set some 20 years ago.

LIST OF SUPPORTING PAPERS

This thesis is based on the following scientific papers.

Paper 1
Jensen TBN*, Dall SM*, Knutsson S, Karst SM, Albertsen M. High-throughput DNA extraction and cost-effective miniaturized metagenome and amplicon library preparation of soil samples for DNA sequencing. *PLOS ONE*. 2024; 19(4):e0301446. *Contributed equally

Theme: Benchmarking of commercially available high-throughput DNA extraction kits, miniaturisation and optimisation of preparations of amplicons and metagenomes combined with a whole genome shotgun sequencing workflow tailored for analysing the taxonomic composition of microbial communities in as complex a sample type as soil. I was leading and I was involved in all analysis of this paper including writing the draft manuscript as well as finalising the manuscript for submission.

Paper 2
Jensen TBN, Sørensen EA, Delogu F, Ramoneda J, Singleton CM, Dall SM, Knutsson S, Dueholm MKD, Jørgensen VR, Ejrnæs R, Frøslev TG, Nielsen PH, Albertsen M. The microbiome mirrors macrobiota across temperate terrestrial ecosystems.

Manuscript in preparation for submission to Nature Microbiology. (2024).

Theme: Evaluation of the response of above-ground and below-ground communities to the major ecological gradients of the temperate terrestrial ecosystems found in Denmark. This paper also provides a foundation for the expansion of commonly used Ecological Indicator Values through the identification of above-ground and below-ground indicator taxa. I was leading this paper and generated all data, except the FL16S reference sequences, based on samples collected by RE and TGF. I did all the analysis except creating the bi-partite network analysis but guided the interpretation of the results. I wrote the draft manuscript and finalised it for submission together with FD based on the comments from the co-authors.

Paper 3
Ramoneda J, **Jensen TBN**, Price MN, Casamayor EO, Fierer N. Taxonomic and environmental distribution of bacterial amino acid auxotrophies. *Nature Communications*. 2023; 14(1):7608.

Theme: Integration of taxonomy and functional traits using 16S rRNA gene sequences and genomic data to explore environmental links with amino acid auxotrophy across free-living and host-associated biological systems, which points to an association to bacterial life history strategies. In this paper I was primarily responsible for contributing methodological ideas combined with the subsequent data generation and analysis. I provided comments and edits for the draft manuscript.

Paper 4
Jensen TBN*, Singleton CM*, Delogu F, Sørensen EA, Jørgensen VR, Karst SM, Yang Y, Knudsen KS, Sereika M, Petriglieri F, Knutsson S, Dall SM, Kirkegaard RH, Woodcroft BJ, Speth DR, Aroney STN, The Microflora Danica Consortium, Wagner M, Dueholm MKD, Nielsen PH, Albertsen M. Microflora Danica: the atlas of Danish environmental microbiomes. *Contributed equally
Manuscript submitted to Nature. (2024).

Preprint available on bioRxiv: https://doi.org/10.1101/2024.06.27.600767

Theme: Generating a census of the free-living environmental microbiomes of an entire country highlighted by the construction of a new improved 16S reference database, evaluating discrepancies of alpha and gamma diversity in managed habitats, evaluating the applicability of human-made habitat classification systems for prokaryotic communities and the investigation of novel and canonical nitrifiers. CS and I led the paper. Together with SK and SMD, I performed all DNA extractions and processing of the NGS workflow. I was responsible for creating an overview of the generated data and performing the microbial 16S gene profiling of all samples, while aiding in the generation of the new 16S reference database. Furthermore, I was responsible for the evaluation of alpha, beta, gamma, core community and distance decay analysis. CS, SK and I wrote the first draft of the manuscript. I finalised the manuscript for submission together with KSK, FD and MA based on the comments from the co-authors.

CONTRIBUTIONS TO PAPERS OUTSIDE THE SCOPE OF THIS THESIS

Sereika M, Petriglieri F, **Jensen TBN**, Sannikov A, Hoppe M, Nielsen PH, et al. Closed genomes uncover a saltwater species of Candidatus Electronema
and shed new light on the boundary between marine and freshwater cable bacteria.
ISME Journal. 2023; 17(4), 561–569.

Kwan, YH, Schauberger C, **Jensen TBN**, Dall SM, Sjimabukuro M, Derycke S, Zeppilli D, Glud RN. Comparative Mitogenomics of common hadal nematodes.
Manuscript in preparation for submission to Frontiers in Ecology and Evolution. (2024).

Schauberger C, Kwan YH, **Jensen TBN**, Dall SM, Albertsen M, Nielsen PH, Glud RN. Charting Microbial Communities in Hadal Trenches: Insights into Biogeography
Manuscript in preparation for submission to Environmental Microbiology. (2024).

Hoffert M, Wilson J, Winfrey C, Coffman M, Ramoneda J, **Jensen TBN**, Yeo E, Clark K, Quispe R, Collins J, Biegert J, Gavin M, Dunn R, Fierer N. Experimental study of microbial dynamics in wild-fermented beers.
Manuscript in preparation for submission. Journal yet to be determined. (2024).

Kondrotaite Z, Petersen J, Singleton CM, Gomez MP, **Jensen, TBN**, Sereika M, Daugberg AOH, Albertsen M, Wagner M, Dueholm, MKD, Nielsen PH. Ecophysiology and niche differentiation of 3 genera of polyphosphate accumulating bacteria in a full-scale wastewater treatment plant
Manuscript In preparation for submission to The ISME Journal. (2024).

Sørensen EA, Karst SM, Sereika M, Overgaard CK, **Jensen TBN**, Knudsen K, Singleton CM, Dueholm MKD, Albertsen M. Single primer enrichment of long-read marker genes
Manuscript In preparation for submission to Nature Methods. (2024).

Petriglieri F, Singleton CM, Yang Y, Jiang C, **Jensen TBN**, Sereika M, Albertsen M, Nielsen PH. Myxococcota in Denmark: a priceless resource for drugs´ discovery
Manuscript in preparation for submission to The ISME Journal. (2024).

Danielsen ACS, Pesch C, Hermansen C, Singleton CM, **Jensen TBN**, Nielsen PH, Greve MH, Sanei H, Rudra A, Weber PL, Arthur E, Gutierrez S, Møldrup P, Normand S, Wollesen de Jonge L. Microbial community composition and diversity in carbon-rich lowland soils is shaped by soil pH. *Manuscript in preparation for submission to European Journal of Soil Biology*. (2024).

In addition, as part of the Danish Covid-19 Genome Consortium
On March 11th, 2020, a nationwide lockdown in Denmark was issued due to COVID19 and our laboratory was repurposed to whole-genome sequencing of SARS-CoV-2 virus. The first samples were processed within the following week and eventually scaled up to sequence >90% of all detected cases in Denmark. Though not included, this affected the original timeline of the PhD project.

O'Toole Á, Hill V, Pybus OG, Watts A, Bogoch II, Khan K, Messina JP et al. **[Danish Covid-19 Genome Consortium (DCGC) [Jensen TBN]]**. Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2 with grinch. *Wellcome Open Res*. 2021; 17(6):121.

Lyngse FP, Mølbak K, Skov RL, Christiansen LE, Mortensen LH, Albertsen M, Møller CH, Krause TG, Rasmussen M, Michaelsen TY, Voldstedlund M, Fonager J, Steenhard N, **The Danish Covid-19 Genome Consortium (DCGC) [Jensen TBN]**, Kirkeby CT. Increased transmissibility of SARS-CoV-2 lineage B.1.1.7 by age and viral load. *Nature Communications*. 2021; 12(1):7251.

Jessen R, Nielsen L, Cohen AS, Gunalan V, Marving E, Alfaro-Núñez A, Polacek C, **The Danish Covid-19 Genome Consortium (DCGC) [Jensen TBN]**, Fomsgaard A, Spiess K. A RT-qPCR system using a degenerate probe for specific identification and differentiation of SARS-CoV-2 Omicron (B.1.1.529) variants of concern. *PLOS ONE*. 2022; 17(10): e0274889.

Michaelsen TY, Bennedbæk M, Christiansen LE, Jørgensen MSF, Møller CH, Sørensen EA, Knutsson S, Brandt J, **Jensen TBN**, Chiche-Lapierre C, Collados EF, Sørensen T, Petersen C, Le-Quy V, Sereika M, Hansen FT, Rasmussen M, Fonager J, Karst SM, Marvig RL, Stegger M, Sieber RN, Skov R, Legarth R, Krause TG, Fomsgaard A, The Danish COVID-19 Genome Consortium (DCGC), Albertsen M. Introduction and transmission of SARS-CoV-2 lineage B.1.1.7, Alpha variant, in Denmark. *Genome Medicine*. 2022; 14(1):47.

Kurana MP, Curran-Sebastian J, Scheidwasser N, Morgenstern C, Rasmussen M, Fonager J, Stegger M, Juul JL, Escobar-Herrera LA, **The Danish Covid-19 Genome Consortium (DCGC) [Jensen TBN]**, Møller FT, Jokelainen P, Lehmann S, Krause TG, Ullum H, Duchêne DA, Mortensen LH, Bhatt S. High-Resolution Epidemiological Landscape from 290K SARS-CoV-2 Genomes from Denmark.
Manuscript submitted to Nature Communications. (2024).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS AND EXPLANATIONS

DNA: Deoxyribonucleic acid

RNA: Ribonucleic acid

SSU: Short subunit

rRNA: Ribosomal Ribonucleic acid

PCR: Polymerase chain reaction

Operon: An operon is a functioning unit of deoxyribonucleic acid containing a cluster of genes under the control of a single promoter.

ITS: Internal transcribed spacer

OTU: Operational taxonomic unit

FL: Full-length

bp: Base pair

Shannon entropy: The Shannon entropy is a commonly used metric to gauge the orderliness of symbol sequences, including DNA sequences. Conceptually, Shannon's Entropy can be intuitively interpreted as the information content.

CPR: Candidate phyla radiation

SAD: Species abundance distribution

Biotic: Consisting of living organisms

Abiotic: Consisting of things in the environment that are not living

BAM: Biotic-Abiotic-Migration

GOS: Global Ocean Sampling

HMP: Human Microbiome Project

MetaHit: Metagenomics of the Human Intestinal Tract

EMP: Earth Microbiome Project

EMP500: Subset of ~500 samples from the Earth Microbiome Project

GSBI: Global Soil Biodiversity Initiative

ICoMM: The International Census of Marine Microbes

IMG/M: The Integrated Microbial Genomes and Microbiomes

SPIRE: Searchable, Planetary-scale mIcrobiome REsource

MAG: Metagenome assembled genome

LUCAS: Land Use/Cover Area frame Survey

NEON: National Ecological Observation Network

WGS: Whole genome sequencing

NGS: Next generation sequencing

MDA: Multiple displacement amplification

ASV: Amplicon sequence variant

Distance, dissimilarity and similarity: These three terms essentially have the same meaning, as in most cases the term distance is used as synonym for dissimilarity. However, a true distance function obeys the triangle inequality. The similarity can be defined as the complement to 1 of x. That is, similarity = 1 - dissimilarity.

PCoA: Principal Coordinates Analysis, also known as multidimensional scaling (MDS)

NMDS: Non-metric multidimensional scaling

RDA: Redundancy analysis

Though trivial one note needs to be addressed on the use of microbial vs prokaryotic as well as the use of the term metagenomic sequencing. Microbial is sometimes used to denote an analysis of bacteria and archaea, when strictly speaking microbial should only be used as a description of all microbial organisms including microbial eukaryotes. Similarly, the term metagenomics is sometimes wrongfully applied within the field to denote taxonomic classification of amplicon sequences obtained from complex samples. Though the template DNA for the amplicon sequencing is of a metagenomic origin, the amplicons themselves do not qualify as metagenomic data.

ENGLISH SUMMARY

Identification of life is essential to our understanding of the world. Traditionally, the focus has been on objects we can recognise with our own eyes, such as plants and animals. However, with the introduction of DNA sequencing methods, it has become apparent that prokaryotes, small cells invisible to the naked eye, pose a vastly larger diversity than observable life. Together with the expansion of the global microbial diversity, it has become increasingly more evident that prokaryotes are deeply involved in the sustainability of the ecosphere by being facilitators of biogeochemical nutrient cycles. Today we know that microbes are everywhere on Earth, and that they impact all aspects of our lives.

Microbes rarely act on their own, but are found as members of intricate communities, which can be imagined as a coherent organism with an enormous functional repertoire. However, identifying the role of the individual community members is not trivial due to the formation of symbiotic relationships as well as interactions with the immediate physical and chemical surroundings. Furthermore, all microbes are not everywhere and understanding the factors that govern the where, how and why is critical for leveraging microbial processes in the management of natural ecosystems, mitigation of climate change, and in civil infrastructure applications.

Compared to the classical natural sciences, microbial ecology is still a young scientific field, and the last 20 years has seen even progressively ambitious large-scale projects to break new ground in microbial ecology. The focus of this PhD project is another such ambitious task; the characterization of the environmental microbiome of an entire country at an unprecedented scale. The first part of this PhD project revolves around how miniaturisation processes can be used for bringing down the costs in amplicon and whole genome shotgun sequencing from complex microbial samples. The second part takes a dive into how full-length 16S rRNA gene amplicons can increase the resolution of metagenomic derived taxonomic profiles. The work also demonstrates how the above-ground and below-ground communities respond in a similar manner to the major ecological gradients of temperate terrestrial ecosystems. Furthermore, the work highlights how key community members of the microfauna yield insights about the above-ground system and allow for the prediction of measurements reflecting the ecological gradients. The third part investigates how taxonomy derived from 16S rRNA gene amplicons can be linked to functional traits and life history strategies represented by amino acid auxotrophies across various environments. Finally, the last part of the

project highlights the major findings from the Microflora Danica project encompassing full-length 16S rRNA gene and operon sequencing of the 16S and 18S rRNA gene of 449 samples, as well as shallow metagenomic sequencing of more than 10.000 samples across Denmark. In summary, this thesis highlights and showcases how amplicon and metagenomic sequencing can complement each other in the disentanglement of environmental microbiomes.

DANSK RESUMÉ

Identifikation af livet er afgørende i forståelsen af verden omkring os. Traditionelt har fokus været på objekter, vi kan observere med vores egne øjne, såsom planter og dyr. Med introduktionen af metoder til sekventering af DNA, er det dog tydeliggjort, at prokaryoter, små celler, der er usynlige for det blotte øje, udgør en langt større diversitet end det observerbare liv. Sammen med udvidelsen af den globale mikrobielle diversitet er det blevet mere og mere evident, at prokaryoter er dybt involveret i økosfærens bæredygtighed ved at være facilitatorer af biogeokemiske næringsstofkredsløb. I dag ved vi, at mikrober er overalt på Jorden, og at de påvirker alle aspekter af vores liv.

Mikrober findes sjældent alene, men indgår som medlemmer af indviklede systemer, som kan forestilles som værende én sammenhængende organisme med ét enormt funktionelt repertoire. Identifikationen af de enkelte medlemmers rolle er dog ikke triviel grundet tilstedeværelsen af symbiotiske relationer, samt interaktioner med de umiddelbare fysiske og kemiske omgivelser. Dog er alle mikrober ikke overalt, og forståelsen af de faktorer, der styrer hvor, hvordan og hvorfor, er afgørende for at udnytte mikrobielle processer i forvaltningen af naturlige økosystemer, afbødning af klimaændringer og til applikationer inden for civil infrastruktur

Sammenlignet med de klassiske naturvidenskaber er mikrobiel økologi stadig et ungt videnskabeligt område, og de sidste 20 år har set progressivt ambitiøse storskalaprojekter for at opnå nye landvindinger inden for mikrobiel økologi. Fokus på dette PhD-projekt er en anden lige så ambitiøs opgave; karakteriseringen af det naturlige mikrobiom for et helt land på en hidtil uset skala. Den første del af dette PhD-projekt omhandler, hvordan miniaturiseringsprocesser kan bruges til at nedbringe omkostningerne ved amplicon- og helgenomsekventering fra komplekse mikrobielle prøver. Den anden del tager et dyk ned i, hvordan fuldlængde 16S rRNA-gen amplicons kan øge opløseligheden af metagenomisk afledte taksonomiske profiler. Arbejdet demonstrerer desuden, hvordan de overjordiske og underjordiske systemer har en lignende respons på de store økologiske gradienter fundet i tempererede terrestriske økosystemer. Desuden fremhæver arbejdet, hvordan centrale medlemmer af mikrofaunaen informere om det overjordiske system og giver mulighed for forudsigelse af målinger, der reflekterer de økologiske gradienter. Den tredje del undersøger, hvordan taksonomi afledt af 16S rRNA-gen amplicons kan kobles til funktionelle træk og livsstrategier repræsenteret ved aminosyre-auxotrofier på tværs af forskellige miljøer. Endelig fremhæver den sidste del af projektet de

vigtigste resultater fra Microflora Danica-projektet, der omfatter fuldlængde 16S rRNA-gen og operon-sekventering af 16S og 18S rRNA-genet på 449 prøver, samt metagenomisk sekventering af mere end 10.000 prøver på tværs Danmark. I sin helhed fremhæver og viser denne afhandling, hvordan amplicon og metagenomisk sekventering kan komplementere hinanden i udredningen af naturlige mikrobiomer.

CHAPTER 1. INTRODUCTION

1.1 CATALOGUING OF INVENTORY

1.1.1 FLORA DANICA

A central part in performing any kind of ecology study is comparison of the sampled inventory to a reference catalogue. The lack of such a Danish national index combined with inspiration by the ideas of the Age of Enlightenment and the Scientific Revolution, was probably what made the medical doctor Georg Christian Oeder propose the creation of such a catalogue in 1753; Flora Danica. The very purpose of Flora Danica was to share the knowledge of botany and by doing so to obtain greater knowledge of both the useful and harmful properties of the various plants native to Denmark. By doing so the goal was to better utilise the local resources and thereby benefit the Danish economy. 122 years later (1761-1883) Flora Danica was finished and ended up as a Scandinavian work due to the inclusion of the Norwegian and the most important Swedish plants [1]. Flora Danica is to this day still the largest printed atlas of botany.

1.1.2 THE AGE OF OMICS

Though the Age of Enlightenment ended with the coming of the 19th century, humans kept the pursuit of expanding scientific knowledge. With the more recent advances and introduction of new cutting-edge molecular techniques we have now transcended into a molecular and data-driven approach which can be described as the Age of Omics. The development of various omics technologies has enabled high-throughput quantitative monitoring of the abundance of distinct biological molecules [2] (**Figure 1**). One such molecule is deoxyribonucleic acid (DNA), with its corresponding suite of methods together termed genomics. With the discovery of the double-helix structure of DNA in 1953 [3] by Francis Crick, Rosalind Franklin, James Watson and Maurice Wilkins the foundation was laid for the strenuous task of deciphering the genetic code. A task that is still today perfected through new approaches within technologies and methods related to sequencing of DNA, data processing, analysis and interpretation. A genomic analysis allows for the genomic-scale analysis of the genetic sequence of a single individual, or of the assemblage of individuals in a community, which is the case when performing a metagenomic analysis. Analysis of the genetic sequence can yield information of the DNA under investigation in response to environmental factors, distance and time [2].

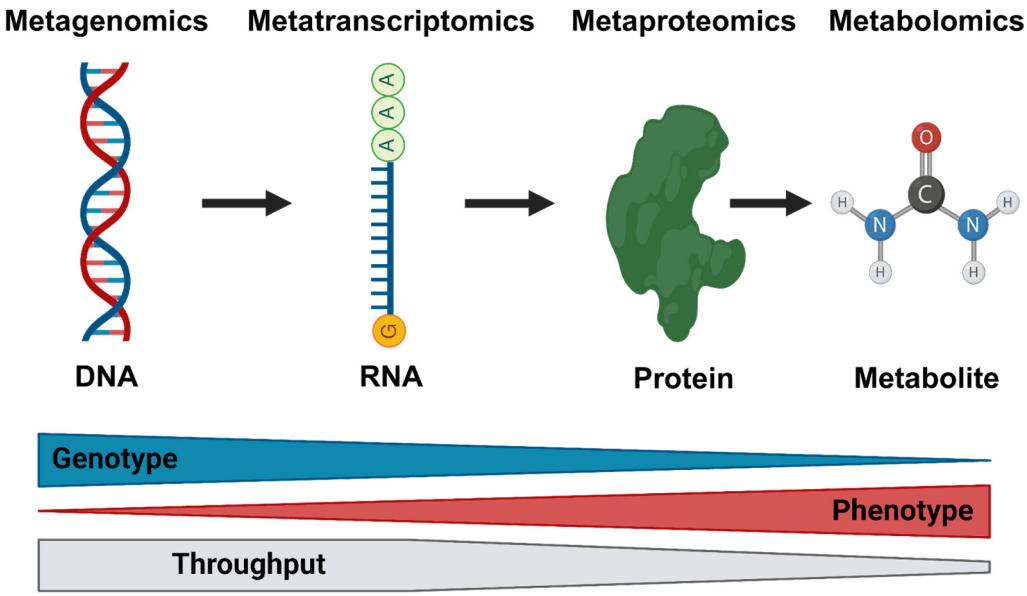

Figure 1: The omics cascade. The prefix "meta-" is used to indicate that the system under study consists of multiple species. This term is often used analogues with the characterisation of the microbial community in an uncultured sample. The aim of a meta-omic analysis is to detect and describe differences in an assemblage of microbial organisms, genes, variants, pathways or metabolic functions [4]. Each biomolecule is associated with its own field of research, each of which again is associated with its own technologies and methodologies. The cascade can be interpreted as the sequential and hierarchical flow of information through biological systems, the associated biomolecules and the effect on analytical throughput when going from analysis of metabolic potential (genotype) to realised metabolism (phenotype). Briefly, the biomolecules and resolution can be listed as: **DNA** (What can happen?), **RNA** (What appears to be happening?), **Protein** (What makes it happen?), **Metabolite** (What has happened and is happening?). Inspired by [5]. Created with BioRender.com**.**

1.2 WE LIVE IN A MICROBIAL WORLD

1.2.1 MICROBIOLOGY FOR PEOPLE IN A HURRY

Since the dawn of humankind, we have both battled and benefited from microbes. Battles in the form of infectious disease and food spoilage, and benefits in the form of fermented foods increasing the nutrient composition and availability together with the storage life [6]. Theories exist that the rise of the Sumerian empire would not have been possible without beer, due to the limited sources of clean water, and the fact that even low amounts of ethanol is enough to kill pathogens [7]. Neither would we today see the Giza pyramid in Egypt as the builders were provided a daily ration of 5 litres of beer [8]. Since the first observation of microbes almost 350 years ago, humans have endeavoured to characterise, understand, cultivate, and take advantage of microbes due to their vast importance and pervasiveness in our environment. From the time of van Leeuwenhoek's initial microscopic observation till today's more molecular-based and data-driven methods, the discipline of microbiology has seen quite a transformation (**Figure 2**).

The father of early modern medicine, Avicenna, suggested in his work of 1020 *The Canon of Medicine*, that tuberculosis and other diseases might be contagious of nature [9]. In continuation of this idea the Ottoman scholar Akshamsaddin noted that: "*This infection occurs through seeds that are so small they cannot be seen but are alive.*" [10]. In 1546 Girolamo Fracastoro similarly noted that epidemic diseases such as syphilis spread by means of *seeds of disease* [11,12]. Athanasius Kircher in 1658 suggested that infectious diseases such as bubonic plague was mediated by a *contagium animatum*, which we today know correspond to *Yersinia pestis* [12]. The first discovery of microorganisms by means of observation of bacteria was recorded by Antonie van Leeuwenhoek in 1677 through his newly invented light microscope [13]. Another early microscopist, Robert Hooke, was the first to use the term *cell* when in 1665 relating it to his findings of hexagonal shapes in cork [14].

A mere 200 years had to pass before Louis Pasteur created the first liquid growth medium (1860) [15] and through a series of experiments rejected the idea of spontaneous generation with his results instead supporting the germ theory of disease (1862) [16]. Linking specific microbes to diseases was pioneered by Robert Koch by linking anthrax to rod-shaped bacteria by microscopy in 1876 [17]. Together with his assistant Walter Hesse and his wife Fannie he developed the first solid medium using agar as a solidifying agent [17]. Simultaneously another of his assistants, Julius Petri, designed

what is now known as the petri dish [17]. These findings together resulted in the plating technique which improved the isolation and cultivation of bacteria. Koch also laid the foundation and development of staining methods, by which he identified *Mycobacterium tuberculosis*, the causing agent of tuberculosis [16]. A large step toward harnessing the potential of the microbes was with the discovery of an antibiotic substance by Alexander Fleming in 1928 [18]. Fleming found that a mould of the *Penicillium* genus inhibited the growth of staphylococci and other gram-positive pathogens [18].

The foundation of genetics was laid in 1944 by Oswald Avery, Maclyn McCarty, and Colin MacLeod, when they identified DNA as the genetic material of *Streptococcus pneumoniae [16]*. The double-helix structure of DNA was identified in 1953 [19] by James Watson, Francis Crick, Rosalind Franklin, and Maurice Wilkins, with the first two continuing their work by deciphering the genetic code [16]. In 1957 Francis Crick stated his first version of the central dogma, which dictates the genetic information flow from DNA to ribonucleic acid (RNA) (and the reverse case), and from RNA to protein. Around the time Marshall Nirenberg and Heinrich Matthaei conducted the first experiments, which showed that RNA sequences code for specific amino acids [20].

In 1977 the genome of the first bacteriophage phi X174 was sequenced [21], and in 1995 the first bacteria, *Haemophilus influenzae* Rd, was sequenced [22]. Also in 1977, Carl Woese and George Fox, discovered a third domain of life, by analysis of the small subunit (SSU) ribosomal RNA (rRNA). Comparing SSU sequences they confirmed the existence of bacteria and eukaryotes, but also observed a group equally distinct from bacteria and eukaryotes, which they named archaebacteria [23]. The domain was renamed to Archaea, due to new SSU rRNA data showing a closer relatedness to Eukarya than Bacteria [24]. With the increase in available sequencing data in the 2000s and 2010s, a two domain of life theory evolved. Central to this theory is the branching of Eukarya from within Asgard archaea, which to this day is still under debate [24–27]. The increase in available sequencing data can be attributed to the invention of the polymerase chain reaction (PCR) method by Kary Mullis in 1983, in which a small amount of DNA can be copied to large quantities over a short period of time [28,29].

These discoveries are only a short and condensed walkthrough of the history of microbiology, and other major discoveries have been left out, and new discoveries published every day.

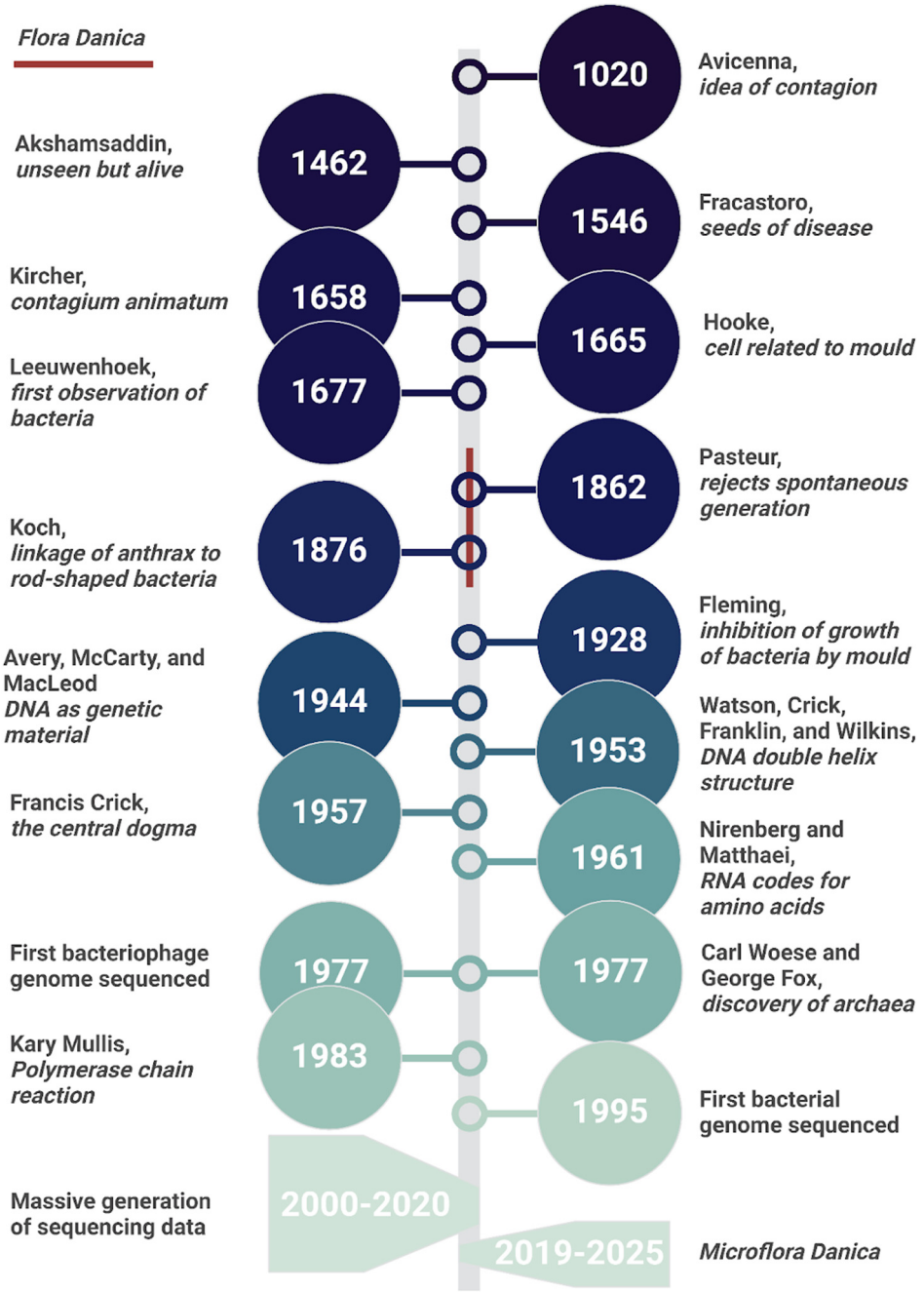

Figure 2: Microbiology for people in a hurry. Historical overview of selected major scientific breakthroughs in microbiology from the last millennium. The timeline for the production of Flora Danica was added to visualise the enormous task it was to compile. Created with BioRender.com

1.2.2 THE MICROBIAL DIVERSITY

When recording the plants (Archaeplastida within Eukaryotes on **Figure 3**) for Flora Danica, Georg Christian Oeder probably did not expect that the environment surrounding the plants harboured a vastly larger diversity in the form of microorganisms. Current estimates set the global number of plant species not to exceed 400,000 of which ~94% are associated with an accepted species name [30,31]. In 2011 Mora and colleagues estimated the planet to harbour ~7.8 million animal species, 298,000 plants, 611,000 fungi, and 63,900 protists of which only 1.2 million at that time were described [32]. Though the number of species in each of these categories should probably be increased, the most drastic difference between their and current estimates is for prokaryotes. Though they acknowledged that their projections of 10,000 bacteria and 500 archaea were likely underestimated, they are very different from later findings putting the global number of microbial species in the range of millions to trillions, which is still a subject of controversy [32–39]. Species of macrobiota, such as plants, can readily be described and distinguished by means of observation, which is not the case for microbes. Furthermore, the species concept as it is defined for macrofauna; *"a group of living organisms consisting of similar individuals capable of exchanging genes or interbreeding.",* does not apply to prokaryotes and some eukaryotes as they perform asexual reproduction and are able to perform horizontal gene transfer - potentially even between the two prokaryotic domains [40]. Hence, an estimation of the global microbial diversity will be influenced not only by the computational methodology used, but also on the definition used for a species representative.

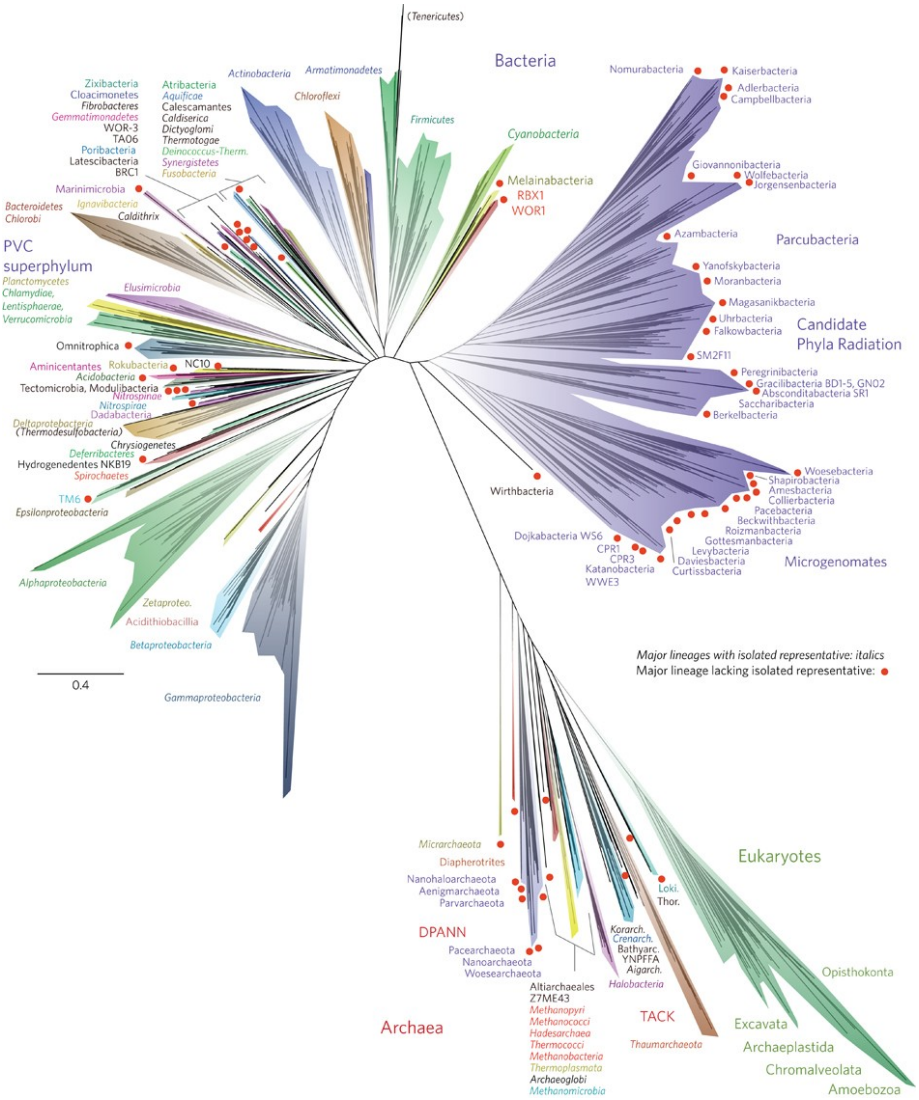

Figure 3: The tree of life. In 2016, Hug and colleagues updated the tree of life using 16 ribosomal proteins derived from genomes and metagenome assembled genomes rather than the ribosomal RNA genes. The tree presents 92 bacterial phyla, 26 archaeal phyla, and all five eukaryotic supergroups, with major lineages colour-coded and named, using italics for well-characterised lineages. A notable aspect of this tree is the numerous major lineages that lack isolated representatives, indicated by red dots. Eukaryotic supergroups are identified, albeit without detailed delineation due to their low resolution. The candidate phyla radiation (CPR) phyla, consisting solely of organisms lacking isolated representatives and still under definition, is assigned a single colour. The relationship between Eukarya, Bacteria, and Archaea remains debated, with the emergence of Eukarya thought as an endosymbiotic event between bacterial and archaeal cells. The updated phylogenetic tree structure shows variations from the findings of Woese and colleagues in 1990 [41], highlighting how the selection of taxa or proteins for phylogenetic analysis can significantly impact results. Adapted from [42].

For the study of the microbial content of a sample, the ribosomal RNA operon is the most widely used phylogenetic marker for taxonomic profiling. The operon is universally present due to its vital role in the metabolic functioning of all forms of life. Since the 18S gene provides lower resolution, it is common to use the internal transcribed spacer (ITS) between the small- and large-subunit rRNA genes to distinguish microbial eukaryotes. The 16S gene is ~1500 bp long and is composed of 9 variable regions (V1-V9) in combination with conserved stretches of nucleotides (**Figure 4**). The highly conserved regions allow for binding of amplification primers as well as regions with high nucleotide variability contributing with taxonomic resolution (**Figure 4**) [43,44]. The V4 region is commonly used as a target for analysis of prokaryotic diversity, especially in soil, and is also recommended by the most comprehensive single study on the microbiome of the planet, the Earth Microbiome Project [45]. V4 operational taxonomic units (OTUs) are often used with a cut-off of 97% sequence similarity to be representable of a species, though data suggest that 100% is more appropriate for the V4 region and 99% for full-length (FL) 16S sequences [46]. The open-reference Earth Microbiome Project comprises 307,572 unique 90 base pair (bp) V4 sequences after subsetting and rarefying to 5,000 reads across 23,828 samples. The subset covering 150 bp of the V4 region amounts to 202,540 sequences [35].

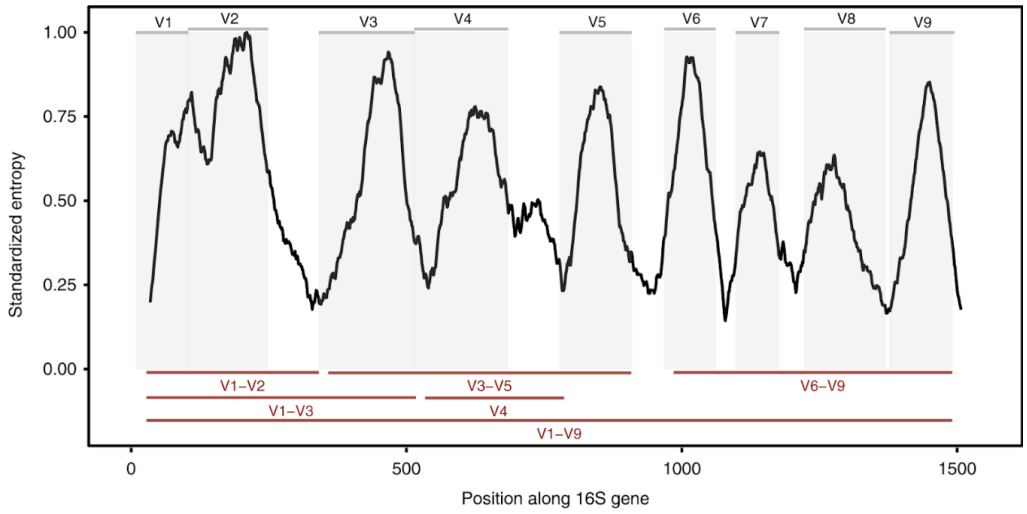

Figure 4: The prokaryotic small-subunit ribosomal RNA gene. Shannon entropy (distance from homogeneity of DNA bases) across the 16S rRNA gene. The figure was constructed based on alignment of a single representative of all species in the Greengenes database v13.5. The 16S rRNA gene sequences were aligned against Escherichia coli K-12 MG1655 (NCBI Gene ID 947777). The variable regions (V1-V9) are highlighted from commonly used primer binding sites (grey panels) and in-silico regions based on entropy calculations (red lines). Adapted from [44].

The current debate on the global prokaryotic diversity seems to be divided between supporting an estimate in the low millions or the low trillions [32–38]. Locey and Lennon estimated that Earth harbours as many as 1 trillion microbial species based on an approach with theory from macroecology and biodiversity [33]. The authors applied a unified scaling law to predict the abundance of dominant species across 30 orders of magnitude. They combined the predicted number of dominant species with the use of the lognormal species abundance distribution model to arrive at their estimate. The scaling laws, also known as power laws, describe the relationship between species richness (S) and number of individuals in an assemblage (N) as $S \sim N^z$, with z being an empirical or theoretical derived constant. The lognormal distribution model suggests that a right-skewed frequency distribution is approximately normal when log-transformed [47]. The existence of ~10^{30} prokaryotic cells on the planet was assumed. For the estimation of z the authors compiled a dataset, which was primarily made up of a 14,615 site subset of the closed-reference 97% V4 OTU dataset from the Earth Microbiome Project [45]. Besides the V4 OTUs from the Earth Microbiome project, they included V3-V5 OTUS from the Human Microbiome project and rRNA amplicons from MG-RAST including 128 samples of fungal ITS. They did not elaborate on the total OTU count in this dataset. Furthermore, Lennon recently published a paper together with Fishman that revealed how richness beyond the previous estimate of 1 trillion is feasible and in agreement with empirical predictions [39].

Contradicting the previous finding, Louca and colleagues compiled a 97% V4 OTU dataset of 34,368 samples across 492 studies and named it the Global Prokaryotic Census. The authors deliberately excluded data from the Earth Microbiome Project for downstream evaluation against this dataset. After filtering for OTUs present in at least two samples of the same study the result was 739,880 prokaryotic OTUs (690,474 bacterial and 49,406 archaeal) [35]. Sequences in this composite data set cover at least 200 base pairs in the V4 hypervariable region. By considering each study an independent sampling unit they estimated the existence of 0.8-1.6 million V4 OTUs on the planet, which they extend to 2.2-4.3 million full-length OTUs clustered at 97% identity. With clustering at 99% similarity, they predicted the existence of 3-9 million V4 OTUs. They further substantiated this claim with fitting the data to a log-normal distribution curve which resulted in the estimation of 886,291 prokaryotic 97% V4 OTUs. The authors performed extrapolation of 6 different statistical richness estimators, utilising sampling theory-based methods to take into consideration the frequencies of low-abundance classes (e.g., singletons, doubletons, etc.) [33,35,36]. Unlike methods based on models of biodiversity, these statistical estimators employ more available data and do not assume anything about biological

processes [33,36]. They acknowledged that these results contrast with the speculations of the vast “rare microbial biosphere” proposed by Locey and Lennon as well as others. Finally they showed how, at 97% similarity, the Global Prokaryotic Census captured 89-96% of prokaryotic sequences in public databases and recaptured 92% of the sequences in the 150 bp Earth Microbiome Project subset mentioned previously [35].

In response to the estimate 6 orders of magnitude lower prokaryotic diversity by Louca and colleagues, Locey and Lennon published another paper in 2020. They attributed the discrepancy to violations of sampling theory, misuse of biodiversity theory and performed a reanalysis of the Global Prokaryotic Census dataset. This led to an estimate of ~10^{14} microbial taxa, which is in accordance with their previously reported finding of 10^{12}. This comes as no surprise as the GPC recaptured 92% of the sequences in the 150 bp Earth Microbiome project dataset, which comprised 70% of the dataset originally used by Locey and Lennon [33]. They elaborate on how the statistical estimators used in microbiome research operate under the assumption that all taxa, even those not observed, are present during sampling and that the samples accurately reflect the ecosystem being studied. They highlight that the Global Prokaryotic Census dataset, despite being a substantial collection of 16S rRNA gene sequences, primarily features samples from central North America, central Europe, and Eastern China, leaving large areas of the world underrepresented or unrepresented. They note that this geographical bias deems this dataset non-representative of the microbiome of Earth, ultimately leading to underestimation of the global microbial diversity from statistical estimators.

It should be noted that the estimate from Louca and colleagues has been disputed in a publication from Wiens, due to underestimation of host-associated species [37]. Specifically, the contribution from species associated with insects, as a large fraction of these are estimated to be currently unrecorded [34,37,48]. In 2017, before the publication of the estimate by Louca and colleagues, Larsen and colleagues estimated 0.209 to 5.8 billion species on Earth, with 66% to 91% being bacteria [34]. Wiens supported this range with an estimate of 0.183 to 4.2 billion species, with 58% to 88% bacteria [37]. In 2021 Louca and colleagues published a response on how the previously estimated host-associated bacterial richness would decrease with better justified statistical models [38]. They finally highlight that all estimates presented are based on 16S rRNA OTUs, and that at finer phylogenetic resolution enabled by whole-genome sequencing (e.g. bacterial strain level or ecostype), might reveal substantial higher bacterial richness than their estimates [38].

As presented in this chapter, there exists a substantial disagreement about the global prokaryotic diversity. The 6 order of magnitude difference is founded on fundamental disagreement on the appropriate approach used for modelling of bacterial diversity of the rare biosphere on a global scale. One thing the two datasets do have in common is that they both suffer from geographical bias, since large geographical areas remain unsampled to this day [49–52]. Furthermore, the published estimates do (almost) not include the suspected vast diversity of microbial eukaryotes. Furthermore, while sub-region targeting was once adequate for identifying taxa at the genus level, it lacks the precision to distinguish species, and the suitability of specific sub-regions depends on the taxa being studied [44]. Today, this compromise is challenged by long-read sequencing technologies, allowing for assembly-free sequencing of full-length small-subunit rRNA genes or transcripts [53,54]. Full-length 16S sequences offer better taxonomic resolution than targeting sub-regions (especially V4), which often fail to accurately reflect the diversity in the original data [44] (**Figure 4**). Finally, a very recent publication on the maintenance of bacterial rRNA operons on plasmids adds for an interesting discussion on the implications of horizontal gene transfer of the prokaryotic phylogenetic marker gene itself [55]. From a genomic perspective the average nucleotide identity (ANI) is used to define the species boundaries of prokaryotes with a value of 95% representing the boundary between two different species. The use of this species definition might alone tremendously affect previous estimates of the global diversity. As a final remark it should be noted, that though there exists disagreements between the number of prokaryotic species by some orders of magnitude, the estimates of microfauna most likely outnumber the macrofauna by quite some margin (**Figure 3**).

1.3 MICROBIAL COMMUNITIES

1.3.1 MICROBES AND THEIR SURROUNDINGS

Microbes are essential for the sustainability of the ecosystem they inhabit by facilitating biogeochemical nutrient cycles (**Figure 5**) [56,57]. Microbes also have a direct influence on the behaviour and health of all animals including humans. It is known that bacteria are necessary for human digestion, nutrient absorption and immunity [58], as well as being capable of causing or treating disease in people [59]. Microbes living below the surface of terrestrial ecosystems affect the biodiversity, composition and health of the aboveground fauna [60–63]. Additionally, evidence points to the relationship between soil biodiversity and the ecological and evolutionary responses of terrestrial ecosystems to recent and future environmental change [63–67]. Understanding the factors that control these processes is critical for leveraging microbial processes in management of natural ecosystems, mitigation of climate change, and in civil infrastructure applications including agriculture, biogas production, wastewater treatment, and value chemical production [57] (**Figure 5**).

Historically, microbes have been characterised by means of morphology enabled by microscopic investigation. However, it has become evident that not all microbes are easily cultured and isolated, due to complex metabolic strategies or the need for the presence of symbiont organisms [68]. This has led to the description of new species solely based on reconstructed genomes derived from environmental samples. However, the processes a microbial cell can carry out, as determined by its genotype, is not necessarily equal to the expressed phenotype under the environmental conditions, from which it was sampled. Linking the genomic derived functional potential to actual function requires validation studies of cultured organisms. Advanced culturing methods have been developed to culture a wider range of microbes [69], but the task is still a tedious process, why most prokaryotes remain uncultured [70]. Adding yet another layer of complexity to this problem, is the fact that many genes are still of unknown function or described as encoding for a hypothetical protein.

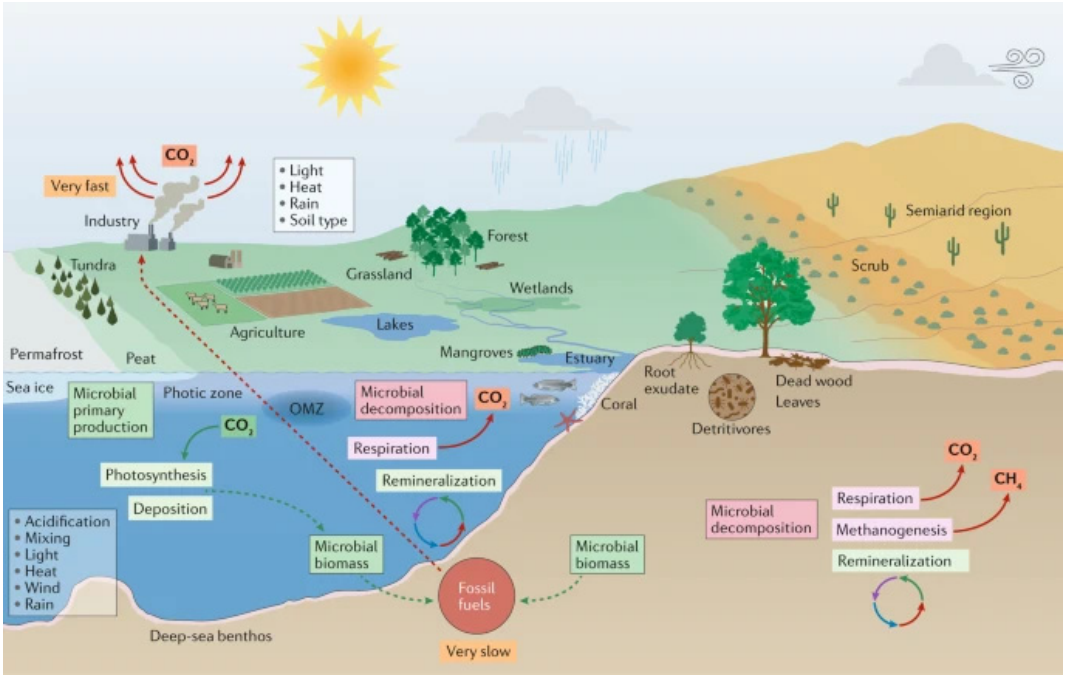

Figure 5: Microorganisms and climate change in marine and terrestrial biomes. Microbial primary production in marine environments plays a significant role in CO_2 sequestration, while also recycling nutrients for the marine food web. On land, microorganisms are crucial decomposers of organic matter, enriching the soil with nutrients for plants while emitting CO_2 and CH_4 into the atmosphere. Over millennia, microbial biomass and organic matter transform into fossil fuels, which, when burned, release greenhouse gases rapidly, disrupting the carbon cycle and causing atmospheric CO_2 levels to rise as fossil fuel consumption continues. Human activities, including agriculture, industry, transport, and consumption, combined with environmental factors such as soil type and light, significantly influence the complex interactions between microorganisms and other organisms. These interactions play a crucial role in how microbes affect and are affected by climate change. In turn, climate change factors such as rising CO_2 levels, increased temperatures, and changes in precipitation patterns also impact microbial responses, with broad implications for the entire ecosystem. OMZ, oxygen minimum zone. Adapted from [71].

1.3.2 MICROBIAL COMMUNITY ASSEMBLY

Assemblages of co-occurring microbial species in space and time is the defining feature of a microbial community [72] (**Figure 6**). The diversity and structure within these communities are highly habitat dependent with few different species present in a sourdough starter culture compared to thousands of species in a gram of soil [73]. Diversity, also known as richness, here refers to the number of different species within the observed community and is linked with the term α-diversity. The term structure relates to the composition and relative abundance of these species in a sample and the differences between samples reflecting β-diversity. The richness can differ significantly between communities, but it generally follows that the majority of taxa are low abundant [74] (**Figure 6**). Furthermore, a large fraction of this rare biosphere might even be in a dormant state [72]. This phenomenon aligns with the universally observed species abundance distribution (SAD), a fundamental ecological pattern that explores the commonness and rarity among species [72,75] (**Figure 6**).

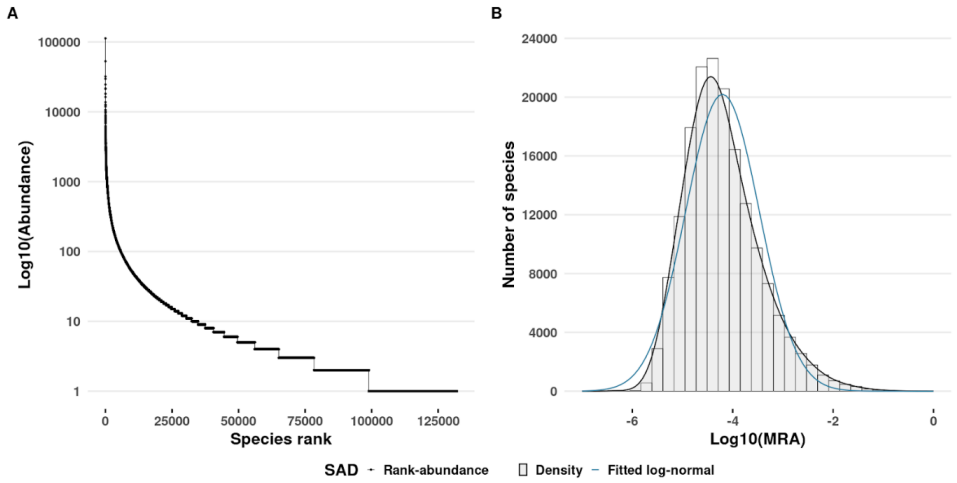

Figure 6: Species abundance distributions. A. Rank-abundance plot, where each point represents the abundance of a single species - here defined based on sequence similarity. The plot highlights the characteristic that a small number of organisms are highly abundant, whereas the majority of taxa occur at low abundances. Inspired by [72]. **B.** Species abundance distribution (SAD) of **A** on a log-scale. A slight positive skew is observed compared to the fitted log-normal distribution. Inspired by [35]. Both figures were generated from the combined Microflora Danica V1-V8 16S rRNA gene sequence datasets from 457 samples [**Paper 4**].

A feature of biological communities is the presence of a disproportionate number of low-abundant species [72]. This feature is even more pronounced for microbial communities, which possibly arise due to the

relative scale that microfauna is examined at, as compared to macrofauna [72]. When the abundances are mapped on a logarithmic scale, they approach a normal distribution (known as the log-normal distribution) [75]. The species abundance distribution is crucial to the existing theories of biodiversity and macroecology, which seek to comprehend patterns of abundance, distribution, and diversity across different spatial and temporal scales [47]. Though there exists debate on the importance of the low abundant taxa, communities with many different taxa harbouring the same genetic potential adds functional redundancy, which can be linked to the overall stability and resilience of the ecosystem [76].

Microbial communities can show a high degree of complexity not only due to the species diversity but also through the numerous metabolic processes they perform. These processes are often linked due to the fact that microbes are interdependent, which results in ecologically important yet complex symbiotic interactions [77] (**Figure 7**). Microbial interactions are also determined by the life history strategy of the members of the community [78,79]. Life history strategies shed light on how bacterial populations balance energy use between growth, resource acquisition, and survival [80]. In relation to different resource acquisition strategies, auxotrophy refers to an organism's reliance on external sources for essential metabolites due to gaps in their biosynthetic pathways [81]. This often indicates a dependence on symbiotic relationships for survival or selection for presence in nutrient-dense environments. Conversely, prototrophy denotes metabolic independence of an organism, as it has the ability to produce all necessary metabolites [81]. These symbiotic interactions can range from mutualistic, where members exchange metabolites for shared growth, to parasitic, where parasites infiltrate living cells for exploitation of the metabolic machinery of the host, inflicting damage in the process (**Figure 7**). As an example, a mutualistic interaction occurs when two or more organisms collaboratively metabolise a compound, with one organism consuming harmful metabolic byproducts produced by the other. This process, known as syntrophic degradation, benefits both parties: it aids the producer by removing detrimental metabolites, thereby encouraging beneficial metabolic pathways, while the consuming organism gains nutrients (**Figure 7**). This framework helps understand the complex behaviours of microbial communities and their adaptive strategies for survival and growth (life history strategies). The wide range of such interactions within these communities has a direct impact on the community's overall structure and stability [82]. The communities can be quite stable over time but might also be sensitive to environmental shifts, which can alter their composition and function, making them useful indicators of environmental health [83].

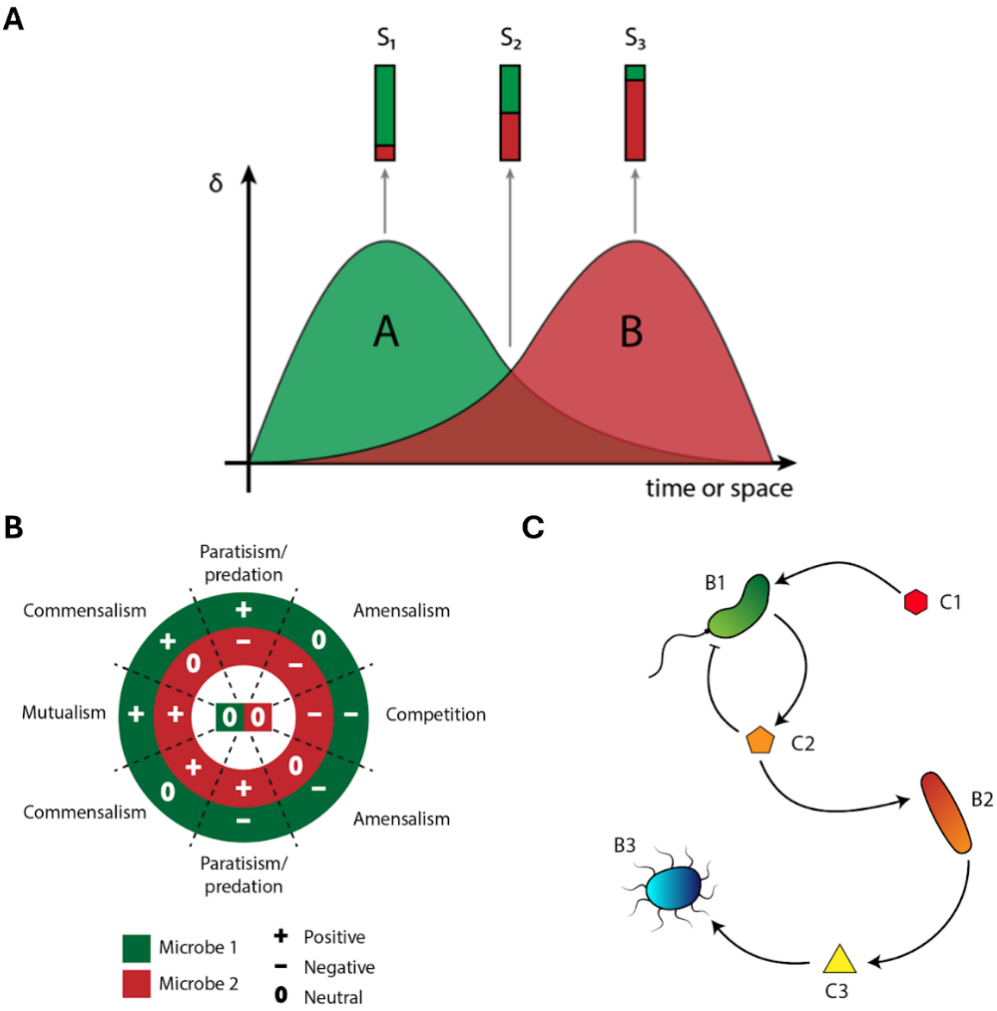

Figure 7: Microbial community assembly. A. Species abundance distribution across space or time. Differences in habitat suitability for different species are influenced by time or geography and generates a positive relationship between the abundance and the distributions of species. Inspired by [75]. **B.** Types of biotic interactions. For every participant in an interaction, three potential outcomes exist: beneficial (+), detrimental (−), and neutral (0). Take parasitism as an example: the parasite gains advantage (+) from the interaction, while the host suffers harm (−), leading to the representation of this relationship with the symbols + −. Inspired by [84]. **C.** Schematic of metabolic interactions. A simple syntrophic schematic for the degradation of the compound C1, bacteria B1 metabolises C1 to the metabolite C2, which inhibits the continued growth of B1. B2 can metabolise C2 to C3, and hence the interaction between B1 and B2 can be categorised as mutualistic. Downstream B3 metabolises C3 leading to a commensalistic relationship between B2 and B3. Inspired by [85].

1.3.3 MICROBIAL ECOLOGY AND BIOGEOGRAPHY

Microbial biogeography

Microbial biogeography explores the spatial and temporal distribution patterns of microorganisms, seeking to decode their presence across Earth. This scientific field examines the reasons why certain microbes are found in particular geographic areas but not others, how microbial communities evolve over time, and what historical, environmental, and evolutionary forces influence these distribution patterns [86,87]. By examining microbial dispersal across continents, environmental gradients, and different habitats, microbial biogeography often considers the roles of dispersal, drift, selection, and speciation in determining where microorganisms are found [86–88]. The aim is to understand why certain microbes are found in particular geographic areas but not others and how factors like climate, geography, and historical events shape these distribution patterns [86,87,89]. Through mapping microbial presence and analysing distribution patterns, this field endeavours to predict how microorganisms respond to environmental changes, providing insights into the ecological distributions of soil microbial diversity, community composition, and functional traits from regional to global scales [87–89]. Additionally, microbial biogeography can provide insights into how microbes adapt to habitat changes, especially with regards to climate change, and thereby attempt at predicting the microbial responses to environmental shifts. By doing so, microbial biogeography can provide insight into the underlying mechanisms that generate and hinder biodiversity [86].

Microbial community ecology

Microbial community ecology is concerned with the relationships between microorganisms and their physical surroundings across space and time [72,90]. It examines how microbes contribute to and are affected by the ecosystems they are found in. This field explores the roles of microorganisms in topics such as biogeochemical cycling (such as carbon and nitrogen cycles), their impact on environmental health and disease, community composition and dynamics, and the ecological functions they perform [71]. Microbial ecology often looks at the diversity and structure of microbial communities, how these communities are structured and function within ecosystems, and how environmental factors and biological interactions influence these communities [72]. Furthermore, microbial community ecology can aim at evaluating the importance of specific community members, predicting environmental habitability for specific taxa, predicting the overall community stability and evaluating how the communities respond to disturbances [91].

Summary of Differences

Microbial biogeography and microbial ecology are closely related yet distinct fields within microbiology that both study various aspects of microorganisms and their environments. Microbial community ecology and microbial biogeography both seek to explain species abundance distributions across space and time. At its core both aim at answering the questions: Where are they, how many are they, why are they here but not there, and what are they doing? Biogeography explores the "where" of microorganisms, mapping their spatial distribution and understanding their geographical patterns. Ecology, however, delves into the "how" and "why," examining the roles of microorganisms, interactions within ecosystems, and their ecological relationships with both living and non-living elements. At their intersection biogeographical insights enhance our understanding of microbial ecology and vice versa. Together, these fields provide a comprehensive understanding of the importance of microorganisms to the ecosystems of Earth, their impact on ecosystem health and stability, and how microorganisms adapt to environmental changes. Given the existence of an overlap in the topics of the two fields, it is problematic that similar patterns and processes are occasionally described using inconsistent nomenclature in both frameworks [76,90].

1.3.4 ECOLOGICAL THEORY

Microbial communities of free-living microbes follow the species-area relationship which dictates that larger areas tend to harbour larger species richness. On the same note turnover in microbial communities are observed across geographical distance and environmental gradients resulting in distance-decay of the community composition [72]. Different conceptual frameworks exist for classifying the drivers of the species-area relationship. Among these is the conceptual synthesis of community ecology as proposed by Vellend [72,92] or the Biotic-Abiotic-Migration (BAM) framework proposed by Soberón and Peterson [93,94] (**Figure 8**).

Vellend's conceptual synthesis of community ecology as presented by Nemergut et al.

A central inquiry in ecological science revolves around how diversity arises and is sustained. The processes shaping genetic diversity within a species are typically seen as evolutionary in nature, encompassing mutation, selection, gene flow, and genetic drift. In contrast, the creation of diversity among species is generally attributed to ecological factors. Vellend has drawn a comparison between these evolutionary processes and four

corresponding ecological mechanisms: diversification (speciation), selection, dispersal, and ecological drift.

Diversification. The generation of new genetic variation. Diversification has been proposed in place of speciation to acknowledge changes in community structure that can occur without the emergence of new species [95].

Dispersal. Movement of organisms across space with time.

Selection. Changes in community structure caused by deterministic fitness differences.

Drift. Stochastic changes in the relative abundance of different species within a community through time.

Vellend's conceptual framework presents several advantages. It integrates both deterministic elements like selection and stochastic elements such as ecological drift, thereby reconciling niche and neutral theories. Three of these processes - dispersal, drift, and diversification - are fundamental to neutral theory. The framework also emphasises the role of evolutionary processes, particularly diversification, in shaping community structures, acknowledging how both evolutionary and ecological forces interplay to affect diversity and structure. Additionally, it offers a flexible model for comparing communities across different ecosystems under a unified theoretical lens. Thus, Vellend's approach has the potential to evolve microbial ecology into a field defined by mechanistic and predictive approaches rather than just descriptive observation. However, while Vellend's framework is well-regarded, especially in microbial ecology, pinpointing the exact effects of the four processes on community assembly is complex. Moreover, converting this theoretical model into a practical, quantitative framework poses even more complex challenges.

Soberón and Peterson’s Biotic-Abiotic-Migration (BAM) framework as presented by Malard and Guisan

The BAM framework builds on an expansion of the concept of environmental niches proposed by G. Evelyn Hutchinson in 1957 [93,96]. A Hutchinsonian niche is conceptualised as a multi-dimensional space where each dimension represents different environmental conditions and resources required for the indefinite survival and reproduction of a species [96,97]. Hutchinson identified two primary types of niches: the fundamental environmental niche and the realised niche. The fundamental environmental niche represents the ideal range of conditions a species

could theoretically thrive in. The realised environmental niche is the more restricted set of conditions within which a species actually lives in the natural world, taking into account interactions with other living organisms such as mutualism and competition [98] (**Figure 8**).

The concept of the realised environmental niche was further expanded to incorporate three principal categories of factors that influence where a species can be found geographically. Firstly, there are abiotic factors (A), such as climate and the physicochemical factors of the environment, which set the physiological bounds for the survival of a species in a locale (**Figure 8**). Secondly, biotic factors (B) involve interactions with other organisms that can either enhance (like mutualism) or hinder (such as competition) the ability of a species to sustain its populations, thus influencing its geographic spread (**Figure 8**). Thirdly, factors related to the capacity of a species to reach and establish itself in new areas are crucial (M) (**Figure 8**). This aspect, which takes into account natural barriers and the dispersal capabilities of a species, helps to discern the actual distribution of a species from potential areas deemed suitable solely based on abiotic and biotic factors. The intersection of these three sets of factors defines the range of environmental conditions truly inhabited by the species, providing a broader definition of the realised environmental niche commonly applied in ecology.

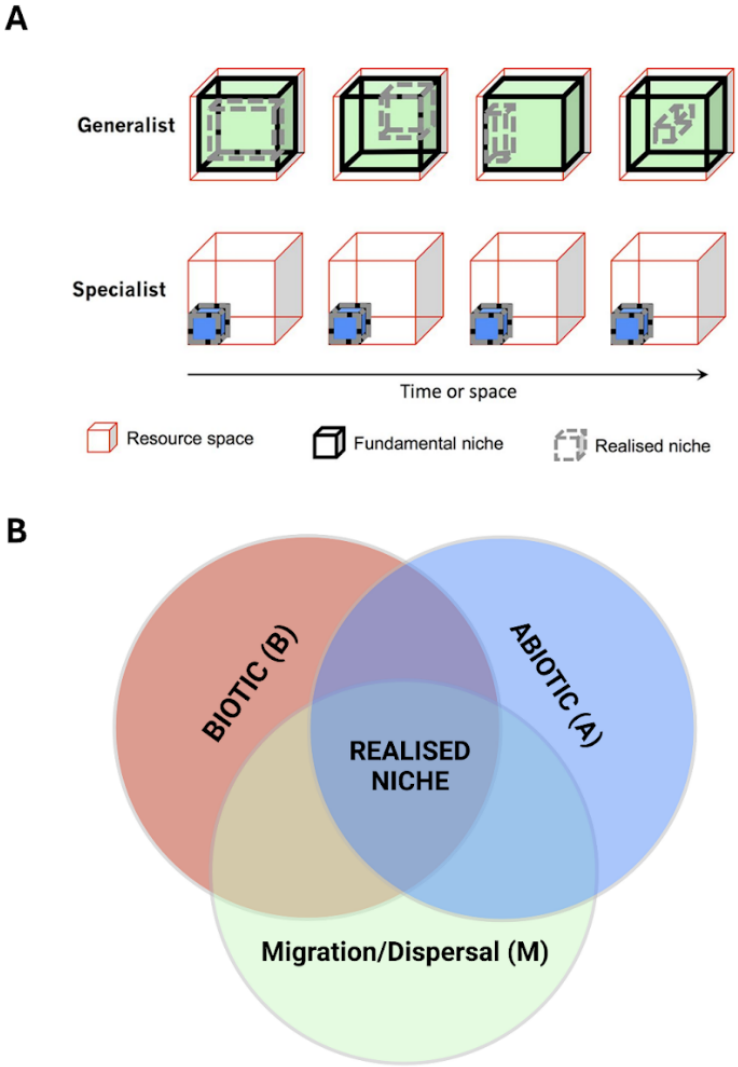

Figure 8: The microbial niche. A. The fundamental and realised niche. The resource space can be thought of as a multidimensional representation of physicochemical factors. The fundamental niche of an organism in the resource space, can be described from the genomic information. The part of the fundamental niche actually occupied at a moment in space and time is described as the realised niche. A generalist will utilise varying portions of its fundamental niche based on the surrounding biotic and abiotic factors. On the other hand, a specialist organism, having a very limited niche breadth, will consistently carry out only the limited functions that its genome encodes. Adapted from [98]. **B.** The Biotic-Abiotic-Migration framework by Soberón and Peterson [93]. The realised environmental niche of a species is shaped by three key sets of factors: abiotic conditions (A), such as climate and physicochemical properties, which set physiological limits; biotic factors (B), including interactions with other species, influencing population maintenance and distribution; and migration (M) to potential colonisation sites, dictated by geographic barriers like oceans or mountains and the dispersal capabilities of the organism. Inspired by [93,94].

1.3.5 FROM MACROBIOTA TO MICROBIOTA

Though there exist many parallels between the ecology of macro and micro fauna, some major differences exist between the two, which might lead to discrepancy in the importance of different processes to the community ecology of microbes. First, the distinction of different microbial species is fundamental for many of the topics in studying ecology of microbial communities. The traditional biological species concept, as proposed by Ernst Mayr, defines species based on reproductive isolation and genetic exchange [76]. However, this concept falls short for prokaryotes which reproduce asexually [76]. An alternative definition is the ecological species concept proposed by Cohan [99]. The ecological species concept presents a species as a group of individuals sharing identical key ecological traits (ecotypes). The hypothesis is based on the assumption that bacteria occupy specific niches where periodic selection eliminates genetic variation within each niche but allows for differentiation of inhabitants in separate niches. Consequently, species that are distinct both genetically and ecologically can emerge when there exist minimal or no genetic recombination. This approach suggests that ecological principles acknowledging these distinct species are relevant to bacteria and predicts a direct correlation between molecular and ecological diversity [76].

A mediator of shared ecological traits between different species is horizontal gene transfer. Horizontal gene transfer can introduce new genes, resulting in different ecological responses of the same species across samples, or the same ecological response of different species under the same favourable conditions [76]. As a result of horizontal gene transfer, it is believed that the prokaryotic genome is divided into two separate components: the core genome and the accessory genome [76]. The core genome contains genes that are typically crucial for survival and may serve as the genetic foundation for species as defined by Ernst Mayr [76]. The accessory genome carries genes that confer specialised ecological functions, of which some may be easily acquired or lost [76]. Even within the same species, categorised by their core genome, individual strains may exhibit a wide variance in their accessory genes, leading to a range of ecological functions. From this perspective, Cohan's concept of ‘ecotypes’ represents transient groups defined by specific combinations of accessory genes, suggesting that an ecological niche alone cannot account for the observed unity within species determined by the lineage of their core genes [76]. Additionally, the phylogenetic and physiological diversity of microfauna is substantially greater than that of macrofauna adding complexity in disentangling inter and intra domain interactions [76].

Besides the problems related to defining a species, the much smaller size of microorganism makes it challenging to define the boundaries of a community. The smaller size also gives rise to the ease by which microbes are passively dispersed by either meso- or macrofauna (**Figure 9**), or by means of e.g. wind currents, which is typically greater as compared to macrobiota [72]. As microbes interact strongly with each other in their local environment and may also interact closely with their immediate physical and chemical surroundings, the definition of the community needs to be defined by taking the scale and the abiotic environment into account. In essence, defining a microbial community requires careful consideration of spatial and temporal scales, interactions among organisms, and the interplay between biotic and abiotic factors. Each study in microbial community ecology must explicitly address these elements to clarify what is meant by community in the specific context.

Finally, for investigation of hypothesis related to the community structure it is important to have in mind that for a marker gene, such as the 16S rRNA gene, relative abundance does not have a 1-to-1 relationship to the number of individuals due to biases associated with sampling and DNA extraction protocols, as well as difference in 16S rRNA or genome copy numbers [100,101] (**Figure 10**). Furthermore, the microbiota data used for analysis is compositional, which is accompanied by its own caveats [102] (**Figure 10**).

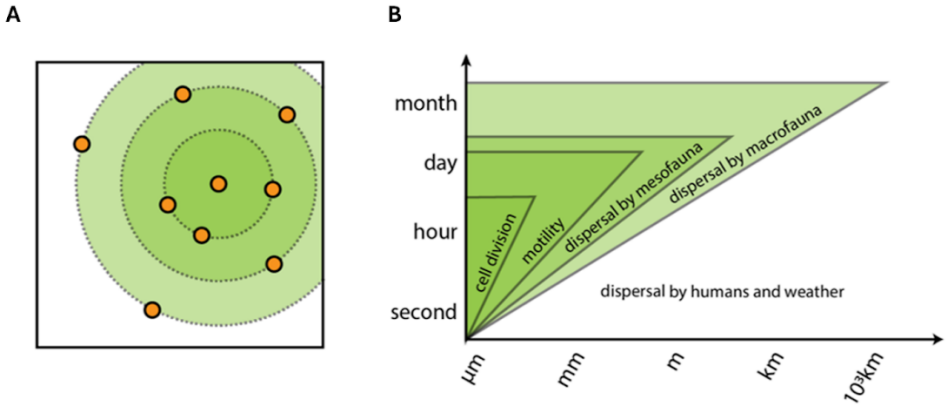

Figure 9: Microbial dispersal in soil habitats. A. Conceptual illustration of microbial dispersal. Colour saturation corresponds to mode of dispersal. **B.** Modes of microbial dispersal. Dispersal of bacteria within the soil matrix can occur through various mechanisms across different spatial and temporal scales. On the smallest scale, bacterial cells move due to Brownian motion and are 'pushed' by the process of cell division. At larger scales, bacteria can disperse passively via water flow or actively by swimming or swarming on moist surfaces. Some bacteria also hitch rides on fungal hyphae, which bridge air-filled gaps in the soil. Nonmotile bacteria depend on passive methods of dispersal, such as being transported by invertebrates, like earthworms, or larger animals. Inspired by [103].

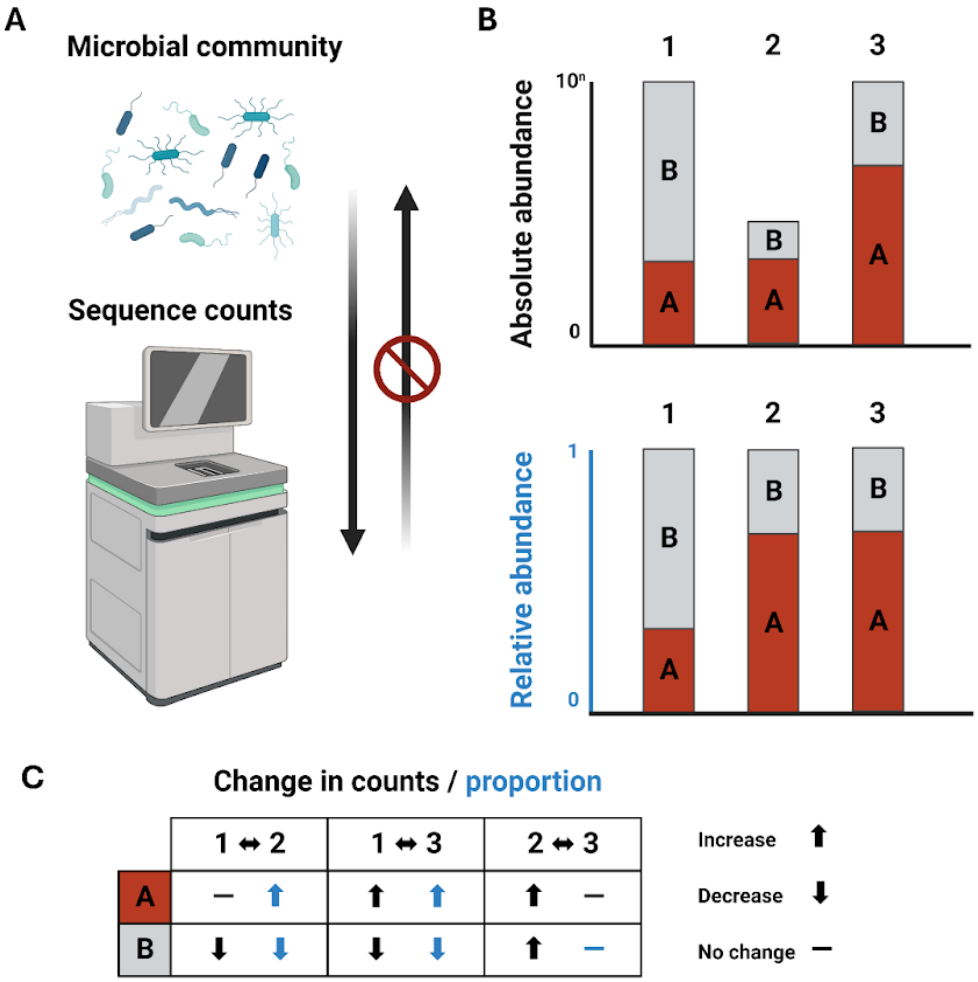

Figure 10: Compositional datasets. A. Sequencing of nucleic acids from a bacterial population only provides data on relative, not absolute, abundance of molecules. The counts in a high-throughput sequencing dataset represent the proportion of counts for each feature (like OTU or gene) in a sample, adjusted by the sequencing depth. Consequently, we can only ascertain the relative abundances from such data. **B.** The bar plots illustrate the discrepancy between the actual count and the proportion of molecules for two features, A (red) and B (grey), across three samples. The top bar graphs display the total counts for each sample, with the height of each colour indicating the count of each feature. Upon sequencing these samples, absolute count data is lost, leaving only relative abundances or "normalised counts", as depicted in the bottom bar graph. Notably, although features A and B show identical relative abundances in samples 2 and 3, the actual counts in the environment differ. This demonstrates how sequencing only provides relative, not absolute, molecular abundance data. **C.** The table displays actual and perceived changes for each sample when comparing one sample to another. Inspired by [102]. Created with BioRender.com

1.4 STANDING ON THE SHOULDERS OF GIANTS

1.4.1 MICROBIAL ECOLOGY ON THE GLOBAL SCALE

Large-scale projects utilising advanced technologies has significantly enhanced our comprehension of the microbial world, uncovering the ubiquitous nature of microbes and their adaptation to various environments, from the ocean depths to high mountain peaks. These studies underscored the pivotal roles microorganisms play in biogeochemical processes, nutrient cycling, and climate dynamics. Overall, such extensive research into microbial ecology and biogeography has reshaped our knowledge about the essential functions within ecosystems of these tiny life forms, their widespread distribution, and their adaptability to changing conditions, which is vital for forecasting biodiversity shifts and developing effective conservation strategies.

Among the first large scale microbial studies was the Global Ocean Sampling Expedition (GOS) (2004-2008) [104,105] and the Tara Oceans expedition (2009-2013) [106]. Both studies have expanded our knowledge of microbial diversity and ecology in the oceans. The studies found immense diversity of marine viruses, bacteria and small eukaryotes. The GOS Expedition identified novel genes and metabolic pathways, thereby shedding light on the genetic basis of key microbial processes that influence nutrient cycling, carbon fixation, and the marine food web. It also provided evidence of widespread horizontal gene transfer among marine microorganisms, suggesting a dynamic exchange of genetic material that contributes to microbial adaptation and evolution in the ocean [104]. The Tara Oceans Expedition provided insight into the global distribution patterns of planktonic organisms and their correlation with environmental factors, advancing our understanding of how marine ecosystems function [107]. It also provided evidence of how climate change affects plankton diversity and distribution, with implications for the broader marine food web and global biogeochemical cycles [108].

Overlapping with the timelines of the global marine studies and an increased interest in the human microbiome led to the conduction of the Human Microbiome Project (HMP) (2007-2016) [109] and Metagenomics of the Human Intestinal Tract (MetaHIT) (2010-2014) [110,111]. Both projects have provided insights into the complex interactions between humans and their microbiomes, highlighting the importance of microbes in association with health and disease. The projects revealed that the human body hosts a vast array of microbes, with bacterial cells outnumbering human cells by

a large margin. Furthermore, the human gut microbiome contains millions of unique microbial genes, vastly outnumbering the human genome in terms of gene content [59]. Both projects have identified associations between the microbiome and several medical conditions, which have opened new avenues for developing microbiome-based therapeutic interventions [112]. MetaHIT researchers proposed the concept of 'enterotypes' - distinct groupings of gut microbiota that may be associated with diet, geography, and health status, suggesting that the human population can be stratified based on gut microbiome composition [113]. Together, the HMP and MetaHIT have greatly enriched our comprehension related to the complexity of microbiomes, its critical role in health and disease, and its potential for developing novel therapeutic strategies.

The same researchers that conducted the HMP proposed to evaluate the microbiome of Earth [45]. The Earth Microbiome Project (EMP) (2010-2018) has revealed an incredible diversity and catalogued many previously unknown microorganisms. The EMP data have shown that microbial communities exhibit specific patterns of distribution and diversity that correlate with environmental factors such as climate, geography, and habitat type. By examining the genetic information of microbial communities, the EMP500 has provided insights into the metabolic pathways and ecological functions performed by microbes in different environments [114]. In summary, the Earth Microbiome Project represents a significant leap forward in our understanding of the microbial world. It emphasises the complexity and importance of microbial life on Earth and its indispensable contributions to ecosystem functions and planetary health.

More initiatives exist such as the Global Soil Biodiversity Initiative (GSBI), which focuses on evaluating soil biodiversity and its contribution to ecosystem functionality worldwide [115]. GSBI highlights the critical role of soil microorganisms in maintaining soil health, enhancing crop yields, and providing ecosystem services. Through the mapping of soil biodiversity, this initiative seeks to guide sustainable land management practices and conservation efforts. The International Census of Marine Microbes (ICoMM) was a project under the Census of Marine Life program [116]. ICoMM sought to catalogue the diversity, distribution, and abundance of microbial life in the oceans. By sampling and analysing microbial communities from various marine environments, ICoMM has expanded our understanding of marine microbial diversity and its role in ocean ecosystems.

Besides studies that collect and categorise new samples, initiatives exist which seek to collect the already generated meta-omic data in a uniform framework. One such initiative is the creation of a genomic reference

catalogue of the microbiomes of Earth from >10.000 metagenomic samples in the IMG/M database and aimed at capturing the extant microbial, metabolic and functional potential. [117]. The samples represent microbiomes of diverse habitats covering all the continents and oceans of Earth, including microbiomes from human and animal hosts, engineered environments, and natural and agricultural soils. The catalogue expanded the known phylogenetic diversity of Bacteria and Archaea by 44% and is broadly available for streamlined comparative analyses, interactive exploration, metabolic modelling and bulk download. The authors demonstrated the utility of this collection for understanding secondary-metabolite biosynthetic potential and for resolving thousands of new host linkages to uncultivated viruses [117]. Another such initiative is the Searchable, Planetary-scale mIcrobiome REsource (SPIRE) [118]. SPIRE consolidates different metagenome-derived microbial data types in a unified manner, spanning across habitats, geographical locations, and phylogenetic categories. The initiative encompasses ~100.000 metagenomic samples from 739 studies covering a wide array of microbial environments in combination with manually-curated contextual metadata. Across a total metagenomic assembly of 16 Tbp, SPIRE comprises 35 billion predicted protein sequences and 1.16 million newly constructed metagenome-assembled genomes (MAGs) of medium or high quality [118].

1.4.2 FINDINGS IN SOIL MICROBIAL ECOLOGY

In 2018 Delgado-Baquerizo and colleagues showed how only 2% of bacterial taxa account for nearly half of the soil bacterial communities across the globe and mapped the global distribution of groupings of ecological clusters of these dominant taxa [119]. In 2021 the same first author evaluated the structure and function of urban greenspaces and neighbouring natural ecosystems across the globe [120]. The study revealed that urban soils are key reservoirs of diversity for soil bacteria, protists, and functional genes, yet they give rise to highly uniform microbial communities globally. Urban greenspaces contain more fast-growing bacteria, algae, amoebae, and fungal pathogens, but fewer ectomycorrhizal fungi compared to natural ecosystems. These urban areas also have higher levels of genes linked to human pathogens, greenhouse gas production, rapid nutrient cycling, and abiotic stress than their natural counterparts. The composition of urban soil microbial communities is significantly influenced by city wealth, management practices, and local climate conditions [120]. Bahram and colleagues showed that bacterial genetic diversity in topsoil was highest in temperate regions, unlike fungal diversity, and that microbial gene composition was more closely related to environmental factors than to geographic distance [121]. It showed the existence of distinct global niches for fungi and bacteria, with diverse responses to precipitation and soil pH. Evidence of bacterial-fungal competition was found through analysis of antibiotic-resistance genes in both topsoils and oceans, underscoring the role of biotic interactions in shaping microbial communities. The results suggested that competition and environmental filtering determined the abundance, structure, and gene functions of these communities, indicating variable contributions to global nutrient cycling across different locations [121]. A recent publication used 2.941 soil metagenomes from NCBI as well as 348 samples from China and 15 from Europe to construct 40.039 MAGs [122]. This soil catalogue includes 3.641 high-quality genomes and identifies 16.530 of 21.077 species-level genomes as novel. From the unknown species-level genomes bins, they identified 43.169 biosynthetic gene clusters as well as 8.545 CRISPR-Cas genes, highlighting the catalogue as a genomic resource [122].

The soil microbiome has also seen publications on the continental scale. In Europe samples from the Land Use/Cover Area frame Survey (LUCAS) evaluated microbial biodiversity metrics and the distribution of potential functional groups across a gradient of land-use intensity [123]. The researchers discovered that the lowest levels of bacterial and fungal diversity were found in less-disturbed environments. In environments with

high disturbance, there was a significant increase in bacterial chemoheterotrophs, along with a higher presence of fungal plant pathogens and saprotrophs (feeders of decaying organic matter), and a decrease in beneficial fungal plant symbionts. The study suggested guidelines for environmental policy and advocated for the simultaneous consideration of both taxonomical and functional diversity in monitoring efforts [123]. For North America a study on the US-based National Ecological Observation Network (NEON) dataset has been conducted [124]. The study analysed average genome size, GC content, codon usage, and amino acid content from soil metagenomes spanning the entire US territory except Hawaii. The researchers identified that the most influential environmental factor affecting the distribution of these genomic traits was soil pH and contributed this relationship to the relationship of pH and various environmental factors, especially soil carbon content. Microbial communities in soils with high pH and lower carbon-to-nitrogen (C:N) ratios exhibited smaller genomes and higher GC content, suggesting a selective pressure against AT base pairs due to their higher C:N ratio as compared to GC base pairs. The study also observed that microbial communities in soils with lower C:N ratios favoured genes coding for amino acids with lower C:N, indicating nutrient conservation in amino acid stoichiometry [124]. Studies from Africa are notably absent, but a recent study tackled the issue of global sampling bias by the conduction of a comprehensive biogeographical survey of top-soil microbiomes from sub-Saharan Africa [52]. The findings indicated that the nine sub-Saharan countries involved in the study possessed distinct soil microbiomes. By incorporating soil chemistry and climate data from the sampled sites, the study showed that the top-soil microbiome was influenced by a variety of environmental factors such as pH, precipitation, and temperature. Using structural equation modelling, a predictive model was developed to assess how soil microbial biodiversity in sub-Saharan Africa might respond to future climate change scenarios, identifying the sub-Saharan African countries that are at risk of significant losses in soil microbial ecology and productivity due to climate change [52]. Finally, the eighth and forgotten continent of Earth, Zelandia [125,126], is represented by a study of more than 3.000 samples from 606 sites in New Zealand [127]. The study showed how 85% accuracy of land use and 83% accuracy for prediction of sites grouped by physico-chemical properties could be obtained from community data obtained from 16S rRNA gene sequencing [127]. The authors further demonstrated that soil bacterial communities offer valuable perspectives on the effects of land use on soil ecosystems and that they can serve as an effective method for assessing soil quality [127].

CHAPTER 2. OBJECTIVES OF THE PHD STUDY

Inspired by the same thoughts that led to the cataloguing of the entire danish flora over a period of 122 years, the drastic proposal of the Microflora Danica project was to catalogue the environmental microbiome of Denmark. The aim, similar to Flora Danica, is the generation of a near-complete reference database of the microflora of Denmark, in the hopes that the microflora of Denmark can be similarly studied, and its riches contribute to the extension of science. The reasoning of the project was, that with the accelerated development related to the sequencing of DNA seen over the last two decades, the time was right to apply novel DNA-based sequencing technologies at country scale. The specific objectives of the PhD study included:

- Setting up a high-throughput sample to sequence workflow for performing metagenomic library preparations of 10,000 environmental samples combined with subsequent downstream bioinformatic processing.

- Merging of data into a national microbial reference database to serve as the backbone for analysis of diversity and composition of microbial communities across an entire country.

- Apply prevalent and novel ecological theory derived from the ecological analysis of macrobiotic communities to increase the understanding of environmental microbial community dynamics.

The main output of this dissertation is a collection of scientific papers. The work presented here, in its essence, attempts to address some of the same basic questions that have captivated scientists since the origin of microbiology: Who are they, where are they, what can they do, and what are they doing?

CHAPTER 3. PREAMBLE FOR BODY OF WORK

3.1 FROM ENVIRONMENTAL SAMPLE TO DNA SEQUENCE

In the upcoming chapter, a detailed context for the key topics of each supporting paper will be provided, thereby highlighting how they fit into broader research themes, enhance current scientific understanding, and impact advancements in the field of environmental microbiology.

3.1.1 DECIPHERING THE GENETIC CODE

The introduction of PCR and new DNA sequencing methods has allowed the detection of taxa from complex samples independently from their cultivation status. To decipher the genomic content of an organism one must implement a strategy for sequencing the base pairs, which the genome is composed of. The field of DNA sequencing has undergone quite some development since its introduction in the late 1970s [128], and different methods can be applied to characterise microbial communities; Amplicon sequencing (a targeted method) and whole genome shotgun sequencing (an untargeted method). Whether to use amplicon or whole genome shotgun (WGS) sequencing is highly context-dependent, with either method having its own advantages. The preparation of amplicon sequencing libraries are cheaper compared to the costs in preparing metagenomic libraries by at least a factor of 12 [**Paper 1**], and provides means of characterisation of low abundant organisms, at the expense of the taxonomic resolution that can be achieved, at a lower sequencing effort as compared to metagenomic sequencing [4,129]. In principle WGS sequencing has the potential to sequence all DNA or RNA in a sample, thereby allowing complete community profiling, which can be linked to functional potential, but requires a much larger sequencing effort [130]. The Illumina and BGI platforms are predominant in whole genome shotgun sequencing due to very high outputs and accuracy, but at the cost of sequencing length. Long-read sequencing platforms such as the ones provided by Oxford Nanopore and Pacific Biosciences (Pacbio) are now more widely used due to recent increases in yield and accuracy [131,132].

3.1.2 CAVEATS

Before beginning the analysis of a microbial community analysis, it is important to consider the limitations and biases inherent in the methodologies used for generation of the sequencing data (**Figure 11**). Contamination can occur at any sample processing step in the forms of microbial contaminants in laboratory or kit reagents [133], cross-contamination between samples during library preparation, cross-over from previous sequencing runs [134] or sample bleeding occurring on the sequencing lane [135]. Low-biomass samples are especially susceptible to all these types of contamination [136]. To quantify possible contamination, it is recommended to include reaction blanks without any added template DNA as controls [133]. To trace any source of contamination such reaction blanks can be added at each processing step (**Figure 11**). Performing technical replicates can aid in assessing variability and separate technical from true variability [137], though this is often not feasible due to the overall cost of conducting sequencing experiments.

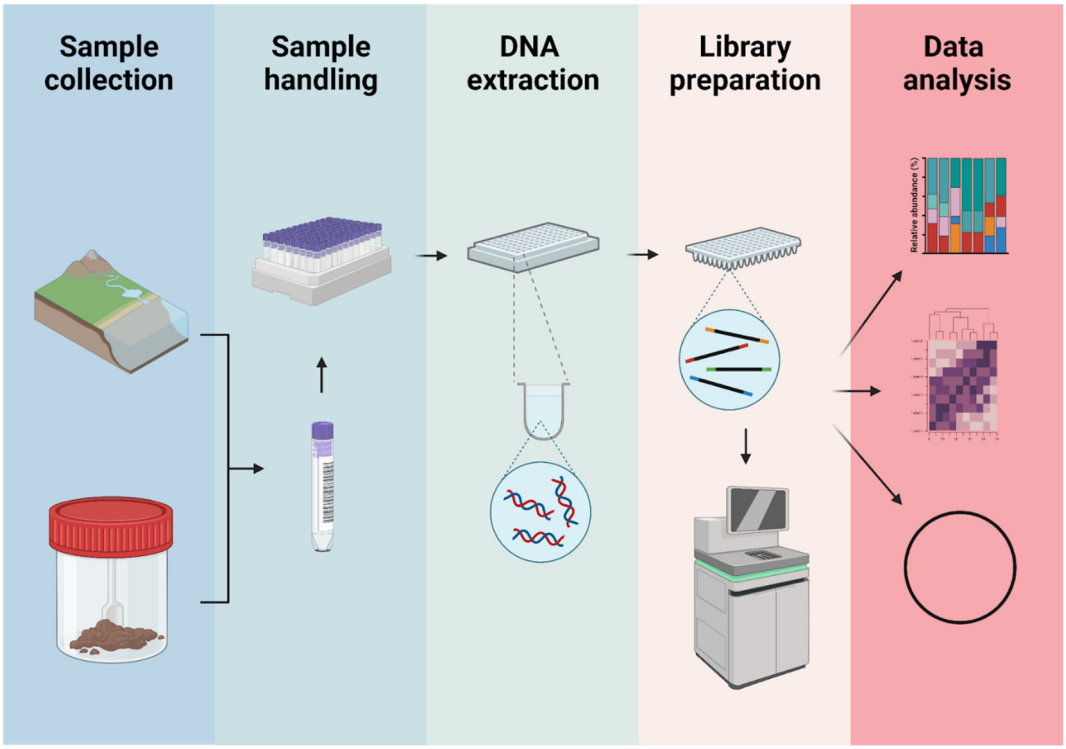

Figure 11: From sample to sequence flowchart. Schematic overview of the steps from acquiring an environmental sample to analysis of the generated sequencing data. **Sample collection.** Environmental material should be collected in a closed container and cooled if not directly processed to minimise effects from external factors. **Sample handling.** The collected environmental material is subsampled (100-500 mg) and should be representative of the environment it originates from. **DNA extraction.** Genetic material from intact cells as well as environmental DNA is isolated by means of cell lysis and purification. The genetic material should represent all community members. **Library preparation.** Before sequencing, the DNA needs to be prepared in accordance with the selected sequencing strategy and method. For sequencing of multiple samples in parallel, identifiers need to be added to the individual samples before multiplexing and subsequent sequencing. **Data analysis.** After demultiplexing the generated raw data needs to undergo appropriate processing to remove sequencing adapters and reads of insufficient quality. Depending on the analysis the data needs to be compared to a reference catalogue. Created with BioRender.com

3.1.3 SAMPLE HANDLING

The first step in any microbial community analysis is the acquisition of the samples of interest. Several factors must be considered to make sure that the sampling reflects a research question or hypothesis. Since many microbial systems are highly dynamic [138], one might consider how to address temporal and spatial effects, or apply stratification to make sure all factorial combinations of categorical variables describing the sampling area

are represented [**Paper 2**]. In other instances, factors such as sampling depth, surrounding vegetation, fertilisation and irrigation practices or partial oxygen pressure might be worth recording or retrieved from external sources [**Paper 2**, **Paper 4**]. However, it might be unfeasible to obtain certain metadata due to technical, time or financial limitations. The next step is the choice of sample collection and preservation protocols, since both may impose effects larger than the biological signal of interest in some cases [139]. The primary objective of collecting sample material is to ensure sufficient biological biomass for downstream processing while minimising the introduction of contamination [137]. Enrichment methods might be applied to samples from environments with scarcity of community members, as protocols for construction of optimal sequencing libraries require anything from nanograms to micrograms of DNA [138]. However, the physical separation of the starting material might come with introduction of bias [140]. Enrichment by means of isolation of cells from the samples, has been done to either increase DNA yields or avoid the coextraction of enzymatic inhibitors (such as humic acids), which can affect the downstream processing [137]. The latter is particularly relevant for soil samples [**Paper 1**]. Finally, factors such as timespan from collection to freezing as well as the number of freeze-thaw cycles a sample is subjected to can influence the observed microbial community [141].

3.1.4 ISOLATION AND PURIFICATION OF DNA

The key objective of the DNA extracted from the biomass is to be representative of all present community members and be of sufficient quantity and quality for preparation of sequencing libraries [138][**Paper 1**]. Many commercial kits specific for the extraction of DNA from environmental samples can be acquired, each differing in their procedure by either utilising chemical or mechanical disruption of the cells. The more vigorous mechanical lysis techniques can result in reduced DNA integrity, which can lead to DNA loss in library preparations involving size-selection [137]. In regard to mechanical lysis, the lysing matrix also exists in a variety of compositions, with the most common being silicate or ceramic spheres. The inorganic material in soil (often silicate minerals), can act as a lysing matrix itself and thereby affect the integrity of the obtained DNA [**Paper 1**]. Increased DNA integrity is observed from DNA extraction performed after separation of cells from the soil matrix, though with differences to the observed community [142]. Bead beating has systematically been shown to affect the observed microbial community when performing 16S rRNA gene amplicon sequencing [100][**Paper 1**]. One finding was that the amount of applied mechanical force had a greater impact on the observed community

(relative abundance of gram-positive bacteria) than had the type of lysing matrix and chemistry used. Extensive bead beating increased the total DNA from hard to lyse microbes, but also decreased overall DNA integrity [100][**Paper 1**].

The presence of a variety of contaminants including proteins, polyphenols and humic substances may inhibit PCR by binding to the Taq polymerase, the template DNA or both [143–145]. Extraction and purification methods differ in the lysis, inhibitor removal and DNA binding chemistry (**Figure 12**). A detergent such as SDS or CTAB is used with the purpose of chemical lysis, and addition of lysozyme, proteinase K and RNAse is commonly used. Inhibitor removal strategies varies among kits and include addition of polyvinylpyrrolidone (PVP), polyvinylpolypyrrolidone (PVPP), cetyl trimethyl ammonium (CTAB), activated charcoal or flocculating agents such as CaCl2, MgCl2, FeCl3 [146,147]. According to Braid and colleagues CTAB and PVP have proven unreliable regarding inhibitor removal [146]. Furthermore, carryover of contaminants from the kit itself (SDS, CTAB, EDTA, phenol, urea, guanidine) may also inhibit the PCR reaction [145]. Other factors affecting DNA extraction from environmental samples are pH and clay content [148]. Due to their opposite charge, the DNA molecules adsorb strongly to clay particles and thereby has the potential to substantially lower the yield from these samples [149]. Similarly, samples with a low content of biomass tend to yield low DNA quantities [**Paper 1**]. The mere number of publications on the topic of a cost-effective method to purify DNA from soil indicate the lack of a perfect protocol. Many of these protocols cannot be used in a high throughput setting due to inclusion of time-consuming steps, addition of hazardous substances such as phenol and chloroform, or steps preventing using the SBS format. The substantial research performed on soil emphasises how important it is to verify the extraction protocol used through benchmarking of multiple methods, and to use the observed microbial community as a target for optimisation [137][**Paper 1**].

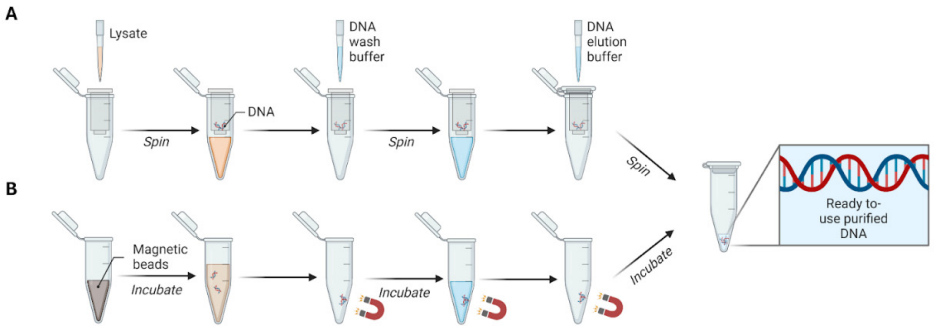

Figure 12: DNA isolation and purification. DNA isolation and purification can generally be achieved by either of two approaches. **A.** Adsorption of DNA to silica membranes happens via the hydrophobic effect under the presence of chaotropic salts. **B.** Alternatively DNA can be bound to beads coated with carboxyl groups, which can then be immobilised in a magnetic field due to the bead core consisting of polystyrene and magnetite. Impurities are removed by means of stepwise washing and finally the isolated and purified DNA is eluted in an appropriate volume. Created with BioRender.com

3.1.5 LIBRARY PREPARATION TECHNIQUES

With the reduction in sequencing costs due to increasingly high outputs from a single run on the Next Generation Sequencing (NGS) platforms and the option to multiplex up to 96 or 384 samples [137], library preparation has become a significant proportion of the total metagenomic project cost [**Paper 1**]. In an attempt to reduce the costs associated with the library preparation, protocols performing dilution of the expensive reagents, or replacing them with cheaper alternatives has been proposed [150,151]. Another possibility is to miniaturise the reaction volumes of the individual protocol steps [152–154][**Paper 1, Paper 2, Paper 4**]. Though specialised micro-dispensing platforms come with a steep entry price, break-even is reached with as little as 2000 metagenomic library preparations [**Paper 1**]. Miniaturising of the Illumina DNA prep library protocol by a factor of 10 reduces the total chemical and plastic costs at least 8-fold [151][**Paper 1**].

If insufficient DNA material was extracted for construction of optimal metagenomic sequencing libraries, the template material can be enriched by means of multiple displacement amplification (MDA). The method can be used to increase DNA yields but might introduce contamination from reagents, formation of chimeric sequences and amplification bias, with the impact depending on the amount of starting material and the number of amplification rounds performed [138,140].

3.1.6 AMPLICON SEQUENCING

Amplicon sequencing, also known as metabarcoding, takes advantage of PCR to amplify a specific region of the genome of interest prior to sequencing [28] (**Figure 13**). However, the inclusion of the PCR step will include some bias in the results as it is reliant on small nucleotide priming sequences. Primers miss certain taxonomic groups, thereby failing at capturing the full microbial community. Furthermore, as standard protocols are resolved around targeting only a small region of the total target gene - primarily determined by the read length of the sequencing platform - the selected region can have a large impact on the resulting observed microbial community [44,100,155,156]. This can be circumvented by using full-length or near full-length sequences, which can be sequenced on the long read platforms [54,156–159][**Paper 2, Paper 4**], or as synthetic long reads, if the sequencing platform prohibits the direct sequencing [53]. Another source of bias is the PCR itself due to differences in amplification efficiency as a result of the nature of the DNA template itself (known as GC bias)[137], or due to the presence of inhibitory substances as a result of poor DNA purification [**Paper 1**].

The marker gene sequences obtained from an amplicon sequencing experiment are traditionally subjected to a clustering algorithm and a predefined similarity threshold, thereby generating operational taxonomic units, which are less prone to sequencing errors [160]. However, the clustering process will be at the expense of taxonomic resolution, as single nucleotide variants distinguishing different species will be lost [161]. Instead, the exact denoised sequences can be used, which is referred to as amplicon sequencing variants (ASVs) [161][**Paper 1, Paper 2, Paper 3, Paper 4**].

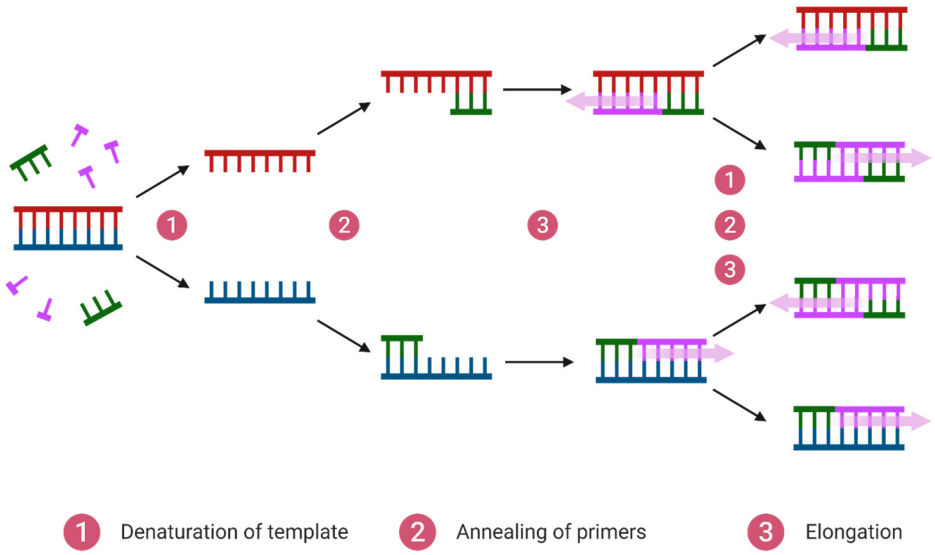

Figure 13: Polymerase Chain Reaction. The PCR reaction involves amplification of a targeted section of the genome using short synthetic DNA fragments known as primers, and then amplifying that segment using several rounds of DNA synthesis. **1.** Denaturation of the template genomic template DNA conferred by increasing the reaction temperature. **2.** Lowering of the reaction temperature allows for the binding of the synthetic DNA oligos to the denatured template DNA. **3.** A DNA polymerase synthesis the remainder of the complement to the template strand by addition of nucleotides. Created with BioRender.com

3.1.7 WHOLE GENOME SHOTGUN SEQUENCING

Metagenomic sequencing can be performed without implementing PCR and thereby avoids the associated bias. However, some standard protocols do implement PCR based on the amount of input material, which allows the construction of metagenomic libraries from samples with very low biomass, or from cells which are hard to extract and purify template material from. Including some cycles of PCR amplification allows the construction of metagenomic libraries from as little as one ng of DNA [**Paper 1, Paper 2, Paper 4**] (**Figure 14**). For sequencing on NGS platforms the input DNA undergoes fragmentation, which can be done by means of restriction enzymes for specific motifs or by means of bead-linked transposomes [162]. If PCR is included, this step will also be prone to amplification bias if inhibitory substances are present.

A community profile is obtained by aligning the metagenomic reads to a catalogue of reference sequences or genomes [**Paper 1, Paper 2, Paper 3, Paper 4**]. Therefore, data obtained from metagenomic sequencing experiments also depends on good reference databases, populated with all representatives from the microbial community under study. Since this is not the case yet, except for host-associated samples [163], effort has been made recently to expand the databases by adding environmental metagenome assembled genomes [117,164–166]. The generation of MAGs involves multiple bioinformatic processing steps [**Paper 4**]. First the individual reads are trimmed, and quality filtered, to remove unambiguous parts of the sequences. The reads are then assembled into longer contigs, by means of computational algorithms. The contigs are then sorted into bins, reflecting the reconstructed genomes, based on different parameters obtained from compositional features [167]. This methodology is currently undergoing development for the inclusion of new features and improvements to the binning algorithms themselves [168–171].

The assembly approach has its own limitations in the form of assembly break due to the presence of long homopolymeric regions and/or high strain heterogeneity causing chimeric and fragmented genomes. Furthermore, obtaining MAGs from the lowest prevalent microbes may require unreasonable sequencing depth to obtain sufficient genomic coverage [74]. Better assembly results can be obtained from libraries constructed from PCR and fragmentation free approaches with subsequent sequencing on the long read platforms [**Paper 4**]. Furthermore, DNA modifications such as methylation, which can be obtained from raw Nanopore signals or Pacbio 5-base sequencing, can be used as features for improving the algorithms for contig binning or for the enrichment of specific taxa of interest [172–175].

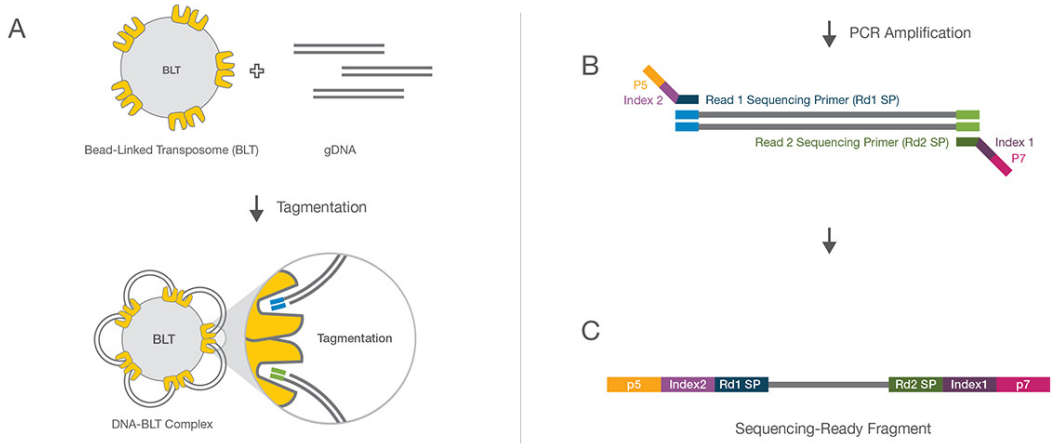

Figure 14: Metagenomic library preparation. A. Bead-linked transposomes facilitate the fragmentation of genomic DNA (gDNA) and the attachment of Illumina sequencing primers simultaneously. **B.** The reduced-cycle PCR amplification enhances sequencing-ready DNA fragments while incorporating indexes and adapters. **C.** Once amplified, these fragments are cleansed and combined into a pool of sequence-ready libraries. Adapted from [176].

3.2 BEYOND THE DNA SEQUENCE

3.2.1 TAXONOMIC DATABASES

Databases for 16S rRNA gene amplicon profiling have existed for a longer time than genome databases and are more populated due to the reference sequences being easier to obtain even for the low-prevalent community members. The population of the databases can be attributed to a number of large scale projects with the aim of describing the microbiomes of a single or multiple biotopes [156,159][**Section 1.4**][**Paper 4**]. For amplicon sequencing the most widely used databases include SILVA [177], Greengenes [43,178], and NCBI RefSeq [179] whereas for MAGs obtained by shotgun sequencing GTDB, custom, or proprietary databases are mostly used. Likewise, the classification of the metagenomic reads themselves relies on a catalogue of reference genomes for mapping [**Paper 4**]. The usage of a database for prokaryotic profiling has the goal to assign a seven rank taxonomy comprising the Kingdom, Phylum, Class, Order Family, Genus and Species level and sometimes even an annotation of the specific strain in question (applicable for whole genomes and not amplicons). Since the official nomenclature is relying on culturing and isolation from pure-culture type strains [180], there is debate on how to classify currently uncultured taxa, which comprise most of the microbial diversity[70]. A framework denoted SeqCode has been proposed to expand the

nomenclature with metagenome-assembled genomes and single-cell amplified genomes obtained from metagenomic sequencing data [181].

A major obstacle in the use of 16S reference databases to perform robust taxonomic assignments of all members of a complex microbial community is the lack of reference sequences with a high identity match (98.7% identity) [157]. Furthermore, the existing reference sequences are prone to have an incomplete or wrong taxonomy assigned [53,54]. The abovementioned widely used reference databases meet neither of these requirements, as most of the uncultured representatives are missing a reference sequence or a complete taxonomy [157]. Though, with the advances in high-throughput culture-independent methods a large number of near-perfect 16S rRNA reference sequences can be generated from either DNA or RNA, the latter circumventing primer bias [53,54,156,157,159][**Paper 2, Paper 4**]. The resulting reference sequences can then be assigned taxonomy based on sequence similarity thresholds, and addition of static de novo names at any taxonomic level when needed [157][**Paper 2, Paper 4**] (**Figure 15**). Another foundational problem in microbial ecology is determining the taxonomy and relative abundance of microorganisms of metagenomic origin. However, recent advances allow for the community composition to be estimated using conserved regions within multiple universal marker genes, not including the 16S or 18S rRNA phylogenetic marker gene [163]. This approach allows for the identification of individual metagenomes with a potential to recover novel metagenome-assembled genomes from lineages of interest, and can incorporate the user-recovered genomes into its reference database to improve the resolution of the profiling [163]. By doing so, the quantification of the total diversity of Bacteria and Archaea in metagenomic data shows that microbial genome databases are still far from saturated [163][**Paper 3, Paper 4**].

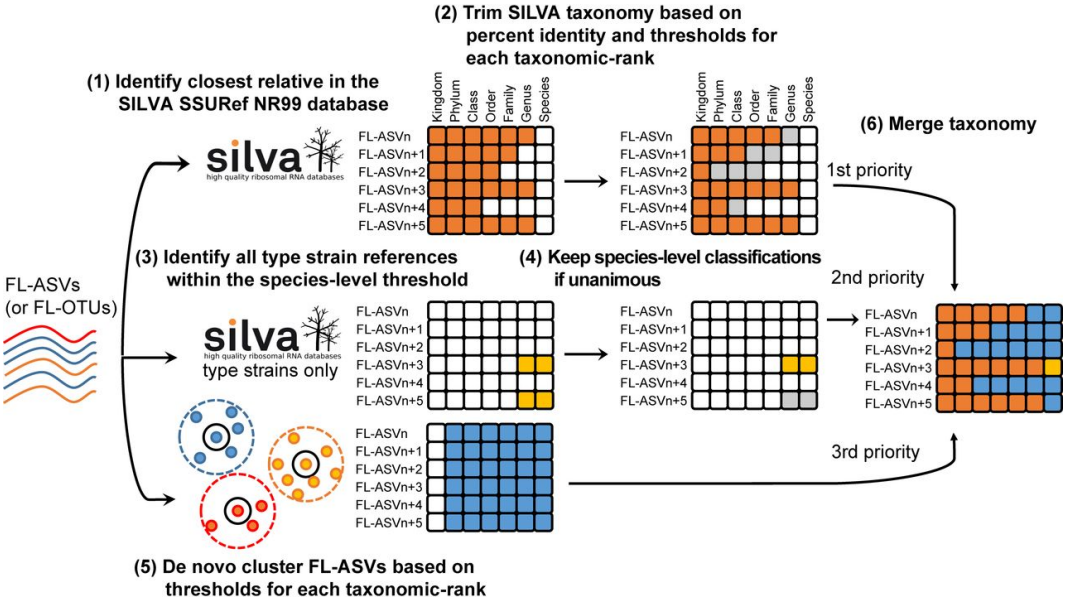

Figure 15: Overview of the AutoTax taxonomic framework. 1. FL-ASVs (or FL-OTUs) are initially mapped to the SILVA 138 SSURef NR99 database [177] to determine the closest relative and the corresponding percent sequence identity. **2.** Based on the sequence identity and the taxonomic thresholds suggested by Yarza et al. [182], taxonomy is assigned after appropriate sequence trimming. **3.** For species-level identification, FL-ASVs are further matched to type strain sequences from the SILVA database **4.** Species names are assigned if the sequence identity surpasses 98.7% and only a single species meets this threshold. **5.** FL-ASVs are grouped at varying percent identities, which helps in establishing a stable de novo taxonomy. **6.** A comprehensive taxonomy is formed by integrating the SILVA-based taxonomy with the de novo taxonomy to fill any gaps. The sources of taxonomic classifications for FL-ASVs are indicated by coloured squares: orange for SILVA SSURef NR99, yellow for SILVA type strains, blue for de novo names, and grey for names excluded during the AutoTax process. Adapted from [157].

3.2.2 LINKING TAXONOMY AND FUNCTIONAL POTENTIAL

Exploration of the functional potential of the microbial communities, allows for hypotheses related to their metabolic function to be made. Analysing these metabolic characteristics enriches our understanding of microbial interactions, ecological functions, and life history strategies in various settings, thereby broadening our comprehension of nutrient cycling, ecosystem behaviour, and microbial adaptability to environmental pressures. Generally, the functional potential of environmental metagenomes can be obtained in two ways. Firstly, microbial metagenomic reads offer a comprehensive snapshot of the genetic material present in environmental samples. The mapping of metagenomic reads against a reference catalogue of functional genes or genomes, can identify specific genes involved in the metabolic pathways of interest [137][**Paper 4**] (**Figure**

16). This process involves aligning short DNA sequences obtained from metagenomic samples to known sequences within the reference catalogue, allowing for the detection of the genes from different microbial taxa [137]. Contrastingly, sequencing of the 16S rRNA phylogenetic marker gene does not directly yield information about the functional potential of the organism it represents. However, the mapping of microbial 16S ASVs to a catalogue of reference genomes allows the functional potential to be inferred from the genome of which the ASV is aligning to [183,184][**Paper 3**] (**Figure 16**). Thereby this method can offer some of the same insights as obtained from metagenomic whole genome shotgun sequencing. The mapping of 16S genes to reference genome catalogues is particularly valuable when metagenomic data from complex communities are scarce, as 16S amplicon sequencing is still the prevalent approach for identifying bacterial communities.

Neither method is without caveats, as any functional potential is limited to the understanding of the metabolic pathways and the overall coverage of the used reference catalogues. Specifically for amplicon-based functional prediction methods, these predictions tend to be skewed towards known reference genomes including at least one copy of the 16S gene, making the identification of rare, environment-specific functions less likely than for the metagenomic approach. Though the latest version of PICRUSt [184] has allowed for the incorporation of user-specified genomes, allowing exploration of specific environments, these genomes will have to be generated from metagenomic sequencing first anyways. Furthermore, amplicon-based methods lack the precision needed to identify strain-specific functionalities because these methods can only differentiate between taxa based on variations in the sequences of the targeted marker gene [184].

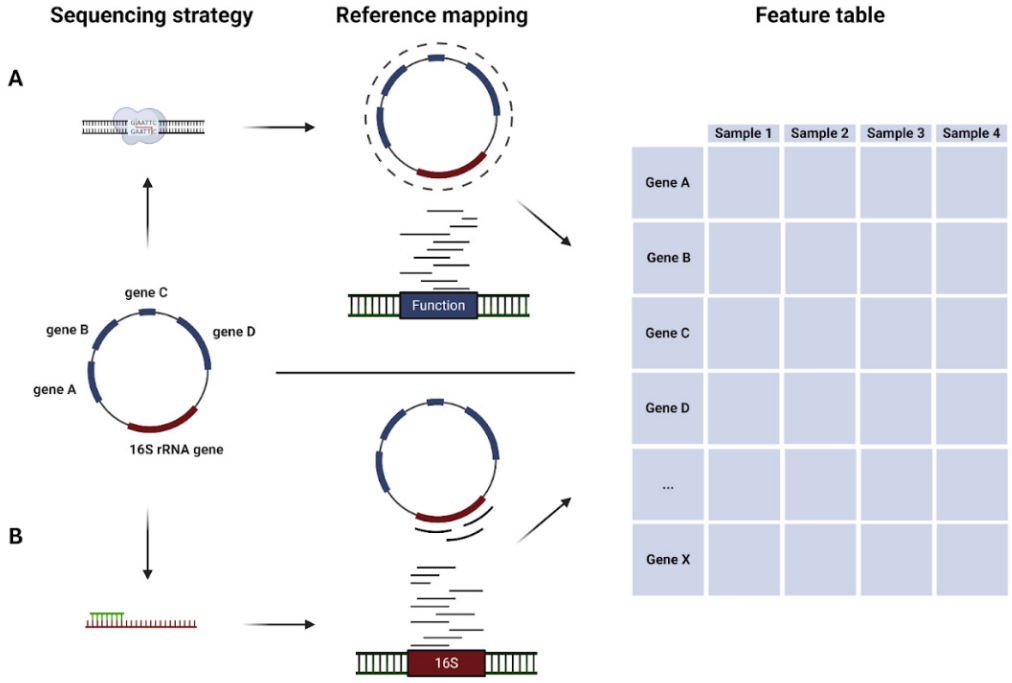

Figure 16: DNA sequence reference mapping. Information about the functional potential can be obtained by two approaches. **A.** Metagenomic reads originate from every part of the genome under investigation, and hence also from functional genes. **B.** Amplicon reads originating from the 16S rRNA gene, can be mapped to a catalogue of reference genomes which include at least one copy of the 16S rRNA gene. The functional potential can then be inferred from the reference genome to which the 16S rRNA gene sequence is mapping.

3.2.3 ANALYSIS OF MICROBIOME DATA

The analysis of microbiome data, that being of either taxonomic or functional origin, aims at dissecting the complexity of microbial communities, thereby illuminating the diversity, composition, and adaptability of microorganisms to environmental conditions (**Figure 17**). This multi-faceted approach provides a detailed examination of the richness and evenness of species within a community, revealing not just the variety of life forms present (diversity) but also identifying the specific species that make up these ecosystems (composition). By delving into how these communities respond to ecological gradients - variations in factors like pH, moisture and disturbance - mechanisms by which environmental pressures sculpt microbial landscapes, or vice versa, can be uncovered [**Paper 2, Paper 4**]. Analysis on microbial diversity within ecosystems can be categorised into alpha, beta, and gamma diversity, each representing and providing different aspects of biodiversity and insights into microbial community structure, distribution, and ecological function. Together, alpha, beta, and gamma diversity provide a comprehensive framework for studying microbial communities, from the intricate species relationships within specific habitats to the broad patterns of diversity across larger geographical scales.

Alpha diversity

Alpha diversity refers to the diversity within a specific area or ecosystem and is a measure of the variety and abundance of species in a single microbial community [185] (**Figure 17**). It encompasses the number of observed taxa or inferred richness from observation, and the abundance (evenness) of species.

Beta diversity

Beta diversity measures the change or turnover in species composition between different areas or ecosystems, providing a comparison of microbial community composition across different environments or along environmental gradients [185] (**Figure 17**). Beta diversity is essential for assessing the effects of environmental gradients and disturbances on community composition [**Paper 1, Paper 2, Paper 4**]. Beta-diversity is quantified using various distance metrics, each highlighting different aspects of community dissimilarity [188]. These metrics essentially calculate the ecological distance between pairs of communities. To calculate these distances, one typically constructs a matrix of presence-absence or relative abundances for the community members under comparison. This is often done after application of a data transformation procedure [188]. Commonly used metrics are the Bray-Curtis dissimilarity, the Jaccard index, the

Sørensen-Dice Coefficient (also known as binary Bray-Curtis dissimilarity) and the UniFrac Distance [189,190]. The chosen metric is then applied to each pair of communities in the dataset, producing a distance matrix that represents the beta diversity among all pairs of communities. This matrix can then be used as input in downstream analysis, such as ordination methods (e.g. PCoA, NMDS and RDA) or hierarchical clustering to visualise and evaluate the ecological distances between communities (**Figure 17**). The dissimilarity matrix can also be used to evaluate community distance decay, which is the ecological principle that geographical separation leads to increased dissimilarity among communities [72]. Similarly, the geographical location can be used to map out geographical patterns in microbial distribution and allows for the partitioning of spatial and environmental drivers of microbial community divergence [**Paper 2, Paper 4**].

Gamma diversity

Gamma diversity is the overall diversity for different ecosystems within a larger region and represents the total richness across all communities within that region. It essentially combines the alpha diversity of individual communities and the beta diversity among those communities. Gamma diversity gives a macroscopic view of the microbial diversity across a landscape, geographical area, or habitat type, highlighting the broad-scale patterns of microbial distribution and the cumulative effects of local and regional processes on microbial diversity. Like the other diversity metrics, gamma diversity has its roots in ecological theory of macrobiota, where the current best practice is using Hill numbers of order q (also known as the effective number of species) [186,187] [**Paper 4**]. The parameter q, in biodiversity measurement, determines the sensitivity of diversity indices to species abundance. With q=0, species abundances are ignored, and diversity (^{0}D) simply equals the total number of species. When q=1, species are weighted by their relative abundances, making ^{1}D a measure of the effective number of common or "typical" species in the community. At q=2, the focus shifts towards abundant species while rare species are less considered, and ^{2}D represents the effective number of “dominant” or highly abundant species. ^{1}D and ^{2}D equals the exponential of Shannon entropy and the inverse of the Simpson concentration, respectively. Hence, these are known as Shannon diversity and Simpson diversity, not to be confused with the Shannon and Simpson diversity indexes. The calculations can be incidence-based rather than abundance-based, thereby using frequencies to determine if a taxon is considered abundant or not [186,187] [**Paper 4**].

Beyond diversity metrics

The abundance of each identified taxa across the investigated samples can itself be thought of as variables (also known as features) and used in univariate or multivariate analysis such as correlations with abiotic and biotic variables (**Figure 17**). The numerical value of the relative abundance of such microbial features is not necessarily directly associated with the feature importance, as both the rare and abundant taxa are important for the overall function of the ecosystem. The abundant taxa drive variation in respiration, metabolic potential, cell yield, and the rare taxa are influencing the capacity of the communities to degrade specific substrates [191].

The core community members, which are taxa consistently present across a type of sample, can serve as a fingerprint of the type of sample under investigation, and serve as indicators of the underlying biological processes that sustain various environmental niches [**Paper 2, Paper 4**]. The core taxa in complex communities are typically identified from the frequency by which a specific taxon is observed in samples from a particular habitat [192]. Additionally, an abundance threshold may be set to select taxa that are likely to have a significant quantitative effect on ecosystem functions [193]. As complex microbial communities harbour a vast low-abundant diversity, this method can also serve to reduce the number of microbial features for subsequent modelling [**Paper 2**]. As a continuation of this the core community members provide means for prediction of location on ecological gradients and classification of the specific habitat a given sample was taken from [**Paper 2, Paper 4**] (**Figure 17**). Finally, the DNA sequences themselves, being derived from either a taxonomic or a functional gene, can serve as phylogenetic markers allowing for investigation of the trait-specific evolutionary development of the investigated taxa [**Paper 2, Paper 4**] (**Figure 17**).

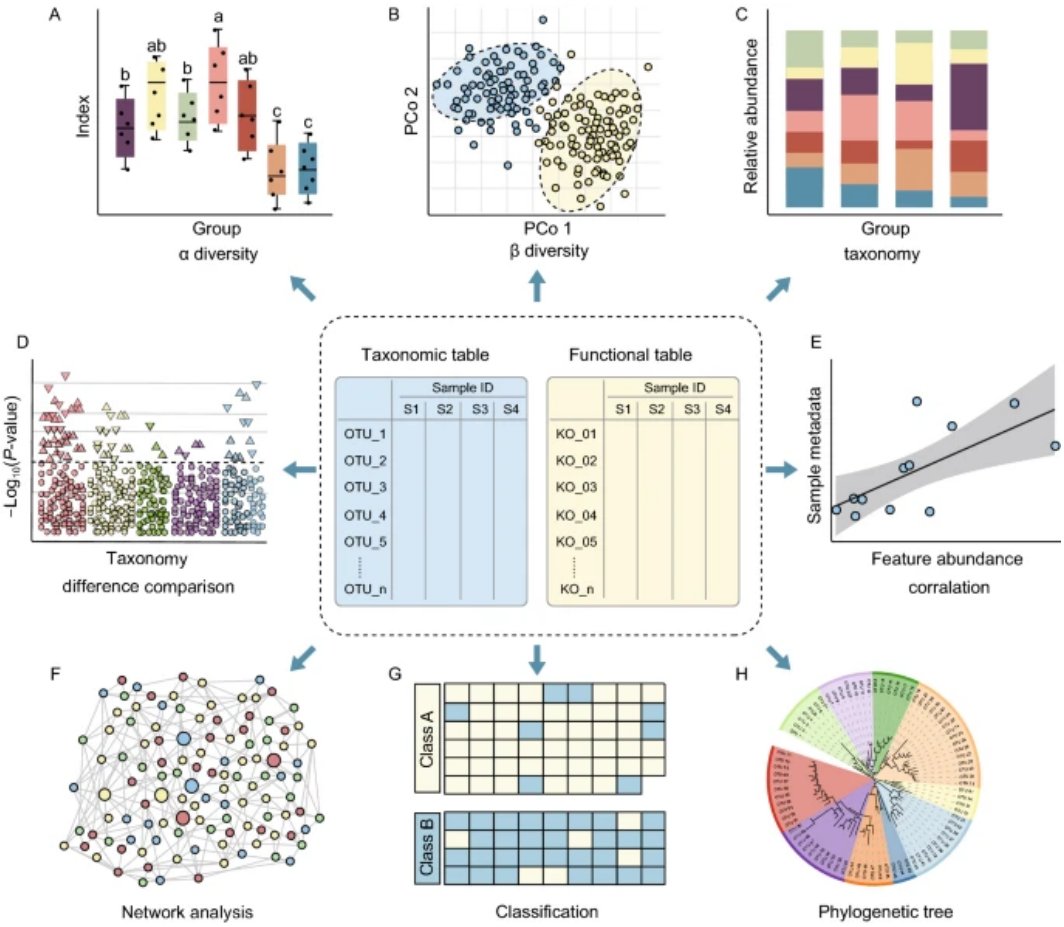

Figure 17: Analysis of microbiome derived feature tables. Overview of visualisation and statistics associated with tables of taxonomic or functional features derived from microbiome sequencing data. The downstream analysis of microbiome feature tables includes **A.** α-diversity (within-sample diversity), **B.** β-diversity (between-sample distance), **C.** taxonomic composition (relative abundance of features), **D.** difference comparison (significantly difference in features between groups), **E.** correlation analysis (correlation of features and sample metadata), **F.** network analysis (global view correlation of features), **G.** classifications from machine learning (group classification or sample metadata regression analysis), **H.** phylogenetic tree (phylogeny tree or taxonomy hierarchy). Adapted from [185].

CHAPTER 4. CONCLUSIONS AND PERSPECTIVES

The common theme for the presented papers in this thesis is “large-scale”, that being used in relation to describe high-throughput processing or in relation to sampling effort. Together, these papers provide a comprehensive view of microbial ecology and practical methodologies for future research. By integrating these studies into the wider scientific landscape, their role in filling research gaps and shaping future inquiries will be demonstrated, thereby emphasising their collective importance in expanding our knowledge of ecological and biological systems.

Paper 1 was conducted to lay the foundation for standardised protocols to be used in the Microflora Danica project. From the planning phase it was evident that a solution that performed well on a range of different soils was needed. In a high-throughput environment, it is not feasible to determine the soil type of each individual sample and perform the DNA extraction accordingly. Any high-throughput solution should perform similar and acceptable on a range of different soils. Therefore, one of the aims of **Paper 1** was to benchmark high-throughput DNA extraction methods, which allows for automation procedures, based on length, quality, quantity and the observed microbial community profile. Furthermore, it was crucial for the conduction of the Microflora Danica project that the cost of library preparations was reduced, to enable the processing of 10.000 samples within the budget of the project. Therefore, **Paper 1** also showcases how a nanoliter drop-dispensing system allows for the miniaturisation of both amplicon and metagenomic library preparations by a factor of 5 and 10 respectively. As more large-scale studies will be needed for future progress within the field of microbial ecology, downscaling will allow for an increase in the number of included samples within a fixed budget.

Paper 2 dives into a subset of the soil samples included in the Microflora Danica project from different terrestrial ecosystems found in Denmark. The samples of this study were subjected to both full-length 16S rRNA gene amplicon and metagenomic sequencing. These data are linked to vast recordings of the above-ground macrobiota as well as physicochemical measurements serving as proxies for the major environmental gradients. The constructed 16S rRNA gene sequences were used to generate an ecosystem-specific reference database, which increased the obtainable resolution of the metagenomic derived community profiles. The aim of **Paper 2** was to evaluate the response of both the above-ground and the

below-ground communities to these major ecological gradients and evaluate to what extent the below-ground prokaryotic communities mirror those of the above-ground assemblages of plants, animals, and macrofungi. **Paper 2** also demonstrates how microbial indicators with great success can be used for prediction of the variables serving for proxies of the environmental gradients as well as in the classification of terrestrial habitats, which is otherwise reliant on a specialised expertise evaluation. This holds potential, as the microbes have previously been overlooked and will be important to consider in future management and conservation efforts.

Paper 3 was an outcome from the external stay at University of Colorado, Boulder. This study took advantage of the vast amount of available 16S rRNA gene amplicon data from more than 3.700 samples spanning a broad diversity of habitats, including Danish municipal wastewater treatment plants. The ASVs were linked to available high-quality genomes in GTDB with the aim of elucidating the taxonomic and environmental distribution of amino acid auxotrophies. **Paper 3** found that three thirds of bacterial taxa can synthesise all amino acids, and that auxotrophies are more prevalent in obligate intracellular parasites and in free-living taxa with genomic attributes characteristic of “streamlined” life history strategies. Furthermore, auxotrophic taxa were discovered to be more prevalent in host-associated environments and fermented food products, whereas they were relatively scarce in soil and aquatic systems. The findings of **Paper 3** enhance our understanding of amino acid auxotrophy across the bacterial tree of life and identifies the ecological settings where auxotrophy may be an advantageous survival strategy. This approach holds the potential to aid in the challenges of bringing yet-to-be-cultivated microbes into culture by examining the specific auxotrophies within generated metagenome assembled genomes of uncultured taxa.

Paper 4 serves as a conclusion to the Microflora Danica sampling campaign and highlights the main findings of the project. The study presents more than 10.000 metagenomic datasets, 415 full-length 16S rRNA gene amplicon datasets, as well as 449 bacterial and eukaryotic rRNA operon datasets. The study is of an unprecedented spatial resolution with one sample per 4 km^2 of the geographical area of Denmark. The samples cover the major habitat types, and each sample is provided with a comprehensive five-tier environmental ontology, which serves as a habitat classification system. **Paper 4** uses this detailed resolution to assess the habitat-specific microbial novelty relative to the existing databases, to investigate the diversity within and between samples across Denmark and evaluate the effect of land-management on microbial diversity and community homogenisation. **Paper 4** also shed light on the distribution and diversity of

nitrifiers, key functional groups within a country that is 60% agricultural land. In doing so, **Paper 4** provides support for niche differentiation of comammox and canonical nitrite oxidising bacteria, which underlines the general need for further investigation of this functional group. The Microflora Danica dataset lays the groundwork for addressing key questions in microbial ecology, such as the factors that influence microbial diversity, distribution, and functionality.

REFERENCE LIST

[1] Knudsen H. Fortællingen om Flora Danica (Preface). Lindhardt og Ringhof Forlag, 2019.

[2] Zhang W, Li F, Nie L. Integrating multiple “omics” analysis for microbial biology: application and methodologies. Microbiology 2010;156:287–301.

[3] Opal SM. A Brief History of Microbiology and Immunology. Vaccines: A Biography November 10. 2009:31.

[4] Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. Computational meta’omics for microbial community studies. Mol Syst Biol 2013;9:666.

[5] Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. Mass Spectrom Rev 2007;26:51–78.

[6] Svanberg U, Lorri W. Fermentation and nutrient availability. Food Control 1997;8:319–27.

[7] Chatterjee I, Somerville GA, Heilmann C, Sahl H-G, Maurer HH, Herrmann M. Very low ethanol concentrations affect the viability and growth recovery in post-stationary-phase Staphylococcus aureus populations. Appl Environ Microbiol 2006;72:2627–36.

[8] A sip of history: ancient Egyptian beer. The British Museum n.d. https://www.britishmuseum.org/blog/sip-history-ancient-egyptian-beer (accessed March 13, 2024).

[9] HEART VIEWS - VOLUME 4 NO.2 JUNE - AUGUST 2003 n.d. https://web.archive.org/web/20110503050312/http://www.hmc.org.qa/hmc/heartviews/H-V-v4%20N2/9.htm (accessed March 13, 2024).

[10] Suddenly I See: How Microscopes Made Microbiology Possible. ASM.org 2022. https://asm.org/articles/2022/june/suddenly-i-see-how-microscopes-made-microbiology-p (accessed March 13, 2024).

[11] Nutton V. The reception of Fracastoro’s Theory of contagion: the seed that fell among thorns? Osiris 1990;6:196–134.

[12] Drews G. The roots of microbiology and the influence of Ferdinand Cohn on microbiology of the 19th century. FEMS Microbiol Rev 2000;24:225–49.

[13] Lane N. The unseen world: reflections on Leeuwenhoek (1677) “Concerning little animals.” Philos Trans R Soc Lond B Biol Sci 2015;370:20140344.

[14] Crouvisier-Urion K, Chanut J, Lagorce A, Winckler P, Wang Z, Verboven P, et al. Four hundred years of cork imaging: New advances in the characterization of the cork structure. Sci Rep 2019;9:19682.

[15] Smith KA. Louis Pasteur, the father of immunology? Front Immunol 2012;3:68.

[16] Opal SM. A Brief History of Microbiology and Immunology. In: Artenstein AW, editor. Vaccines: A Biography, New York, NY: Springer New York; 2009, p. 31–56.

[17] Blevins SM, Bronze MS. Robert Koch and the “golden age” of bacteriology. Int J Infect Dis 2010;14:e744–51.

[18] Gaynes R. The Discovery of Penicillin—New Insights After More Than 75 Years of Clinical Use. Emerg Infect Dis 2017;23:849.

[19] Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 1953;171:737–8.
[20] Deciphering the Genetic Code, 1958-1966. Francis Crick - Profiles in Science n.d. https://profiles.nlm.nih.gov/spotlight/sc/feature/deciphering (accessed March 13, 2024).
[21] Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. Nucleotide sequence of bacteriophage phi X174 DNA. Nature 1977;265:687–95.
[22] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995;269:496–512.
[23] Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 1977;74:5088–90.
[24] Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 2017;541:353–8.
[25] Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 2015;521:173–9.
[26] Williams TA, Cox CJ, Foster PG, Szöllősi GJ, Embley TM. Phylogenomics provides robust support for a two-domains tree of life. Nat Ecol Evol 2020;4:138–47.
[27] Eme L, Tamarit D, Caceres EF, Stairs CW, De Anda V, Schön ME, et al. Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. Nature 2023;618:992–9.
[28] Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 1986;51 Pt 1:263–73.
[29] Mullis KB. The unusual origin of the polymerase chain reaction. Sci Am 1990;262:56–61, 64–5.
[30] Home n.d. https://www.worldfloraonline.org/ (accessed March 14, 2024).
[31] World Plants: How many species? n.d. https://www.worldplants.de/world-plants-complete-list/total-species-count (accessed March 14, 2024).
[32] Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. How many species are there on Earth and in the ocean? PLoS Biol 2011;9:e1001127.
[33] Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. Proc Natl Acad Sci U S A 2016;113:5970–5.
[34] Larsen BB, Miller EC, Rhodes MK, Wiens JJ. INORDINATE FONDNESS MULTIPLIED AND REDISTRIBUTED: THE NUMBER OF SPECIES ON EARTH AND THE NEW PIE OF LIFE. Q Rev Biol 2017;92:229–65.
[35] Louca S, Mazel F, Doebeli M, Parfrey LW. A census-based estimate of Earth's bacterial and archaeal diversity. PLoS Biol 2019;17:e3000106.
[36] Lennon JT, Locey KJ. More support for Earth's massive microbiome. Biol Direct 2020;15:5.
[37] Wiens JJ. Vast (but avoidable) underestimation of global biodiversity. PLoS Biol 2021;19:e3001192.

[38] Louca S, Mazel F, Doebeli M, Parfrey LW. Response to “Vast (but avoidable) underestimation of global biodiversity.” PLoS Biol 2021;19:e3001362.
[39] Fishman FJ, Lennon JT. Macroevolutionary constraints on global microbial diversity. Ecol Evol 2023;13:e10403.
[40] Ghaly TM, Tetu SG, Penesyan A, Qi Q, Rajabal V, Gillings MR. Discovery of integrons in Archaea: Platforms for cross-domain gene transfer. Sci Adv 2022;8:eabq6376.
[41] Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 1990;87:4576–9.
[42] Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol 2016;1:16048.
[43] McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 2012;6:610–8.
[44] Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 2019;10:5029.
[45] Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth’s multiscale microbial diversity. Nature 2017;551:457–63.
[46] Edgar RC. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics 2018;34:2371–5.
[47] Shoemaker WR, Locey KJ, Lennon JT. A macroecological theory of microbial biodiversity. Nat Ecol Evol 2017;1:107.
[48] Wiens JJ. How many species are there on Earth? Progress and problems. PLoS Biol 2023;21:e3002388.
[49] Cameron EK, Martins IS, Lavelle P, Mathieu J, Tedersoo L, Gottschall F, et al. Global gaps in soil biodiversity data. Nat Ecol Evol 2018;2:1042–3.
[50] Cameron EK, Martins IS, Lavelle P, Mathieu J, Tedersoo L, Bahram M, et al. Global mismatches in aboveground and belowground biodiversity. Conserv Biol 2019;33:1187–92.
[51] Guerra CA, Heintz-Buschart A, Sikorski J, Chatzinotas A, Guerrero-Ramírez N, Cesarz S, et al. Blind spots in global soil biodiversity and ecosystem function research. Nat Commun 2020;11:3870.
[52] Cowan DA, Lebre PH, Amon C, Becker RW, Boga HI, Boulangé A, et al. Biogeographical survey of soil microbiomes across sub-Saharan Africa: structure, drivers, and predicted climate-driven changes. Microbiome 2022;10:131.
[53] Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, Albertsen M. Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. Nat Biotechnol 2018;36:190–5.
[54] Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, et al. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. Nat Methods 2021;18:165–9.

[55] Anda M, Yamanouchi S, Cosentino S, Sakamoto M, Ohkuma M, Takashima M, et al. Bacteria can maintain rRNA operons solely on plasmids for hundreds of millions of years. Nat Commun 2023;14:7232.
[56] Lyons TW, Reinhard CT, Planavsky NJ. The rise of oxygen in Earth's early ocean and atmosphere. Nature 2014;506:307–15.
[57] Gilbert JA, Neufeld JD. Life in a World without Microbes. PLoS Biol 2014;12:e1002020.
[58] Blaser M, Bork P, Fraser C, Knight R, Wang J. The microbiome explored: recent insights and future challenges. Nat Rev Microbiol 2013;11:213–7.
[59] Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nat Med 2018;24:392–400.
[60] Wardle DA, Bardgett RD, Klironomos JN, Setälä H, van der Putten WH, Wall DH. Ecological linkages between aboveground and belowground biota. Science 2004;304:1629–33.
[61] De Deyn GB, Van der Putten WH. Linking aboveground and belowground diversity. Trends Ecol Evol 2005;20:625–33.
[62] Cardinale BJ, Duffy JE, Gonzalez A, Hooper DU, Perrings C, Venail P, et al. Biodiversity loss and its impact on humanity. Nature 2012;486:59–67.
[63] Bardgett RD, van der Putten WH. Belowground biodiversity and ecosystem functioning. Nature 2014;515:505–11.
[64] Wagg C, Bender SF, Widmer F, van der Heijden MGA. Soil biodiversity and soil community composition determine ecosystem multifunctionality. Proc Natl Acad Sci U S A 2014;111:5266–70.
[65] Delgado-Baquerizo M, Eldridge DJ, Ochoa V, Gozalo B, Singh BK, Maestre FT. Soil microbial communities drive the resistance of ecosystem multifunctionality to global change in drylands across the globe. Ecol Lett 2017;20:1295–305.
[66] Sayer EJ, Oliver AE, Fridley JD, Askew AP, Mills RTE, Grime JP. Links between soil microbial communities and plant traits in a species-rich grassland under long-term climate change. Ecol Evol 2017;7:855–62.
[67] Wagg C, Hautier Y, Pellkofer S, Banerjee S, Schmid B, van der Heijden MG. Diversity and asynchrony in soil microbial communities stabilizes ecosystem functioning. Elife 2021;10. https://doi.org/10.7554/eLife.62813.
[68] Wade W. Unculturable bacteria--the uncharacterized organisms that cause oral infections. J R Soc Med 2002;95:81–3.
[69] Stewart EJ. Growing unculturable bacteria. J Bacteriol 2012;194:4151–60.
[70] Steen AD, Crits-Christoph A, Carini P, DeAngelis KM, Fierer N, Lloyd KG, et al. High proportions of bacteria and archaea across most biomes remain uncultured. ISME J 2019;13:3126–30.
[71] Cavicchioli R, Ripple WJ, Timmis KN, Azam F, Bakken LR, Baylis M, et al. Scientists' warning to humanity: microorganisms and climate change. Nat Rev Microbiol 2019;17:569–86.
[72] Nemergut Diana R., Schmidt Steven K., Fukami Tadashi, O'Neill Sean P., Bilinski Teresa M., Stanish Lee F., et al. Patterns and Processes of Microbial Community Assembly. Microbiol Mol Biol Rev 2013;77:342–56.

[73] Fierer N. Embracing the unknown: disentangling the complexities of the soil microbiome. Nat Rev Microbiol 2017;15:579–90.
[74] Matthews TJ, Whittaker RJ. REVIEW: On the species abundance distribution in applied ecology and biodiversity management. J Appl Ecol 2015;52:443–54.
[75] Explaining General Patterns in Species Abundance and Distributions n.d. https://www.nature.com/scitable/knowledge/library/explaining-general-patterns-in-species-abundance-and-23162842/ (accessed April 9, 2024).
[76] Prosser JI, Bohannan BJM, Curtis TP, Ellis RJ, Firestone MK, Freckleton RP, et al. The role of ecological theory in microbial ecology. Nat Rev Microbiol 2007;5:384–92.
[77] Graham EB, Knelman JE, Schindlbacher A, Siciliano S, Breulmann M, Yannarell A, et al. Microbes as Engines of Ecosystem Function: When Does Community Structure Enhance Predictions of Ecosystem Processes? Front Microbiol 2016;7:214.
[78] Stearns SC. Life-history tactics: a review of the ideas. Q Rev Biol 1976;51:3–47.
[79] Stearns SC. Trade-Offs in Life-History Evolution. Funct Ecol 1989;3:259–68.
[80] Barnett SE, Egan R, Foster B, Eloe-Fadrosh EA, Buckley DH. Genomic Features Predict Bacterial Life History Strategies in Soil, as Identified by Metagenomic Stable Isotope Probing. MBio 2023;14:e0358422.
[81] Madigan MT, Bender KS, Buckley DH, Sattley WM, Stahl DA, editors. Genetics of Bacteria and Archaea. Brock Biology of Microorganisms. Fifteenth, Global edition, 330 Hudson Street, NY NY 10030: Pearson; 2018, p. 343–4.
[82] Butler S, O'Dwyer JP. Stability criteria for complex microbial communities. Nat Commun 2018;9:2970.
[83] Fierer N, Wood SA, Bueno de Mesquita CP. How microbes can, and cannot, be used to assess soil health. Soil Biol Biochem 2021;153:108111.
[84] Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol 2012;10:538–50.
[85] Cavaliere M, Feng S, Soyer OS, Jiménez JI. Cooperation in microbial communities and their biotechnological applications. Environ Microbiol 2017;19:2949–63.
[86] Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, et al. Microbial biogeography: putting microorganisms on the map. Nat Rev Microbiol 2006;4:102–12.
[87] Green JL, Bohannan BJM, Whitaker RJ. Microbial biogeography: from taxonomy to traits. Science 2008;320:1039–43.
[88] Logares R, Deutschmann IM, Junger PC, Giner CR, Krabberød AK, Schmidt TSB, et al. Disentangling the mechanisms shaping the surface ocean microbiota. Microbiome 2020;8:55.
[89] Chu H, Gao G-F, Ma Y, Fan K, Delgado-Baquerizo M. Soil Microbial Biogeography in a Changing World: Recent Advances and Future Perspectives. mSystems 2020;5. https://doi.org/10.1128/mSystems.00803-19.

[90] Konopka A. What is microbial community ecology? ISME J 2009;3:1223–30.
[91] van den Berg NI, Machado D, Santos S, Rocha I, Chacón J, Harcombe W, et al. Ecological modelling approaches for predicting emergent properties in microbial communities. Nat Ecol Evol 2022;6:855–65.
[92] Vellend M. Conceptual synthesis in community ecology. Q Rev Biol 2010;85:183–206.
[93] Soberon J, Peterson AT. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodivers Inf 2005;2. https://doi.org/10.17161/bi.v2i0.4.
[94] Malard LA, Guisan A. Into the microbial niche. Trends Ecol Evol 2023;38:936–45.
[95] Zhou Jizhong, Ning Daliang. Stochastic Community Assembly: Does It Matter in Microbial Ecology? Microbiol Mol Biol Rev 2017;81:10.1128/mmbr.00002–17.
[96] Holt RD. Bringing the Hutchinsonian niche into the 21st century: ecological and evolutionary perspectives. Proc Natl Acad Sci U S A 2009;106 Suppl 2:19659–65.
[97] Hutchinson GE. Concluding remarks. In: Cold Spring Harb Symp Quant Biol 1957.
[98] Muller EEL. Determining Microbial Niche Breadth in the Environment for Better Ecosystem Fate Predictions. mSystems 2019;4. https://doi.org/10.1128/mSystems.00080-19.
[99] Cohan FM. What are bacterial species? Annu Rev Microbiol 2002;56:457–87.
[100] Albertsen M, Karst SM, Ziegler AS, Kirkegaard RH, Nielsen PH. Back to Basics--The Influence of DNA Extraction and Primer Choice on Phylogenetic Analysis of Activated Sludge Communities. PLoS One 2015;10:e0132783.
[101] Větrovský T, Baldrian P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS One 2013;8:e57923.
[102] Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol 2017;8:2224.
[103] Vos M, Wolf AB, Jennings SJ, Kowalchuk GA. Micro-scale determinants of bacterial diversity in soil. FEMS Microbiol Rev 2013;37:936–54.
[104] Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, et al. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 2007;5:e77.
[105] Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, et al. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 2007;5:e16.
[106] Sunagawa S, Acinas SG, Bork P, Bowler C, Tara Oceans Coordinators, Eveillard D, et al. Tara Oceans: towards global ocean ecosystems biology. Nat Rev Microbiol 2020;18:428–45.

[107] Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh H-J, Cuenca M, et al. Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. Cell 2019;179:1068–83.e21.
[108] Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Ocean plankton. Structure and function of the global ocean microbiome. Science 2015;348:1261359.
[109] Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature 2007;449:804–10.
[110] Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010;464:59–65.
[111] Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol 2014;32:834–41.
[112] NIH Human Microbiome Portfolio Analysis Team. A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007-2016. Microbiome 2019;7:1–19.
[113] Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature 2011;473:174–80.
[114] Shaffer JP, Nothias L-F, Thompson LR, Sanders JG, Salido RA, Couvillion SP, et al. Standardized multi-omics of Earth's microbiomes reveals microbial and metabolite diversity. Nat Microbiol 2022;7:2128–50.
[115] Global Soil Biodiversity Initiative. Global Soil Biodiversity Initiative n.d. https://www.globalsoilbiodiversity.org/ (accessed April 18, 2024).
[116] Qiu J. It's a microbial world. Nature Publishing Group UK 2010. https://doi.org/10.1038/news.2010.190.
[117] Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. Nat Biotechnol 2021;39:499–509.
[118] Schmidt TSB, Fullam A, Ferretti P, Orakov A, Maistrenko OM, Ruscheweyh H-J, et al. SPIRE: a Searchable, Planetary-scale mIcrobiome REsource. Nucleic Acids Res 2024;52:D777–83.
[119] Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, et al. A global atlas of the dominant bacteria found in soil. Science 2018;359:320–5.
[120] Delgado-Baquerizo M, Eldridge DJ, Liu Y-R, Sokoya B, Wang J-T, Hu H-W, et al. Global homogenization of the structure and function in the soil microbiome of urban greenspaces. Sci Adv 2021;7. https://doi.org/10.1126/sciadv.abg5809.
[121] Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, et al. Structure and function of the global topsoil microbiome. Nature 2018;560:233–7.
[122] Ma B, Lu C, Wang Y, Yu J, Zhao K, Xue R, et al. A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. Nat Commun 2023;14:7318.
[123] Labouyrie M, Ballabio C, Romero F, Panagos P, Jones A, Schmid MW, et al. Patterns in soil microbial diversity across Europe. Nat Commun 2023;14:3311.

[124] Chuckran PF, Flagg C, Propster J, Rutherford WA, Sieradzki ET, Blazewicz SJ, et al. Edaphic controls on genome size and GC content of bacteria in soil microbial communities. Soil Biol Biochem 2023;178:108935.
[125] Mortimer N, Campbell HJ, Tulloch AJ, King PR, Stagpoole VM, Wood RA, et al. Zealandia: Earth's hidden Continent. GSA Today 2017;27:27–35.
[126] Mortimer N, Williams S, Seton M, Calvert A, Waight T, Turnbull R, et al. Reconnaissance basement geology and tectonics of North Zealandia. Tectonics 2023;42. https://doi.org/10.1029/2023tc007961.
[127] Hermans SM, Buckley HL, Case BS, Curran-Cournane F, Taylor M, Lear G. Using soil bacterial communities to predict physico-chemical variables and soil quality. Microbiome 2020;8:79.
[128] Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet 2011;52:413–35.
[129] Tessler M, Neumann JS, Afshinnekoo E, Pineda M, Hersch R, Velho LFM, et al. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. Sci Rep 2017;7:6589.
[130] Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, et al. Sequencing and beyond: integrating molecular "omics" for microbial community profiling. Nat Rev Microbiol 2015;13:360–72.
[131] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016;17:333–51.
[132] Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods 2022;19:823–6.
[133] Tanner MA, Goebel BM, Dojka MA, Pace NR. Specific ribosomal DNA sequences from diverse environmental settings correlate with experimental contaminants. Appl Environ Microbiol 1998;64:3110–3.
[134] Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J. Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. PLoS One 2014;9:e94249.
[135] Mitra A, Skrzypczak M, Ginalski K, Rowicka M. Strategies for achieving high sequencing accuracy for low diversity samples and avoiding sample bleeding using illumina platform. PLoS One 2015;10:e0120520.
[136] Motley ST, Picuri JM, Crowder CD, Minich JJ, Hofstadler SA, Eshoo MW. Improved multiple displacement amplification (iMDA) and ultraclean reagents. BMC Genomics 2014;15:443.
[137] Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 2017;35:833–44.
[138] Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp 2012;2:3.
[139] Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vázquez-Baeza Y, et al. Meta-analyses of studies of the human microbiota. Genome Res 2013;23:1704–14.

[140] Probst AJ, Weinmaier T, DeSantis TZ, Santo Domingo JW, Ashbolt N. New perspectives on microbial community distortion after whole-genome amplification. PLoS One 2015;10:e0124158.
[141] Cuthbertson Leah, Rogers Geraint B., Walker Alan W., Oliver Anna, Hafiz Tarana, Hoffman Lucas R., et al. Time between Collection and Storage Significantly Influences Bacterial Sequence Composition in Sputum Samples from Cystic Fibrosis Respiratory Infections. J Clin Microbiol 2020;52:3011–6.
[142] Delmont TO, Robe P, Clark I, Simonet P, Vogel TM. Metagenomic comparison of direct and indirect soil DNA extraction approaches. J Microbiol Methods 2011;86:397–400.
[143] Analysis of the Effect of a Variety of PCR Inhibitors on the Amplification of DNA Using Real Time PCR, Melt Curves and STR Analysis. National Institute of Justice n.d. https://nij.ojp.gov/library/publications/analysis-effect-variety-pcr-inhibitors-amplification-dna-using-real-time-pcr (accessed March 18, 2024).
[144] Funes-Huacca ME, Opel K, Thompson R, McCord BR. A comparison of the effects of PCR inhibition in quantitative PCR and forensic STR analysis. Electrophoresis 2011;32:1084–9.
[145] Schrader C, Schielke A, Ellerbroek L, Johne R. PCR inhibitors - occurrence, properties and removal. J Appl Microbiol 2012;113:1014–26.
[146] Braid MD, Daniels LM, Kitts CL. Removal of PCR inhibitors from soil DNA by chemical flocculation. J Microbiol Methods 2003;52:389–93.
[147] Sharma S, Sharma KK, Kuhad RC. An efficient and economical method for extraction of DNA amenable to biotechnological manipulations, from diverse soils and sediments. J Appl Microbiol 2014;116:923–33.
[148] Hurt RA Jr, Robeson MS 2nd, Shakya M, Moberly JG, Vishnivetskaya TA, Gu B, et al. Improved yield of high molecular weight DNA coincides with increased microbial diversity access from iron oxide cemented sub-surface clay environments. PLoS One 2014;9:e102826.
[149] Yankson KK, Steck TR. Strategy for extracting DNA from clay soil and detecting a specific target sequence via selective enrichment and real-time (quantitative) PCR amplification. Appl Environ Microbiol 2009;75:6017–21.
[150] Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. Inexpensive multiplexed library preparation for megabase-sized genomes. PLoS One 2015;10:e0128036.
[151] Gaio D, Anantanawat K, To J, Liu M, Monahan L, Darling AE. Hackflex: low-cost, high-throughput, Illumina Nextera Flex library construction. Microb Genom 2022;8. https://doi.org/10.1099/mgen.0.000744.
[152] Mora-Castilla S, To C, Vaezeslami S, Morey R, Srinivasan S, Chousal JN, et al. Miniaturization Technologies for Efficient Single-Cell Library Preparation for Next-Generation Sequencing. J Lab Autom 2016;21:557–67.
[153] Minich JJ, Humphrey G, Benitez RAS, Sanders J, Swafford A, Allen EE, et al. High-Throughput Miniaturized 16S rRNA Amplicon Library Preparation Reduces Costs while Preserving Microbiome Integrity. mSystems 2018;3. https://doi.org/10.1128/mSystems.00166-18.

[154] Mayday MY, Khan LM, Chow ED, Zinter MS, DeRisi JL. Miniaturization and optimization of 384-well compatible RNA sequencing library preparation. PLoS One 2019;14:e0206194.
[155] Rausch P, Rühlemann M, Hermes BM, Doms S, Dagan T, Dierking K, et al. Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. Microbiome 2019;7:133.
[156] Dueholm MKD, Nierychlo M, Andersen KS, Rudkjøbing V, Knutsson S, MiDAS Global Consortium, et al. MiDAS 4: A global catalogue of full-length 16S rRNA gene sequences and taxonomy for studies of bacterial communities in wastewater treatment plants. Nat Commun 2022;13:1908.
[157] Dueholm MS, Andersen KS, McIlroy SJ, Kristensen JM, Yashiro E, Karst SM, et al. Generation of Comprehensive Ecosystem-Specific Reference Databases with Species-Level Resolution by High-Throughput Full-Length 16S rRNA Gene Sequencing and Automated Taxonomy Assignment (AutoTax). MBio 2020;11. https://doi.org/10.1128/mBio.01557-20.
[158] Overgaard CK, Tao K, Zhang S, Christensen BT, Blahovska Z, Radutoiu S, et al. Application of ecosystem-specific reference databases for increased taxonomic resolution in soil microbial profiling. Front Microbiol 2022;13:942396.
[159] Dueholm MKD, Andersen KS, Petersen A-KC, Rudkjøbing V, MiDAS Global Consortium for Anaerobic Digesters, Nielsen PH. MiDAS 5: Global diversity of bacteria and archaea in anaerobic digesters. bioRxiv 2023:2023.08.24.554448. https://doi.org/10.1101/2023.08.24.554448.
[160] Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol 2018;16:410–22.
[161] Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 2017;11:2639–43.
[162] Bruinsma S, Burgess J, Schlingman D, Czyz A, Morrell N, Ballenger C, et al. Bead-linked transposomes enable a normalization-free workflow for NGS library preparation. BMC Genomics 2018;19:722.
[163] Woodcroft BJ, Aroney STN, Zhao R, Cunningham M, Mitchell JAM, Blackall L, et al. SingleM and Sandpiper: Robust microbial taxonomic profiles from metagenomic data. bioRxiv 2024:2024.01.30.578060. https://doi.org/10.1101/2024.01.30.578060.
[164] Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2017;2:1533–42.
[165] Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data 2018;5:170203.
[166] Singleton CM, Petriglieri F, Kristensen JM, Kirkegaard RH, Michaelsen TY, Andersen MH, et al. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. Nat Commun 2021;12:2009.

[167] Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 2013;31:533–8.
[168] Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol 2018;3:836–43.
[169] Nissen JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJA, Grønbech CH, et al. Improved metagenome binning and assembly using deep variational autoencoders. Nat Biotechnol 2021;39:555–60.
[170] Lamurias A, Sereika M, Albertsen M, Hose K, Nielsen TD. Metagenomic binning with assembly graph embeddings. Bioinformatics 2022;38:4481–7.
[171] Wang Z, You R, Han H, Liu W, Sun F, Zhu S. Effective binning of metagenomic contigs using contrastive multi-view representation learning. Nat Commun 2024;15:585.
[172] Beaulaurier J, Zhu S, Deikus G, Mogno I, Zhang X-S, Davis-Richardson A, et al. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. Nat Biotechnol 2018;36:61–9.
[173] Wilbanks EG, Doré H, Ashby MH, Heiner C, Roberts RJ, Eisen JA. Metagenomic methylation patterns resolve bacterial genomes of unusual size and structural complexity. ISME J 2022;16:1921–31.
[174] Enam SU, Cherry JL, Leonard SR, Zheludev IN, Lipman DJ, Fire AZ. Restriction Endonuclease-Based Modification-Dependent Enrichment (REMoDE) of DNA for Metagenomic Sequencing. Appl Environ Microbiol 2023;89:e0167022.
[175] Cao L, Kong Y, Fan Y, Ni M, Tourancheau A, Ksiezarek M, et al. mEnrich-seq: methylation-guided enrichment sequencing of bacterial taxa of interest from microbiome. Nat Methods 2024;21:236–46.
[176] Illumina DNA Prep n.d. https://emea.illumina.com/products/by-type/sequencing-kits/library-prep-kits/illumina-dna-prep.html (accessed March 18, 2024).
[177] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 2013;41:D590–6.
[178] McDonald D, Jiang Y, Balaban M, Cantrell K, Zhu Q, Gonzalez A, et al. Greengenes2 unifies microbial data in a single reference tree. Nat Biotechnol 2023. https://doi.org/10.1038/s41587-023-01845-1.
[179] 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:D733–45.
[180] International Code of Nomenclature of Prokaryotes. Int J Syst Evol Microbiol 2019;69:S1–111.
[181] Whitman WB, Chuvochina M, Hedlund BP, Hugenholtz P, Konstantinidis KT, Murray AE, et al. Development of the SeqCode: A proposed nomenclatural code for uncultivated prokaryotes with DNA sequences as type. Syst Appl Microbiol 2022;45:126305.

[182] Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol 2014;12:635–45.
[183] Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 2013;31:814–21.
[184] Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, et al. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol 2020;38:685–8.
[185] Liu Y-X, Qin Y, Chen T, Lu M, Qian X, Guo X, et al. A practical guide to amplicon and metagenomic analysis of microbiome data. Protein Cell 2021;12:315–30.
[186] Chao A, Gotelli NJ, Hsieh TC, Sander EL, Ma KH, Colwell RK, et al. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecol Monogr 2014;84:45–67.
[187] Chao A, Chiu C-H, Jost L. Phylogenetic Diversity Measures and Their Decomposition: A Framework Based on Hill Numbers. In: Pellens R, Grandcolas P, editors. Biodiversity Conservation and Phylogenetic Systematics: Preserving our evolutionary heritage in an extinction crisis, Cham: Springer International Publishing; 2016, p. 141–72.
[188] Ramette A. Multivariate analyses in microbial ecology. FEMS Microbiol Ecol 2007;62:142–60.
[189] Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 2005;71:8228–35.
[190] Magurran AE. Measuring biological diversity. Curr Biol 2021;31:R1174–7.
[191] Rivett DW, Bell T. Abundance determines the functional role of bacterial phylotypes in complex communities. Nat Microbiol 2018;3:767–72.
[192] Astudillo-García C, Bell JJ, Webster NS, Glasl B, Jompa J, Montoya JM, et al. Evaluating the core microbiota in complex communities: A systematic investigation. Environ Microbiol 2017;19:1450–62.
[193] Saunders AM, Albertsen M, Vollertsen J, Nielsen PH. The activated sludge ecosystem contains a core community of abundant organisms. ISME J 2016;10:11–20.